|  |
| --- |
| IU |
| Advanced Data Analysis |
| DLBDSEDA01 |

# Learning Objectives

The most powerful asset an organization has at its disposal today is data. How to collect the necessary information and how to use that information to directly benefit the organization is one of the main focus points in large companies. Choosing the right indicators to gather insights for future processes is proving to be financially and organizationally beneficial for a large number of organizations.

Your **Advanced Data Analysis** journey will begin by gaining knowledge about the key performance indicators (KPIs), the way a company can choose the best indicators to measure, and how to gain the desired outcome by utilizing these KPIs. Enhancing business processes analytics and process mining provides the best way to understand complicated business processes, find possible bottlenecks, and suggest improvements.

Once the metrics have been selected, you may notice that a significant proportion is in textual form. Professionals identify and analyze crucial information in their everyday working process. But what if the amount of information is ever expanding and it is impossible to examine each piece of data? Machines are proving able to perform complex text processing by identifying topics, emotions, and more. Knowing the right techniques and tools to process and automatically gain knowledge from the textual data at one’s disposal is the best asset to have in this world dominated by an excessive amount of information. This proliferation of data is exemplified by online reviews and opinions; the analysis of these data is impacting fields such as marketing and e-commerce.

By the end of this course, you will be able to understand and utilize web analytics techniques and have an overview of the techniques and tools used in social media analytics. Finally, the techniques and methods linked with experimentation and testing will provide the required knowledge to utilize the newest approaches used in online business and marketing.

# Unit 1 – Business Performance Analytics

**Study Goals**

On completion of this unit, you will be able to …

… define the term KPI.

… determine the KPI design process.

… interpret the most common business performance indicators.

… assess business process mining.

… break down the steps from event logs to process models.

# 1. Business Performance Analytics

## Case Study

Toyota is a Japanese automobile manufacturer. It is one of the largest automobile manufacturers in the world (List of manufacturers by motor vehicle production, 2021) and also one of the largest companies in the world in terms of revenue (Fortune, 2021). Toyota has contributed to sustainable development since its founding in 1937 by leading in manufacturing and providing society with innovative and high-quality goods and services (Toyota Motor Corporation., n.d.). They believe in the “Toyota Way,” which is a coherent system of values based on two pillars: continuous improvement and respect for people (Toyota Industries Corporation, n.d.-a).

Toyota has linked its performance, measured by key performance indicators, to three pillars: social performance, environmental performance, and economic performance (Toyota Motor Corporation., n.d.). Toyota’s financial review for the financial year 2012, which followed four years of global economic decline, was successful in motivating managers to give their KPIs more attention and ownership, as they commit to short-term and long-term goals. They consolidated the usage of KPIs (reducing the number of KPIs from 111 to less than 30), resulting in fewer but stronger indicators. Toyota’s corporate culture continues to focus on providing customers with top quality and kaizen (continuous improvement), by relying on experienced and qualified employees (Toyota Industries Corporation, n.d.-a; Toyota Industries Corporation, n.d.-b).

KPIs linked with social performance include indicators such as injury frequency rates, which concerns the health and safety of their employees; the percentage of women in the total workforce, which concerns equality and non-discrimination; employee training days; and making financial contributions to philanthropic activities. Meanwhile, environmental KPIs measure the company’s ability to minimize environmental impact. They include CO2 emissions, resource efficiency (such as water usage), the quantity of waste being sent to landfills (the target is zero), energy usage, and other indicators. (Toyota Industries Corporation, n.d.-a; Toyota Industries Corporation, n.d.-b).

## 1.1 KPI Design Considerations

Key performance indicators (KPIs) are one of the most powerful tools for gaining an understanding of whether a company is following the right approach to effectively achieve its main strategic goals or not. A KPI is a quantifiable measure used to evaluate the success of an employee, organization, or activity in terms of whether or not the performance objectives have been met (Oxford University Press, n.d.). Choosing the correct indicators among the industry-recognized pool of KPIs is rather challenging, leading to a high failure rate for KPI implementation projects in a considerable number of organizations.

Before developing a strategy for designing KPIs, it is important to understand the company’s strategy and the main objectives the company is seeking to achieve. Far too often, companies create a vast collection of performance indicators, which are easy to measure, but which have a low significance and sometimes even prove to be worthless. It is important to note that each KPI design process is unique and must be tailored to the company’s circumstances and the general industry context.

The SMART (specific, measurable, attainable, relevant, time-bound) criteria is one method that can be used to evaluate the relevant KPI. The SMART criteria approach is commonly attributed to Peter Drucker’s book, *The Practice of Management* (Drucker, 1954), but the first known usage was in 1981 (Doran, 1981). It is beneficial for an organization to ask some questions before trying to integrate an indicator into the KPI design process.

* Is the objective of using this performance indicator specific?
* Is it possible to continuously measure the progress made during the KPI’s integration?
* Is the objective attainable, having regard to the personnel required, financial factors, and other constraints?
* Does this indicator match the strategic goals of the organization? Is it relevant to the organization’s focus?
* What is the time frame? What is going to happen in one day, week, or month?

### The Types of Performance Measures

Managers in small and large organizations include a large number of measures in what is called a balanced scoreboard. The balanced scoreboard, developed in 1992, is considered as a basis of the strategic management of an organization (Kaplan & Norton, 1996). It is organized around four perspectives: finance, customers, internal processes, and learning, which are the foundation of most modern management strategies. By trying to integrate everything into their scoreboard, some organizations tend to incur significant expenses without providing meaningful insights, such an approach can cause confusion and disorganization. According to Parmenter (2019), it is not correct to consider all imaginable measures as KPIs. To provide a better implementation process, it is important to categorize all measures into two main groups: indicators of result (RI) and indicators of performance. Meanwhile, to further emphasize the importance of some measures, the word “key” can be added to each of these terms.

**Result indicators**

A metric, which shows the result of a finished job.

**Result indicators** provide information about the final result of the work performed by different teams in the organization. However, they do not help to fix problems that have been identified. It is impossible to verify the impact each team by only considering the final metrics.

Result indicators are a good way to understand the progress of an organization during a monthly or yearly period. Because these measures are delivered on a monthly or yearly basis, they do not enable the workflow to be improved and the end results do not improve. It is important to mention that results indicators only tell us about what has happened in the past; they cannot help us to understand the current or future situation.

Some of the key result indicators for both the private sector and government or nonprofit organizations include net profit, customer and employee satisfaction, sales percentage, and services availability.

**Performance indicators**

A metric, which directly measures the output of a specific group or team.

On the other hand, **performance indicators** provide a better way to measure each team’s work. These indicators are directly linked with the work of a specific team, making it easier to link the success or failure in relation to the measured metric with a specific team’s work.

The main characteristics of KPIs are (Parmenter, 2019)

* non-financial. A financial measure is primarily considered to be an indicator of the result.
* frequently measured. For a KPI to provide useful feedback, it needs to be measured frequently (on an hourly, daily, or weekly basis), in order to leave room for improvement.
* simple to understand. The chosen KPI should be easily understood by each of the organization’s employees.
* uniquely linked to a team’s work. It is a good practice to link a KPI with a specific team in the organization, or to directly link a developing problem with a team.
* impactful in reaching the organization’s strategy goal. KPIs form the basis of the strategy that is intended to ensure that the organization’s goals are reached.
* minor negative aspects. Before making use of a KPI, it needs to be tested in order to see what kind of negative consequences may arise. The significance and severity of the possible consequences must be assessed.

Key performance indicators, for both the private sector and government or nonprofit organizations, include the number of initiatives, late projects, response times for problems that arise, complaints that remain unresolved, and many other measurables.

### Differences between KRIs and KPIs

To better understand the terms “key result indicators” and “key performance indicators,” it is best to compare their main characteristics. The table below shows the main differences between these two types of measurements.

|  |  |
| --- | --- |
| **Comparison of KRIs and KPIs** | |
| **Key result indicators** | **Key performance indicators** |
| Financial or nonfinancial | Nonfinancial |
| Monthly or yearly | Hourly, Daily, or Weekly |
| Reported to meetings to show the organization’s progress | Constantly made available to the staff |
| Gives no information about the possible points that needs to be fixed | Highlight problems which can later be fixed by the responsible staff |
| Is a summary of many activities | Is focused on a specific action |

Source: Alba Haveriku, based on Parmenter (2019)

It is important to round out the above comparison by pointing out that a financial indicator, or an indicator that is measured more often than monthly, can never be a KPI (Parmenter, 2019). As opposed to result indicators, KPIs are useless if they are not measured constantly during a day or week. Only by providing a timely measurement of KPIs can an organization increase their profit and thereby improve their overall performance.

### KPI Design Process

Organizations often fail to implement meaningful key performance indicators. This usually happens when an organization selects KPIs without preparing a well thought plan of implementation. It is important to first create the perfect environment in which the KPIs are going to be developed. Each KPI needs to be customized to the organization’s current situation and objectives.

The following are some of the most important questions to be asked before including any indicator:

* What is the desired outcome of using this KPI?
* How is using this KPI going to affect the organization?
* Which employee or team is responsible for this outcome?
* How frequently is the progress going to be measured?
* If the outcome is not contributing to the organization’s strategic goals, how can the KPI be modified?

The difficulties linked with the implementation of KPIs are well known, and practices have been developed to help organizations and employees better understand KPIs and their implementation. The first manual was developed by AusIndustry (an Australian Government Department) and Baker in 1996 (AusIndustry & Baker, 1999) This manual proposed 12 important steps that need to be taken during the KPI design process. However, there is variation in the advice, for example, Parmenter (2019) specifies a three-stage process, which does include all 12 steps, but is more user friendly and easier to implement.

The three stages are as follows (Parmenter, 2019):

1. The organization should fully commit to the KPI integration and an in-house KPI team should be established. The KPI team should be composed of highly qualified members, who are well-acquainted with the strategic goals of the organization.
2. The critical success factors of the organization must be determined.
3. The correct measures that affect the organization’s performance must be chosen.

The following are the primary considerations in this three-stage process:

* Collaborate with the staff and other interested third parties. A KPI is useless if it is not communicated in the right manner. When considering a KPI, the management team should ensure that all relevant staff and stakeholders know what they need to measure and what benefits will flow from this measurement. Communicating and discussing which key metrics to measure will help the employees to better understand the organization’s goals. The employees may even come up with bright new ideas themselves.
* Empower staff and delegate decision-making power to the relevant team. The organization’s employees whose work is directly linked with measured KPIs should have the power to make immediate changes and improvements, depending on the situation they are observing. This leads to staff empowerment and better problem management.
* Measure and report the main performance indicators. Each measure should be important and linked to a critical success factor. Everything that gets reported should be followed up in a timely manner. Reports should be concise and focused on performance changes. Unnecessary reporting processes should be disposed of or modified to avoid time being wasted in the creation of reports that will never be read.
* Link the KPI with the critical success factors of the organization. The main reason for using performance measures is to ensure that each employee is entirely focused on the critical factors associated with the organization. Employees can utilize the relevant KPIs and success factors to ensure that they are performing as expected.
* Do not retain processes that are not productive. Ignoring or abandoning some processes can be seen as a sign of good management (Drucker, 1954). Abandoning a process that no longer provides a tangible benefit, frees up time that can be allocated to more important indicators. Reports and meetings that do not result in constructive feedback can be reduced or canceled, if they do not contribute to the achievement of the strategic goals of the organization.
* Establish an in-house KPI team. If an organization utilizes their KPIs in their everyday performance measurements, they need to build a KPI development team. This team should be composed of in-house staff who are familiar with the strong and weak points of the organization. This team’s members and the management team should collaborate closely while determining the strategic direction of the company. Dean Spitzer suggests the establishment of a new position “chief measurement officer,” who is responsible for the KPI implementation (Spitzer, 2007). In a large organization, this officer should be working full-time on the KPI project; in smaller organizations this person should dedicate the largest proportion of their workload to KPIs. Following the theory “never give a new job to a new person” (Drucker, 1954), the KPI team leader should be well respected, multifunctional, and collaborative.
* Ensure that everyone in the organization fully understands the chosen KPIs. Each chosen KPI should be clearly understood by employees on all levels. The organization’s staff should be properly trained, so that they understand what a KPI is and how it is going to influence the organization’s overall performance.

In addition to the main points described thus far, a further important aspect is the data at the organization’s disposal. Once the management team has developed the questions that must be asked, they need to identify the necessary data that could provide the answers. A proper evaluation of the gap between the actual information in the company’s existing systems and the data that should be further collected is of high importance. Large and mid-sized companies are usually already collecting all sorts of information directly linked with KPIs. They just need to figure out what they have and how they can use it.

### Self-Check Questions

1. Please list the three main stages composing the KPI design process.

*Creation of the in-house KPI team.*

*Determination of the critical success factors of the organization.*

*Selecting the correct indicators to measure.*

1. Please mark the correct statements.

* *Net profit is a possible key result indicator for an organization.*
* Key performance indicators can be financial in nature.
* It is best to measure a KPI on a yearly basis.

1. Please define the sentence below as true or false.

To better manage the KPI design process, there is a need to create a KPI team, which should be composed of individuals who are not currently working within the organization. True/*False*

## 1.2 Common Business Performance Indicators

When choosing the best indicators, there is no general “best” KPI. The organization’s management team should select the best KPIs, according to the organization’s main objectives. Choosing the right set of performance indicators will directly impact the performance of the organization and emphasize the weakest points, where measures need to be taken. At first, each organization struggles while trying to identify these indicators. They may have a large amount of data at their disposal, but they are not clear how to use this data to their benefit.

The most common business performance indicators are discussed below. A comprehensive review of all KPIs is beyond the scope of this course; however, there is a large body of literature on this topic. The interested reader is referred to ServiceNow, n.d.; Marr, 2012; and Marr, 2014.

### Understanding Customers

The following performance indicators have been chosen according to the requirements and characteristics mentioned in the KPI design process. They are grouped into two main categories “understanding customers” and “understanding the internal processes of the company.” For each of the indicators chosen, the five SMART criteria are considered and explained.

#### Net promoter score (NPS)

How likely is it that a customer will recommend the business to a friend? (Marr, 2012).

Specific. Understanding the satisfaction rate of the customer directly defines their level of loyalty. This indicator is rather difficult and expensive to measure. To provide a better insight into the customer’s perception of the organization, it is necessary to conduct in-depth interviews, which usually return results that are difficult to interpret. The NPS is a measurement based on one question—“how likely is that a consumer will recommend your product or service?”

Measurable. The NPS can be measured by using a survey delivered by e-mail, online, or in-person. A scale from 0—10 can be used to calculate the organization’s NPS. Customers can be divided into different groups, those who are loyal to the company and are keen on promoting the company’s work, and those who may be unhappy or dissatisfied and thus focus on damaging the brand’s name. The NPS can be calculated by subtracting the percentage of unsatisfied customer from the percentage of satisfied customers.

Ein Bild, das Text, Screenshot, Zahl, Software enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Marr (2022).

Attainable. The data collected to calculate the NPS comes from surveys completed by current customers. The cost depends on the method chosen to collect this data and will be particularly high if the company choses to use hard copy or in-person surveys. To reduce the cost and staff involved in this process, the company should find a way (if they do not already have a method in place) to automate the data collection process.

Relevant. This simple question provided above helps the organization to understand what percentage of people positively promote the organization and what percentage of people try to harm the organization’s reputation. To get a more in-depth perspective on why the customers are happy or unhappy, the organization can add further questions to the survey, this can help the organization to identify and understand the proper measures to take.

Time-bound. Not all organizations collect customer data frequently. It is a good approach to collect data related to the calculation of the NPS on a continuous basis, at least for a subset of clients. This approach prevents working with outdated data and provides information that can be used to improve the organization-customer relationship.

Example:A company interviewed 2000 customers in order to calculate their NPS. The customers could indicate how likely they were to recommend the product or service by choosing a number from 0 to 10 (with 0 representing not likely, 5 representing a neutral stance, and 10 representing likely).

**Customers’ Scores for a ProductorService**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Score*** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ***Count*** | 20 | 10 | 10 | 100 | 10 | 500 | 50 | 550 | 250 | 250 | 200 |

Source: Alba Haveriku (2021)

The following formula is used to calculate the NPS:

#### Customer retention rate (CRR)

How loyal are the customers to the brand, organization, or service? (Marr, 2012).

Specific. In order to generate profit, an organization must have a sufficient number of customers. It is easier and cheaper to retain the current customers, than it is to find new ones. Consequently, once customers have purchased a service or product, it is important to retain them as customers and to sell them further products or services. This indicator is very important, because it documents the loyalty of customers. Usually, customers that tend to buy more frequently from the same company, are going to continue to do so, even in the future. However, it is vital to avoid looking at the customer retention rate in isolation, it should always be considered together with other customer related performance indicators.

Measurable. The data linked with the retention rate should be easily collected from the organization’s CRM (customer relationship management) system. Otherwise, in-depth surveys should be used to calculate this metric.

The following formula is used to calculate the customer retention rate:

Attainable. The cost and effort required to evaluate the customer retention rate depends on the organization’s data, including the quantity and quality. For example, for a bank that collects and stores large amounts of information about its clients, it is inexpensive to calculate the retention rate. In contrast, for organizations which do not have large amounts of customer data at their disposal, such as restaurants or markets, a survey may be necessary, leading to a higher cost and additional staff engagement.

Relevant. Unfortunately, not all current customers are profitable for an organization. That is why this indicator should be considered alongside other performance indicators and should be carefully evaluated according to the organization’s points of interest.

Time-bound. In terms of time, or frequency of measurement, the retention rate depends on the purchasing cycle or the duration of the contract for a specific product or service. Depending on the type of the product or service, it can be calculated daily, weekly, monthly, or yearly. The longer the timeframe of the contract or purchasing cycle, the more difficult it is to correctly measure or keep track of the CRR.

#### Customer complaints

The question is, how satisfied are the customers?

Specific. An unhappy customer can potentially damage the reputation of an organization, diminishing the quantity of new customers. Unhappy clients usually tell more people about their negative experience, whereas satisfied clients tell only a few other people on average. People who have received a personal recommendation are more likely to purchase and become customers than those who become aware of the organization through other means.

Measurable. Surveys can be used to measure customer satisfaction and determine customer complaints. Small organizations can gather information from short customers surveys, while larger organizations should establish a formal complaint management process. Determining a formula for calculating customer complaints is rather difficult, but organizations may use several metrics to provide a final measure, including

* the total number of complaints made by customers within a day, week, or month.
* the average time required to solve a customer’s complaint.
* the frequency of different types of complaints.
* the percentage of complaints that were resolved.

Attainable. The effort linked with the measurement of the complaint rate is linked with the capabilities of organization. If an organization has a well-established data management system, the measurement of this indicator should not involve any extra costs or effort from the employees. A survey of a large number of customers may result in additional costs for the organization. In the case of large organizations, the existence of a customer service department would be highly beneficial, but it comes with costs, including personnel and equipment costs.

Relevant. In most companies, this indicator is directly linked with the strategic goals of the company. By providing insights into the customers’ opinions about the product or service that they have purchased, the organization can improve faulty or imperfect products or services, in order to gain new customers and to retain current customers in the future.

Time-bound. Customer complaints can be continuously collected (daily or weekly), or more infrequently (monthly or yearly). This depends primarily on the scale of the organization and the type of product or service that is being monitored.

#### Customer turnover rate

What proportion of customers will switch to a different provider over a specific period of time?

Specific. The customer turnover rate is used to estimate the loss of customers. For telecommunication and financial services, for example, it is important to minimize the customer turnover rate, because their customers are highly likely to change the service provider due to more favorable services or better service plans (customers of telecommunications companies may easily change their service provider if they find a better package). Identifying customers that are likely to leave and winning them back should be a focus area for all related companies.

Measurable. For organizations with customer contracts, this measure can be calculated by evaluating how many customer contracts have been renewed. CRM and business intelligence tools can be useful for organizations that have a larger customer base. Sophisticated data mining processes can be used to detect customers that are likely to switch to a different provider, and also determine which the factors dictate this abandonment. By detecting possible turnover and leaving factors, the organization can determine the necessary measures to implement, in order to win back their customers.

Attainable. The cost linked with measuring this indicator depends on the type of industry and the total number of customers. Organizations that provide products or services without establishing significant rapport with their customers may have difficulty keeping track of the number of lost customers. Technological tools and systems (such as CRM tools) are highly beneficial in reducing the costs and personnel required to measure this metric.

Relevant. This performance indicator should be used carefully. The organization must have a clear understanding of why each customer has chosen to stop purchasing the product or service. The reason may have been that the customer received a poor service from the company, but the reason could also be external. For example, a customer may have relocated or perhaps the contract was canceled by the company itself due to non-payment by the customer.

Time-bound. This performance indicator may need to be assessed more frequently by organizations operating within fast-moving sectors (for example, telecommunications organizations); slower moving organizations can assess the turnover rate on a monthly basis.

### Understanding the Internal Processes of the Organization

For each of the indicators chosen, the five SMART criteria are considered and explained.

#### Capacity utilization rate (CUR)

Is the maximum potential attained, taking into consideration the full workload and available resources?

Specific. Capacity utilization provides a way to measure how effectively an organization uses its production capacity. It measures the difference between the actual quantity of product produced or services provided, with the potential output that would have been produced if the organization fully utilized its capacity.

This indicator is significant, because it determines whether an increase in production would result in an increase of production costs. If a company is working at 80 percent utilization rate, it means the production can be increased by 20 percent without incurring any additional costs, because the organization already has this capacity that is currently unused.

Measurable. The full capacity, also called the “potential capacity,” of the organization must be manually calculated and then entered into the following equation. Subsequently, the actual capacity can be calculated based on the information collected from the machine in a given time period.

Attainable.The cost to collect and calculate the data necessary to define the CUR indicator is high, unless a software program is used to automate the entire process.

Relevant. The level of capacity of machines can be estimated by the US Federal Reserve Board’s definition for the Federal Reserve Board’s capacity utilization index (Board of Governors of the Federal Reserve System, n.d.). This index shows the maximal level of production. A proper benchmark that can be used is the economy-wide CUR, which is around 80 percent in the US (federal reserve) and around 82 percent in Europe (Marr, 2012).

Time-bound. CUR can be measured daily or weekly, depending on the production line taken under consideration. If the organization needs to calculate the CUR for a single machine, it can be measured hourly, while the CUR for the whole factory can be measured daily, weekly, or monthly.

**Example:**A production site can produce a maximum of 30,000 units per day. The actual production is 15,000 units per day. To calculate the capacity utilization rate

#### Project schedule variance (PSV)

Are internal projects delivered on time?

Specific. A considerable number of initiatives in small- and large-scale companies are delivered in terms of projects. The timeframe and budget are determined before the project starts. It is necessary to frequently monitor and report on the progress of the project in order to evaluate the current performance and to help identify potential problems in the future performance.

Measurable. When measuring the project variance linked with scheduling time, a general overview is possible. This is because delays usually entail extra costs. For example, the Sydney Opera House was scheduled to open in 1963 at a cost of $7 million and opened in 1973 at a cost of $102 million (Marr, 2012).

The measurement of this indicator is simple, it is just a comparison between the scheduled completion date and the actual completion date. The time can be measured in terms of days or weeks.

Attainable. The costs depend on the company’s tools. If the company is using a management tool for projects, then it will automatically calculate the schedule variance. Otherwise, everything needs to be calculated manually and is going to involve higher costs.

Relevant. Constantly evaluating projects variance against schedule or budget enables future planning and budget approvals.

Time-bound. Depending on the length of the projects, the organization can choose to measure PSV weekly or monthly.

#### Delivery in full, on time (DIFOT) rate

This indicator measures what proportion of customer orders are filled in full, and on time, compared to the total number of customer orders (Marr, 2012).

Specific. On-time delivery provides insight into the need for an organization to fulfill their customer’s requirements. Consumers have increasingly busy lives, and they expect everything to arrive on time. This measurement involves a composite of different factors, ranging from shipping time until the product arrives at the consumer.

Measurable. The delivery in full, on time (DIFOT) rate can be measured as the quantity of units delivered on time as a proportion of all orders made. The data can be collected by using an order tracking system, which tracks the order from dispatch to delivery to the customer.

Attainable. The data can be collected by using the order tracking system, if it is currently available in the organization, which does not add any extra costs. Organizations that do not have a pre-existing system may choose to rely on third parties to deliver their products, which incurs additional costs. Alternatively, the organization may need to conduct a survey (which adds costs) to calculate the DIFOT.

Relevant. The DIFOT rate should not slip below 95 percent, but some companies require a higher DIFOT rate such as 98—99 percent (Marr, 2012).

Time-bound. In order to obtain an accurate DIFOT rate, the organization needs to gather data for each order, meaning the data must be gathered as each order is delivered.

#### Process downtime level

What is the estimated amount of production time lost due to technical breakdowns or staff absence due to illness?

Specific. Downtime is any time, in which, due to machine failures, faults, maintenance, or other problems, the working process is stopped. Downtime is not only associated with manufacturing. It can be linked also, for example, with hospitals, where the hospital needs to measure the time during which different pieces equipment were unavailable.

Measurable. The data necessary to measure this indicator can be collected directly from a machine or can be determined from production process reports. For example, the machine downtime level can be calculated by using the following formula:

Attainable. The costs are moderate, because the machines usually automatically generate data related to different faults or problems.

Relevant. By measuring downtime levels, the organizations can estimate the effectiveness of their internal processes (either equipment or staff).

Time-bound. The process downtime can be measured constantly and should be reported each time this downtime exceeds a pre-defined threshold.

### Self-Check Questions

1. Please list the two primary KPI categories.

*understanding customers*

*understanding the internal processes of the organization*

1. Please mark the correct statement.

* The customer turnover rate specifies how many customers like the company within a specific timeframe.
* *The cost of calculating the machine downtime level is not high, since the information can be directly collected from the machine data*.

1. Please complete the following sentence:

The cost of collecting the data necessary to calculate the CUR (*capacity utilization rate*) is *high*.

## 1.3 Business Process Mining

Traditional methods such as Business Process Management (BPM), Business Intelligence (BI), and Big Data technologies have received attention by improving data models and systems, but they do not offer a way to improve processes based on the event data extracted from the company’s data sources. Process mining is situated in the gap between the analysis based on process models that do not consider real data, and data analyses that do not consider the relevant processes.

### Chemours Case Study

The Chemours Company is a global leader company in chemistry. The company was established in 2015, has approximately 6,500 employees, and serves almost 3,300 customers in 120 countries (Davenport & Spanyi, 2019). The company has initiated a strategic review to drive shareholder value and portfolio focus (Cision PR Newswire, 2021).

DuPont, a global chemical manufacturer, established Chemours as an independent company intended to manage the production of specific products previously produced by DuPont. Chemours inherited all of its business processes, including an ERP system, from DuPont (Davenport & Spanyi, 2019). The attempt to transform the company and to operate with greater efficiency stimulated the interest in process mining. The company acquired the software company Celonis and targeted the order-to-cash (OTC) process (Davenport & Spanyi, 2019). Prior to the process mining project, few people in the organization could properly determine the entire OTC process performed in Chemours, as each employee focused on the role they had to play in the process, without having a broader overview of the process. Over a four-month period, Chemours was able to determine what was really happening in this process and to make the process visible and understandable for the entire company. The organization discovered gaping holes, such as unnecessary credit holds that were placed on some strategic customers, slowing down the business (Davenport & Spanyi, 2019).

Key business benefits from the implementation of process mining techniques were (Lee & Hesel, 2019)

* transparency. The process mining techniques provided a better understanding of the overall OTC process.
* collaboration enhancement. All business units gained access to the same information and can collaborate to make data-driven decisions, based on the information provided by the process mining.
* encouraging development. The insights gathered from process mining leveraged changes in mindsets and behaviors.

The benefits of applying process mining in the OTC at Chemours were directly visible. The company continues to work towards fixing issues and providing simplified and automated processes. Sung Lee, the director of Business Process Transformation, mentioned that process mining contributed to a better understanding of roles and collaboration (Davenport & Spanyi, 2019). The final goals of the company are to focus the workforce on customer-oriented activities and implement better business analytics (Davenport & Spanyi, 2019).

### What is Process Mining?

Interest in data science has been growing rapidly in recent years. The collection, storage, and analysis of data is of fundamental importance to organizations. Despite collecting, storing, and analyzing data, a data scientist needs to understand the correlation between data and operational processes. Process mining links traditional process analysis (business process management tools) with data analysis techniques (machine learning and data mining), providing a new approach with which to improve processes (Aalst, 2016). Event logs, together with process mining, enable an organization to find and fix performance related problems.

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

Process mining emerged as a discipline in the last decade (Aalst, 2016), offering a complete set of tools that support the improvement of business processes. This discipline is built upon process modeling and data mining, but it cannot be seen simply as a fusion of these two disciplines. The techniques associated with process mining make use of event logs to discover the implicated processes, analyze bottlenecks, compare process traces, and propose possible improvements (Aalst, 2016).

The ultimate goal of process mining techniques and tools is the automatic discovery of process models and the possibility of extracting information about each process.

### What is a Process Model? What is it Used For?

A process model determines the flow of work that leads towards the accomplishment of a specific strategic goal, ideally in a graphic format. Different notations and techniques exist to create or visualize a process model. Some of the commonly used notation to model operational business processes are Petri nets, BPMN (Business Process Management Notation), UML (Unified Modeling Language), EPCs (Event-driven Process Chain), and YAWL (Yet another workflow language).

The advancement in technology has resulted in remarkable changes in the business world. Innovations in computing are changing the way business processes are being organized. These business processes have become complex and rely on the organization’s information systems. Using process modeling techniques has become important to manage the complex processes and improve what is called business alignment, the mechanism by which a company aligns its business processes and information systems. The analysis of real processes can show the need for improvement in current information systems, in accordance with the needs of managers and employees.

Making good process models is rather difficult, but fortunately process mining facilitates the faster construction of preferable models. A transition system is one of the basic process modeling notations. It consists of states and transitions from a request until the goal is achieved (Aalst, 2016). The notions used for transition systems can be translated into high-level languages such as Petri nets, BPMN, and UML (Aalst, 2016).

Petri nets are one of the oldest process modeling languages that are suited to model cases of concurrency, synchronization, sharing of limited resources, and other purposes. They are intuitive and executable in the form of a bipartite graph, which can be represented as an n-tuple (P, T, I, O, M) where:

* P sets of places. Places represent the different conditions within the system that is being modeled.
* T set of transitions. Transitions represent events happening in the system that provide a change in the given conditions.
* I, O Input Arcs and Output. Arcs connect places to transitions or the opposite. They are grouped into input and output arcs, where the first one represents the conditions to be fulfilled before an event and the second represents the conditions that exist as a result of the event.
* M marking. A place can be associated with tokens, which condition the enabling or disabling of a specific transition. The marking of a Petri net is none other than e vector which lists the number of tokens in each place of the specific Petri net, represented as (m1 m2 … mp) where p is the number of places.

The figure below represents the basic components of a simple Petri net.

Ein Bild, das Screenshot, Diagramm, Kreis, Text enthält.

Automatisch generierte Beschreibung

Source: Soós Máté (2006). CC-BY-SA-2.5.

When talking about business processes, we can consider a type of Petri nets known as Workflow Nets (WF-nets). In the context of BPM (business process management), WF-nets have a uniquely defined start and end. WF-nets provide a representation of the whole lifecycle of a process instance. Multiple tokens can start from the input place, without getting mixed with each other. Each token represents a case, which follows a road from the input place to the output place, isolated from all other cases. The same model can be used for different instances, such as insurance claims, job or credit applications, and similar activities, meaning that WF-nets are particularly suitable for these types of tasks.

### The Value of Process Mining

When considering process mining, organizations need to define some main questions related to their organization’s performance goals. The proper data need to be collected, and then the process mining process should be performed to provide useful answer to each of the important questions.

The following are the main questions associated with the organization operational processes that can be defined by each organization:

* What happens in a specific business process?
* Why does it happen?
* Are there deviations from the initial models? What are the bottlenecks linked to these deviations?
* How can the processes be better controlled?
* How can the current processes be optimized to improve performance?

Ein Bild, das Text, Screenshot, Diagramm, Design enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Aalst (2016).

The mining analysis process is comprised of four main activities.

1. Gathering the data. The first step towards process mining is collecting the information needed from event logs. These data can be collected from whichever data system that an organization uses, such as Excel spreadsheets, databases, message logs, transactional logs, and ERP systems. However, data can also come from different unstructured sources such as e-mails, PDF documents, web pages, and other sources. That is why the data extraction process should be determined in reliance on the above questions, rather than just collecting whatever data are available. For the collection of both structured and unstructured data, it is important to extract the data and convert it into event logs. Some of the typical formats to store event logs are XES (extensible event stream) and MXML (mining extensible markup language) (Aalst, 2016). Depending on the initial questions, the organization can generate different event logs, which are based on the same set of data extracted. One important point to consider is that after the process mining, new questions may arise, prompting further extraction and analysis of the company’s data.
2. Process discovering and understanding. The process discovering techniques automatically produce a representation of the process model, which determines the current processes and data involved. In this part, a proper understanding of all the resources and interactions among resources provides an insight into how the business operations are linked together. The process discovery process is based on the generated event logs, constructing a model by analyzing event logs.
3. Compliance checking. After the process discovery phase, a compliance check is necessary to identify possible deviations. A comparison between an ideal process model and the derived model from the process discovering techniques is obligatory. This checks and records whether the information system is complying with the standards and project policies, or to what extent they are not compliant. In this process, it is supposed that the process model and event logs are initially given. The goal of compliance checking is to find deviations between the modeled behavior and the observed behavior (Aalst, 2016). This process is relevant to business auditing, evaluating the performance of the techniques used for process discovery or repair models that do not represent reality.
4. Improvement of current business processes. The improvement or enhancement of business processes consists of analyzing the performance of processes in terms of time and avoiding possible deadlocks. In this part of the process mining process, traditional data mining techniques can be used, as all of the information and data are available and can be used to enable better decision making.

### Why Do Organizations Need to Use Process Mining?

Traditional business process management techniques struggle to manage the high volumes of data that are being continuously generated. Process mining provides a method for those organizations that want to make the best use of the data they collect.

The most important reasons to switch toward process mining are

* the alignment of processes. Process mining highlights the variations between the initial process design and what happens during the execution of the processes. This provides a better insight into the impact each activity has on all other areas of the organization’s business processes.
* the perfect end-to-end view of the organization process model. Due to biases and lack of information, traditional methods fail to give the proper understanding of the process behavior.
* elimination of data differences. Data may be generated or saved in a wide range of formats. Process mining determines a way to gather information from different sources.
* reduced processing time. Process mining can automatically generate the process model in a reduced timeframe and with lower costs than traditional process analysis projects.
* data-driven decisions. The essential information is generated automatically, providing the opportunity to make faster changes in different projects.

### Process Mining Tools

To successfully apply process mining, an organization has to use the most appropriate tools relevant to their final goal. Many traditional business intelligence (BI) tools and data mining or machine learning tools do not cover all of the phases of process mining; however, some dedicated process mining tools exist. They provide the best possibility to transform event logs into insights about business processes. ProM is an open-source tool, which supports all of the techniques discussed above, from the process discovery process to conformance and final recommendations for improvements (Aalst, 2016). ProM also allows different formats of data CSV files, MXML, and XES to be loaded. It is one of the most commonly used tools within the academic community. Other tools exist such as Celonis, Disco, Minit, and myInvenio (Aalst, 2016) that are easier to use, but they provide less functionality than ProM in terms of performance analysis and scalability issues.

### Self-Check Questions

1. Please list at least three reasons to switch to process mining.

*the perfect end-to-end view of the organization’s process model*

*elimination of data differences*

*data-driven decisions*

1. Please mark the correct statements.

* Process mining cannot determine the impact each activity has on different areas of the organization.
* *The techniques associated with process mining use event logs to find bottlenecks.*
* *Process mining can automatically generate the process model with lower costs.*

1. Please mark the following statement as true or false.

Process mining techniques provide a better understanding of the overall OTC process. *True*/False

Summary

KPIs are one of the crucial tools used to determine the efficiency and performance of an organization. Choosing the right KPIs and implementing them in the everyday processes of the organization is challenging and for this reason an in-house team, comprised of team members who are familiar with the overall business processes, should be created. This team is then responsible for making changes and improvements, in accordance with the insights gained from the KPIs.

It is important to communicate and discuss each of the metrics chosen with all employees, so that each member of the organization has a clear view of the organization’s main critical success factors. Continuous reporting allows the suitability of the KPIs to be assessed on an on-going basis and improvements to be made when necessary.

Some of the most common KPIs were mentioned, providing information related to their importance and the way we can measure their value. The KPIs discussed fall into two main categories “understanding customers” and “understanding the internal processes of the organization.” In the first category, the main KPIs are the NPS (net promoter score), CRR (customer retention rate), customers complaints, and customer turnover rate. The KPIs in the second category (understanding the internal processes of the organization) are CUR (capacity utilization rate), PSV (project schedule variance), DIFOT rate (delivery in full, on time rate), and process downtime level.

Finally, process mining is situated in the gap between the analysis based on process models that do not consider real data, and data analyses that do not consider the relevant processes. The related tools enable traditional process analysis to be combined with new data analysis techniques. The techniques linked with process mining are focused on event logs, the discovery of business processes, and identification of possible problems and improvements.

# Unit 2 – Text Analytics

**Study Goals**

On completion of this unit, you will be able to …

… define industry-relevant terms such as bag of words and TF-IDF.

… review the best analytical approaches to text mining.

… understand the techniques related to semantic analysis.

… estimate the steps in a natural language processing pipeline.

… solve text related problems using text mining tools.

# 2. Text Analytics

## Case Study

Reading and identifying information is an everyday process for a vast number of professionals. What about machines? How are machines today able to perform complex text processing, identify places, people, emotions and more? Across industries, business owners are valuing their raw text, such as call center transcripts, customer surveys, and other business-related documents. Text mining solutions are leveraging sources of information, providing meaningful insights or predicting future events.

E-commerce websites, such as Amazon, provide a platform to collect and analyze customer’s reviews and generate appropriate ratings. Mellachervu and Minukuri (2018) conducted a study, using the programming language Python, to analyze customer reviews, related to five Bluetooth speakers on Amazon. Python was used to extract the Amazon Standard Identification Number (ASIN) and the range of review pages (Mellachervu & Minukuri, 2018). Among the variables involved in the dataset were author, user rating, and comments. A term by document matrix was generated to identify the most frequently occurring words in each review (such as speaker, charge, sound, and quality). After applying some word filters the terms were able to be combined into 20 main topics. The thorough analysis conducted provided an insight into the bestselling product and the responsible features that are linked with a higher ranking (quality of sound, battery life, and stability of the Bluetooth connection). Similar analyses help companies and retailers to better understand customer expectations and improve their brand’s reputation.

## 2.1 Word and Document Frequency (TF-IDF)

Text communication is one of the most commonly used forms of communication. Different activities such as messages, e-mails, and social media platforms generate a significant amount of text in unstructured form. Organizations are trying to find the best way to analyze textual data, to better understand their customers’ opinions. Text analytics or “text mining” are terms used to describe the whole process of identifying patterns within textual data. Natural language processing (NLP) is one of the main components of text mining used to generate insights for an organization. Detecting the frequency with which a word appears in a text or document gives an oversight of the most important words in a corpus. A corpus is the collection of written texts or documents where we apply different natural language processing techniques.

### The Semantic Space

To perform analytics and machine learning techniques on text, the documents that compose a corpus should be transformed into numerical representations. Specifically, the words are transformed into a vector representation and the process of vectorization is called “feature extraction.” The features of a text or document are information such as length, source, and date of publication.

In natural language processing, we change the way in which we see a language. Now, a document cannot be simply considered as a sequence of words, but as a point in a multidimensional space (semantic space). The closer together the points in the space are, the higher the connection between the documents. Therefore, the semantic space can be considered as a map, which defines the similarities in the meaning of documents in a corpus. Distance between points is the measure that is used to derive decisions in this semantic space.

**Bag of Words (BoW)**

The bag of words represents, in a vectorial form, the word count.

Two commonly used encoding techniques in the semantic space are **bag of words** **(BoW)** and term frequency times inverse document frequency (TF-IDF) (Lane et al., 2019). A bag of words represents, in a vectorial form, the word counts and frequencies; while TF-IDF vectors represent word importance as determined by word scores. These two techniques can be part of an NLP process or they can be applied separately in different contexts. The applications in which they can be used include spam filtering and sentiment analysis.

### The Bag of Words Technique

Bag of words (BoW) is a technique used in NLP for text modelling and for extracting features from the given documents (Lane et al., 2019).Words in a text are represented as a bag (collection), providing a method of feature extraction that can be further used in the document’s classification. This model does not take into consideration information about the structure or the order in which the words are found in a document. It represents each document in a corpus as a vector with a length equal to the vocabulary of the whole corpus (Lane et al., 2019).

Each machine learning model must be able to access data in numerical form, thus the BoW technique is well-suited to the task. The first step entails converting the text into a vector of numbers, which provide a structured input for the machine leaning algorithms that can be used in subsequent steps. This vector is also known as “word frequency vector,” because the order in which the words are saved is not important. What is important is to keep track of the count or frequency of each word.

To better understand how this technique works, let’s take a practical example and explore each of the steps followed in this approach.

#### Example

Two reviews related to a portable Bluetooth speaker have been collected as follows:

**Stop words**

A word which is filtered out in the pre-processing phase of natural language processing, such as articles, prepositions and pronouns.

Review 1: “Loved the sound, no battery issues”

Review 2: “Sound quality is good; battery life not good”

The initial step is the pre-processing phase, the method of which varies in different cases, depending on the expected final result. In this example, during the pre-processing phase, the sentences will be converted into lower case and all special characters will be removed from the text. In this case, **stop words** will not be removed, because we do not want to remove the stop words “not” and “no,” which have an important meaning in these two sentences.

Following the pre-processing phase, the sentences will be

Review 1: “loved the sound no battery issues”

Review 2: “sound quality is good battery life not good”

Note that the words “sound” and “Sound” are now transformed both in lowercase, so that they can be considered as having the same meaning. Also, as you may notice, the punctuation marks have been removed.

After pre-processing and removing the unnecessary information, the number of unique words in the above two review is extracted. The 11 unique words are loved, the, sound, no, battery, issues, quality, is, good, life, and not.

These words compose our model’s vocabulary. Having deducted the length of the vocabulary, we can now create a fixed length vector representation of each review. A vector with 11 dimensions is going to be used for each review; one position in the vector is linked with each word.

For the first review, the count of each word is represented in the below table, detailing the index of each unique word, its value, and the number of times it is found in the review.

**Review 1: Word Frequency**

|  |  |  |
| --- | --- | --- |
| **Index** | **Word** | **Frequency of word** |
| 0 | loved | 1 |
| 1 | the | 1 |
| 2 | sound | 1 |
| 3 | no | 1 |
| 4 | battery | 1 |
| 5 | issues | 1 |
| 6 | quality | 0 |
| 7 | is | 0 |
| 8 | good | 0 |
| 9 | life | 0 |
| 10 | not | 0 |

Source: Alba Haveriku (2021)

The frequencies vector generated for the first review is

review1 = [1 1 1 1 1 1 0 0 0 0 0]

For the second review, the count of words is as follows:

**Review 2: Word Frequency**

|  |  |  |
| --- | --- | --- |
| **Index** | **Word** | **Frequency of word** |
| 0 | loved | 0 |
| 1 | the | 0 |
| 2 | sound | 1 |
| 3 | no | 0 |
| 4 | battery | 1 |
| 5 | issues | 0 |
| 6 | quality | 1 |
| 7 | is | 1 |
| 8 | good | 2 |
| 9 | life | 1 |
| 10 | not | 1 |

Source: Alba Haveriku (2021).

The frequencies vector generated for the second review is

review2 = [0 0 1 0 1 0 1 1 2 1 1]

Now that we have both the reviews represented in the form of a vector, the Euclidean distance between the vectors can be calculated to check the percentage of similarity between the vectors.

To calculate the Euclidean distance, we can create a function which takes as parameters the two vectors and returns the value of the above formula. This calculation can also be performed by using a function from SciPy.

from scipy.spatial.distance import Euclidean

review1 = [1 1 1 1 1 1 0 0 0 0 0]

review2 = [0 0 1 0 1 0 1 1 2 0 0]

distance= euclidean(review1, review2)

In this short example, all the words present in the reviews are part of the vocabulary, but in the larger corpus, only the most important words are chosen to compose the vector and implement the BoW technique. Other steps that can be used to clean the text are (Bengfort et al., 2018)

* the removal of stop words. When using this technique, it is important to ensure that the meaning of the sentence is not altered.
* stemming, a technique that consists of removing the ends from various words (e.g., the removal of “ing” from the word “staying” results in the word stem “stay”).
* Lemmatization, which aims to remove inflectional endings, returning what is called a “lemma” (the base of the word). Lemmatization is more accurate than stemming, because it is based on the meaning of the word. This method handles irregular cases more effectively. For example, in the case of the word “troubled,” lemmatization identifies “trouble” as the base form of the word. Meanwhile, stemming would cut “ed,” converting this word into ‘troubl’, which does not have a proper meaning.
* fixing spelling errors.

The figure below represents all the steps required for the transition to vector representation, starting from the cleaning of text, tokenization (a token can be a character, word or group of words), and the creation of a word vocabulary.

Ein Bild, das Text, Screenshot, Schrift, Marke enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021).

Even though the BoW model is simple to implement, problems can arise in some cases.

* Sparsity. Vectors with a large number of zeros are called sparse vectors. Sparse vectors can make the process challenging, they require significant computational resources.
* Vocabulary. To diminish the sparsity of the vector, the vocabulary needs to be chosen carefully, ensuring that only the most important words are included.
* Meaning. When using BoW, the word order or the structure of a sentence is not taken into consideration, leading towards problems in the general context of the document.

To capture more meaning from a text, vocabularies with n-token sequences can be taken into consideration. An **n-gram** is a sequence that is composed of n elements, extracted from a string. Each token is considered a “gram,” by so constructing a n-gram model (Lane et al., 2019). Groups of words, such as “ice cream,” have both a meaning when read separately and when read together. By considering 2-grams (bigrams) we can insert a new entry into the vocabulary (the two words grouped together). The same applies for 3-grams or more. Using n-grams can improve machine learning processes, such as auto-completing sentences, or chatbot training, to better understand the way a sentence is composed.

### Python Implementation of the BoW Technique

The BoW techniques explained in the above section, can be implemented in Python. Three steps are required to create the vocabulary and vector representation for each of the reviews.

#### Step 1: Importing Python modules

In this phase, there is a need to import the Python modules responsible for transforming words into lower case and the tokenization of sentences.

From the nltk.tokenize package we need to import the method word\_tokenize to tokenize a string and remove punctuation. Meanwhile, defaultdict, part of the collection’s module, is a specialized container, similar to Python dictionaries. The only exception is that defaultdict does not raise KeyErrors, returning the default value for keys that do not exist.

import numpy as np

from nltk.tokenize import word\_tokenize

from collections import defaultdict

#### Step 2: Define the sample corpus and preprocess the text

In this step, the sample corpus is defined and the methods from the Python modules mentioned in the first step are used to clean the text. After the tokenization of the sentences, the functionsw.lower() and w.isalpha() are used to transform the text into lowercase and to check whether the string is textual so that punctuation can be removed. The vocabulary, which is composed of the unique words that appear in the corpus, is created and printed.

corpus = ['Loved the sound, no battery issues','Sound quality is good; battery life not good']

reviews=[]

vocabulary=[]

for r in corpus:

x=word\_tokenize(r)

rev=[w.lower() for w in x if w.isalpha()]

sentences.append(rev)

for w in rev:

if w not in vocabulary:

vocabulary.append(w)

len\_vocabulary = len(vocabulary)

print(vocabulary)

index = {}

i = 0

for w in vocabulary:

index[w] = i

i += 1

The index dictionary is created in such a way that each word in the created vocabulary is linked to a value in the index from 0 to 10.

#### Step 3: Create the BoW

To create the bag of words for this example, we create a cycle which iterates in each of the reviews found in the corpus. A vector representation is created for each review that was given as input. The method defaultdict(int) automatically assigns a value of 0 to any key that does not yet exist in the dictionary. Thenp.zeros returns a new array of a given length, filled with zeros.

for i,x in enumerate(reviews):

vector = np.zeros(len\_vocabulary)

count\_dictionary = defaultdict(int)

for j in x:

count\_dictionary[j] += 1

for k,t in count\_dictionary.items():

vector[index[k]] = t

print("BoW for review "+str(i)+":")

print(vector)

In the end as output, we get the vector representation for each of the reviews, which should correspond with the table we created before for this example.

#### BoW model by using Sklearn

The above implementation shows each step to help better explain the logic behind this algorithm; however, there are more efficient ways of performing this implementation. One would be by using the Sklearn package with its predefined methods and the pandas package which would provide a better final representation of the vector representations. The main function in Sklearn used to create the BoW model is CountVectorizer(). The n-gram range in this case can be defined, depending on the need to implement 2-grams, 3-grams or more. (CountVectorizer(ngram\_range=(1,1))

The method fit\_transform learns the vocabulary dict and returns a matrix which links documents (in this case, reviews) with words (terms). Meanwhile pandas.DataFrame constructs a dataframe representation of the created matrix (data). The dataframe can label each column in the table with the words from the vocabulary and link them with the corresponding value.

import pandas as pd

from sklearn.feature\_extraction.text import CountVectorizer

review\_1="Loved the sound, no battery issues"

review\_2="Sound quality is good; battery life not good"

vect = CountVectorizer()

data = vect.fit\_transform([review\_1,review\_2])

data=pd.DataFrame(data.toarray(),columns=vect.get\_feature\_names())

print(data)

Using the get\_feature\_name() method defines the terms for each column. The final result is the same as in the previous method but is printed in a more structured format.

Ein Bild, das Text, Schrift, Zahl, Webseite enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021).

If specific words should be removed, the English stop words can be downloaded from thenltk module or be defined as a parameter in the CountVectorizermethod.

from ntlk.corpus import stopwords

nltk.download(‘stopwords’)

Alternatively, we can define the stop words we want to remove ourselves, and in doing so retain more control over the way the unnecessary text is cleaned from out of the information.

stopwords=["the","is","be"]

### Term Frequency-Inverse Document Frequency (TF-IDF)

After the vocabulary is created, the occurrence of words is defined, where scoring methods include the counting of the words occurrence in a document, or the frequency with which the word appears related to other words. The BoW represents a document alone, but not the context of the whole corpus. Frequently used words are not necessarily the most important ones in a document. For example, word such as “the” do not provide any information relevant to the content of a document and even though it is commonly found in each document, it cannot be considered significant.

The two components of Term Frequency times Inverse Document Frequency (TF-IDF) are defined as follows (Bengfort, 2018):

* Term Frequency (TF). The number of times a word occurs in a document (Bengfort, 2018).
* Inverse Document Frequency (IDF). Scoring of how rare a word is across all documents (Bengfort, 2018).

The TF-IDF method is an extension of bag of words, providing a numerical statistic that reflects the importance of a specific word relative to the whole document collection. For example, suppose the word “computer” was found 10 times in document A and 100 times in document B. It may seem that the word “computer” is more relevant to document B. But what if we compare the importance of the word with the whole document? The perspective changes when it is discovered that document A is just a 20-word document, meanwhile document B is a book containing 100,000 words.

Calculating the term frequency demonstrates that TF gives a larger value for less frequent words in a document.

The inverse document frequency (IDF) is used to calculate the weight of rare words across the entire corpus. It is estimated logarithmically as the inverse fraction of the documents that contain a specific word (Lane et al., 2019).

To better understand IDF, imagine a document where a word appears many times. The same word is very rare in other documents. The final assumption is that this word is specifically important to the initial document, that’s what IDF is all about.

**Example:** The word “laptop” is found in just 1 out of 100,000 documents that compose the whole corpus. Meanwhile, the word “computer” is found in 10 documents. To calculate the IDF

The final formulas that we have for a word w, in a document doc, in a corpus C are:

Although these formulae may not be used on a day-to-day basis, it is important to understand them as the steps within natural language processing. TF-IDF relates a specific word to a specific document in the whole corpus, and assigns an importance to this word, leading towards topic analysis. The higher the value of the TF-IDF, the more significant the word is, which makes it beneficial to use in search engines such as Google, as a factor in ranking content.

#### Example:

We consider again the example analyzed to understand the BoW model. The vocabulary created for the example was composed of the words loved, the, sound, no, battery, issues, quality, is, good, life, and not. We recall that the two reviews taken in consideration were:

Review 1: “Loved the sound, no battery issues”

Review 2: “Sound quality is good; battery life not good”

The total number of words for the first review is six and for the second review is eight. Taking into consideration the second review, we can calculate the term frequency for each word.

Likewise

Similarly, we can calculate the term frequencies for words in the first review. The inverse document frequency is calculated for each of the words found in the vocabulary.

Words which have an IDF value reduced to 0 are of little importance, while words with higher values have a larger significance. Having collected both the TF and IDF value, we can now calculate the TF-IDF value for each of the words in the reviews.

### Python Implementation of TF-IDF

The scikit-learn package can be used to build a TF-IDF matrix. The method fit\_transform() from theSklearn API creates a matrix of the documents in the corpus and includes the TF-IDF value for each word. Pandas Dataframe is used to generate the matrix and to provide an overview of the calculated values for each word.

### Ein Bild, das Text, Screenshot, Schrift, Zahl enthält. Automatisch generierte BeschreibungSelf-Check Questions

1. Please list three problems that may arise while working with the bag of words model.

*Sparse vectors requiring more computational resources.*

*The structure of a sentence is lost.*

*The meaning of the text is not considered.*

*It can be difficult to choose the right words to include in the vocabulary.*

1. Please mark the correct statement.

* TF determines the total number of times a word is found in a corpus of documents.
* *The higher the value of the IDF, the higher the significance of a word.*
* The fit\_transform method returns a matrix, which links documents with topics.

1. Please fill in the empty spaces in the following sentence.

The nltk.tokenize package, includes the *word\_tokenize* method to tokenize a string and remove *stopwords/punctuation*.

## 2.2 Semantic Analysis

Knowing people’s opinion is the most precious information in business analytics and beyond. Semantic analysis is an ever-growing field, which is being adapted in fields such as marketing, social media, and electronic commerce to gain insight from the content generated all over the internet. It allows machines to interpret the meaning of textual documents, and not just check the word’s count or frequency of appearance. The application of a well-defined NLP pipeline can, for example, be implemented in financial or business analytics processes in an organization. The features generated can be beneficial for making financial predictions, generating statements, or even creating new articles out of the previously processed data.

**Semantic analysis**

The process of extracting topics from natural language texts.

Terms such as “meaning,” “topic,” and “semantic” are usually used interchangeably in natural language processes. **Semantic analysis** describes the process of understanding natural language text (Lane et al., 2019). Different to lexical analysis, which focuses on the meaning of a specific word, semantic analysis focuses on larger parts of texts. It can identify the relationships between words in sentences, paragraphs, or even full documents.

Semantic technology analyses the structure of a sentence or a document to identify the most relevant information and understand the topic that is being discussed. For example, even if the word “economics” does not appear frequently in a text, semantic analysis will be able to recognize that the topic of the text is “economics” by identifying other, related words such as “tax,” “recession,” “money,” and “profit.” By using semantic analysis and NLP, machines are learning to better understand text and provide valuable information linked with customer service or business intelligence.

### Topic Modeling

Topic modeling is a technique which assists in exploring large amount of textual data, finding similarities in documents, and discovering abstract topics. It is generally used in search engines to match a string with the most relevant results. Topic modeling can also be used in recommender systems to provide user specific services, in resume reviews to filter candidates according to the needed skills, in the improvement of customer service to better delegate complaints to respective departments, and in other information retrieval systems.

Two important points can be made about topic models. (Lane et al., 2019).

* A range of topics are discussed in the various documents.

**Latent Semantic Analysis (LSA)**

An unsupervised technique to understand the relationship between documents and topics.

* Each of the topics discussed in the corpus is composed of a collection of specific words.

The semantics of a document can be described by “latent topics” that we need to observe, in order to give a meaning to a specific document. To better understand what happens during the topic modeling process, we are going to examine two popular techniques linked with semantic analysis—**Latent Semantic Analysis (LSA)** and Latent Dirichlet Allocation (LDA)(Lane et al., 2019).

### Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an unsupervised technique in NLP for information retrieval, which analyzes the relationship between a set of words or documents and was first introduced in the late 1980s (Dumais et al., 1988). This technique associates topics in a space with lower dimensions than the space of words. The reduction of dimensions is used to retain the most meaningful properties of the original data. Each word in each document is linked with a topic, contributing to the association of topics with specific documents.

Having analyzed what a bag of words and TF-IDF vectors are, the importance of each word was found related to the whole text. The next step is letting machines understand the meaning or the topic of a text. The frequency scores from the TF-IDF vectors are going to be used to calculate the “topic scores,” grouping words related to their topic into topic vectors. Topic vectors help in the searching process, based on meaning, or what is known as “semantic search.”

LSA is an algorithm that can work on bag of words vectors or TF-IDF vectors. The result when using TF-IDF vectors is slightly better, because it provides a proper insight into the significance of a specific word. After the classification using TF-IDF is finished, a matrix B (m x n) is formed, where mis the number of documents and n is the total number of unique words. The document-term matrix is usually sparse and noisy, leading to the need to find latent topics and to reduce the dimensions of this initial matrix. Latent Semantic Analysis makes use of the matrix constructed and implements Singular Value Decomposition (SVD) to decompose it in a product of three matrices (Blum et al., 2020).

The two main components of the LSA algorithm are

* Singular Value Decomposition
* distributional hypothesis, meaning that words having similar meaning are usually grouped together.

The document-term matrix B (m x n) has already been established, where mis the number of documents and n is the total number of unique words. The user specifies the number of k topics that will be extracted from this matrix. A table can be created in which each document is linked with one, two, or more topics, each with a particular weight. Another aspect would be to define a linkage between the topics and the words in the vocabulary in a tabular form in which rows are the terms and the columns are the topics.

The last correlation that we need to consider is finding a numerical representation of how much a topic explains our corpus. We construct an array of singular values, which helps determine the distribution of topics in our corpus and determine whether we need to consider more than t topics.

B -> m x n matrix

P -> m x k left singular matrix

L -> k x k diagonal matrix

K -> n x k right singular matrix

KT -> k x n, the transpose of K

Ein Bild, das Text, Screenshot, Rechteck, Diagramm enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Blum et al. (2020)

In this case, SVD reduced the dimensions by selecting a k value. k is a parameter chosen to reflect the number of topics that we want to find. This parameter can be determined by using a silhouette coefficient or topic coherence.

LSA, like IDF, describes the dimensions in a vector, which are important to the meaning of the documents. By giving a low variance to a topic, unimportant dimensions (distractions and noise) are identified. It is important to state that the topics are not known beforehand, there is simply a hypothesis that there will be t number of topics, each topic discovered by the model can be subsequently explored.

To sum up, LSA provides more meaningful information in less dimensions, generalizing the representation of the vector for subsequent machine learning related processes. The SVD algorithm identifies words that are commonly seen together and categorizes them under the same topic. It properly defines the dimensions that are more relevant to the semantics or the meaning of the documents in the corpus. We can further compare the similar documents by implementing the cosine similarity method.

### Python Implementation

The Sklearn package can be used to implement the LSA algorithm. In this implementation, LSA will be implemented with the TF-IDF values generated from the TF-IDF example. Using the TfidfVectorizer method; we can try on different values for the max\_feature parameter. The truncatedSVD function from Sklearn is used to implement LSA. The parameter n\_components takes the number of topics that we want to extract from our corpus.

### Ein Bild, das Text, Screenshot, Schrift, Dokument enthält. Automatisch generierte BeschreibungLatent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA) technique uses Dirichlet distribution, a kind of probability distribution. It was introduced in 2003 by Blei, Ng, and Jordan (2003) and is considered one of the most famous topic modelling algorithms. LDA is an unsupervised algorithm for extracting topics from a collection of documents. The LDA model supposes that the topics that make up a document in a corpus each have a probability distribution over words. When LDA is given a corpus composed of different documents, it backtracks to find out which topics have created these documents.

To better understand the logic behind LDA, we take a group of sentences to compose our corpus.

1. I want to buy a laptop and a cell phone.
2. I forgot my cell phone at home.
3. She was eating a banana for breakfast.
4. Orange juice is her favorite drink.
5. He was drinking apple juice while working on the computer.

LDA automatically discovers the topics that these sentences contain. The user needs to define the total number of topics to be discovered. The above sentences would be categorized as follows:

Sentences 1 and 2: 100% Topic A (Gadgets)

Sentences 3 and 4: 100% Topic B (Food)

Sentence 5: 70% Topic B (Food) and 30% Topic A (Gadgets)

Meanwhile, the words linked with each of the topics are also determined during the LDA algorithm. In this case the words related to Topic A are laptop, cell phone, computer; the words linked to Topic B are banana, orange, apple, juice, and drinking.

Ein Bild, das Text, Screenshot, Zahl, Diagramm enthält.

Automatisch generierte Beschreibung

Source. Alba Haveriku (2021)

The steps followed to create such a representation are as follows:

1. Choose a number K of topics to discover.
2. Review each document and assign each word to one of the topics randomly. This assignment gives topic representations and the distribution of words.
3. In each document, consider each word in the document. For each topic calculate
4. the words in this document that are assigned to this topic
5. the proportion of topic assignments in all documents, linked with word w
6. reassign a new topic to the word using the following probability

All topic assignments, despite the current word that is being checked, are considered correct.

1. Repeating the third step will provide better results after a specific number of tries, or until a specific threshold has been achieved.

Using the LDA algorithm, it is possible to extract the topics from a document in our corpus and associate each topic with the words that characterize it.

#### Ein Bild, das Text, Screenshot, Schrift, Rechteck enthält. Automatisch generierte Beschreibung

Source: Alba Haveriku (2021).

#### Example

To better understand each of the steps explained above, let’s look at a simple example. Consider four documents and a vocabulary composed of nine words, the words being represented in the table below.

**Example of Vocabulary**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Airplane | Laptop | Orange | Travel | Earphones | Train | Best | Drink | Good |

Source: Alba Haveriku (2021).

Step 1: Determine the three topics to be extracted from these documents.

Step 2: Assign the words randomly to topics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topic 1 | Airplane | Orange | Drink |  |
| Topic 2 | Laptop | Travel | Earphones | Best |
| Topic 3 | Train | Good |  |  |

**Topic-Word Relationship**

Source: Alba Haveriku (2021).

Step 3: Define a topic-term matrix and a document-topic matrix.

**Topic-Term Matrix**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Airplane | Laptop | Orange | Travel | Earphones | Train | Best | Drink | Good |
| Topic 1 | 40 |  | 2 | 10 |  | 3 |  |  | 4 |
| Topic 2 | 42 |  |  |  | 2 | 5 | 3 |  |  |
| Topic 3 | 10 |  |  | 4 |  |  |  | 5 |  |

Source: Alba Haveriku (2021).

**Document-Topic Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | T1 | T2 | T3 |
| D1 | 25 | 20 | 10 |
| D2 | … | … | … |
| D3 | … | …. | … |
| D4 | … | … | … |

Source: Alba Haveriku (2021)

Reassign a new topic to a word. Consider the word airplane. Suppose the number of words each document is as follows:

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

Automatisch generierte Beschreibung

The following formula is used to calculate the relationship between a topic and a document:

The following formula is used to calculate the significance of a specific word within a topic:

The two parameters α and β are chosen based on the final results that we want to achieve. The parameter is defined based on the mix of topics that are expected to be extracted from the document. Meanwhile, defines the control of the word’s distribution for each topic. These two parameters usually have a value less than 1.

The following three equations are used to determine the relationship between topic 1 and documents 1, 2, and 3, respectively.

The following three equations are used to determine the relationship between the word “airplane” and the three topics, respectively.

To reassign the word “airplane” to the appropriate topic, the two probabilities are multiplied and the final values are compared.

Finally, after the above calculations for the word “airplane” have been completed the probability that “airplane” belongs to topic 2 is the highest of the three probabilities; therefore, “airplane” is reassigned to topic 2.

These procedures included in the third step should be repeated for each of the words in the document.

Step 4: In the beginning, each of the topics was assigned randomly to the words that compose the vocabulary, now it is necessary to repeat the third step for a finite set of iterations, until the proper assignment is achieved for each of the words in the vocabulary.

### Python Implementation of Latent Dirichlet Allocation

The process for implementing the LDA algorithm is similar to that used for LSA, but the LatentDirichletAllocation must be imported from the Sklearn package.

from sklearn.decomposition import TruncatedSVD

from sklearn.decomposition import LatentDirichletAllocation

review\_1 = "Loved the sound, no battery issues"

review\_2 = "Sound quality is good; battery life not good"

vectorizer = TfidfVectorizer(use\_idf=True,

max\_features=5,

smooth\_idf=True)

model = vectorizer.fit\_transform([review\_1, review\_2])

lda\_model=LatentDirichletAllocation(n\_components=2,learning\_method='online',random\_state=42,max\_iter=1)

lda\_top=lda\_model.fit\_transform(model)

print("Review 1: ")

for i,topic in enumerate(lda\_top[0]):

print("Topic ",i,": ",topic\*100,"%")

### Self-Check Questions

1. Please list the steps followed for the Latent Dirichlet Allocation.

*Choose a number K of topics to discover.*

*Control each document and randomly assign each word to a topic.*

*Review each word in each document.*

*Calculate the words in the document assigned to each topic and the proportion of topic assignments in all documents linked with this word.*

*Reassign a new topic to the word.*

*Repeat the above steps until the desired assignment is achieved.*

1. Please mark the correct statements.

* *LSA is an unsupervised technique in natural language processing*.
* LSA supposes that the topics that make up a document each have a normal distribution over words.
* The two parameters α and β, used in the formula to calculate the significance of a word in the LDA algorithm, should be considered with a value larger than one.

1. Please calculate the proportion of words in a document that are assigned to T1. Let: , no\_topics=2, no\_words\_document=100 and the number of words linked to T1 in this document is 20.

**Summary**

Text mining techniques are used to understand the workflow from data pre-processing to finding meaning in textual information. Two common techniques used to transform text into vectorial representation are bag of words (BoW) and term frequency times inverse document frequency (TF-IDF). A logical explanation of the way BoW and TF-IDF, together with the main Python functions to implement these techniques give a proper view of one of the first steps followed in data mining.

Both bag of words and TF-IDF vector representations provide meaningful information to analyze and compare different documents in a corpus. Difficulties can arise when we want to check the similarity between synonyms, or better understand the topic of a specific document. To give a meaning to the whole document in disposal a thorough semantic analysis of the text is conducted. Different Python modules can be used to implement BoW, TF-IDF, and semantic analysis such as nltk.tokenize, pandas, and sklearn, with methods such as CountVectorizer(), TfidfVectorizer(), truncatedSVD(), and fit\_transform().

Semantic analysis describes the process of extracting meaning or specific topics from natural language texts. Latent Semantic Analysis and Latent Dirichlet Allocation are two of the main techniques used to extract topics from documents in a corpus. LDA is one of the most commonly used algorithms to identify the theme of a document, in recommendation systems, content generation, and other types of textual data. The application of a well-defined NLP pipeline can also be implemented in financial or business analytics processes and its application can prove beneficial when making financial predictions.

# Unit 3 – Web Analytics

**Study Goals**

On completion of this unit, you will be able to …

… define the main metrics linked with web analytics.

… understand clickstream analytics.

… review the best techniques for web data analytics.

… experiment with the metrics in Google Analytics.

… determine the main approaches of recommender systems.

# 3. Web Analytics

## Case Study

Marketo is a leading company in marketing automation software that supports large and small enterprises with their marketing campaigns (Google, 2015). Meanwhile, to do their own marketing, they needed a flexible platform to secure user engagement and reengagement. Marketo had two goals—to “improve customer engagement and generate more conversions” (Google, 2015, p. 1) and to “tie data obtained from Marketo’s real-time personalization product to an online remarketing campaign” (Google, 2015, p. 1). Their aim was to use the data collected from Marketo’s real time personalization product (RTP) in the process of retargeting marketing campaigns. Marketo engaged Google Analytics to provide a better overview of their communication channels (Google, 2015).

The approach chosen by Marketo was to utilize both their real-time personalization product (RTP) and services provided by Google Analytics to get the best experience possible. The website’s visitor’s information was passed from RTP to Google analytics in the form of an event, allowing Marketo to view customer’s data and visits number (Google, 2015). With the help of all these accurate data in hand, it was easy for the company to segment their market by visitors’ interest. Since Google analytics can be integrated with AdWords (Google Marketing Platform, n.d.), it was easier to provide users personalized remarketing ads.

The application of customized data resulted in increased user engagement for Marketo. The conversion rate was estimated to be ten times higher when compared to other traditional marketing campaigns (Google, 2015). The integration of Marketo’s RTP with Google Analytics provided the company the possibility to better segment markets, learn about user’s interests, understand the behavior of customers and therefore the overall decisions of the company were dramatically improved. The company now plans to further optimize current campaigns and create new remarketing strategies.

## 3.1 Web Metrics

The evolution of the Web has expanded the quantity and quality of information, providing the best possibility to gather useful data and generate insights for analysts and marketing experts. It is important to collect clickstream data, but it is of much higher importance to have people who can make sense of this data.

**Web analytics** includes the collection and analysis of data from your own website to determine success or failure and define future goals (Kaushik, 2010). This analysis is important for the improvement of user engagement and to achieve the company’s desired results. Since the reports that can be generated by measuring data (visits, page views, the time spent by a user in our website, etc.) are various, it is important for a company to be focused on the final outcome.

**Web analytics**

The collection and analysis of data from a website to determine future goals.

We can continuously count the number of visits in a website, but what is the outcome or the benefits the company gets from these visits? We need to be focused and understand that each metric considered should have an impact or a positive outcome for the company. Three of the most important types of outcomes that a website can generate for a company are: improved revenue, reduced costs, and customer satisfaction. No matter which area the business is operating in, whether it be technical, e-commerce, social media, or something else, focusing on these three outcomes can help you to better understand which metrics you need to be focused on. Google Analytics, Omniture, Clicktracks, and Xiti are some of the tools used by companies to have a view of what is happening in their websites (Kaushik, 2010).

A web metric is a term used for the types of measurement used to monitor the activity on a website and which help in the evaluation of online marketing campaigns (Kaushik, 2010). It is a measure which describes the main trends in a website. We will discuss six of the most important web metrics, that can be calculated and analyzed, to give positive outcomes to a company.

### Number of Visits

The number of persons that visit a website is one of the key metrics to be calculated, since it also serves as a basis to calculate other metrics. While talking about visits, it is important to distinguish between types of visits—visits by the same person or visits by unique visitors. A website can be visited several times by the same person, so the difference between these two measures is quite important to know, especially when the number of newcomers in a website is an important metric. Two common terms are important while studying these metrics—**sessions** and **cookies**.

**Sessions**

A unique session, with a unique ID, is created for each user that enters a website.

Sessionsare linked with visits to the website. They typically are linked with one visitor’s experience in our website. Let’s consider the usage of JavaScript tags in a website. In the moment a user enters the website, the analytics tool creates a unique session for that user. The user is linked with a unique session ID, and each of the actions they take in the website is associated with this unique session ID. Running reports for a given time period would give us a count of the total visits, by calculating the number of all sessions created during this specific period of time.

**Cookies**

Each new user is associated with a unique cookie, saved in the user’s web browser.

Cookieson the other side, are used to identify unique visitors to a website. When a user requests an item in a website, they are associated with a unique cookie, which is saved on the user’s web browser. Even when the user leaves the website, the unique cookies is still saved in the browser. This way, if the user later visits again the same website, they are recognized by the unique cookie previously stored in their browser. In this case, when we want to gather reports about the unique visitors in a given time frame, the number of unique cookie IDs represents the number of distinct visitors to the website.

Depending on the analytical tool used, the number of unique visitors can be measured at different time intervals, such as daily, weekly, or monthly.

All the screenshots in this unit are taken from the Google Analytics demo account which shows data from the Google Merchandise Store. This account is freely accessible and provides a good way to experiment with business data .(Google Analytics Help, n.d.).

### Ein Bild, das Text, Screenshot, Zahl, Software enthält. Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Google Analytics (2021g)

### Time Spent on The Website

The time a user spends on a website usually determines their engagement and their interest towards the products and services offered in it. Analytical tools help in measuring the time a user spends navigating the website (during a session), or the time a user typically spends viewing an individual page.

Each time a user makes a request to enter the home page of a website, an entry is created in the event log of this site showing the date and the time of the request. While the user clicks and enters new pages on the website, a new entry is created in each page access. This way, we can calculate the time spent on the first page as the difference between the times saved in the entry file:

T (time\_spent\_first\_page) = T(access\_second\_page)-T(access\_first\_page)

The same formula can be used to calculate the time spent in each of the pages that the user is going to access. The only problem stands in the inability of calculating the time spent in the last page of the website that the users visit, since there is no consecutive page visited next and the formula can’t be applied properly. The same problem arises when there is just a one page click, or a single-page session, a case where it is challenging to calculate the time that the user has spent on this only visited page.

In Google Analytics, it is possible to check the average session duration, which reports the average time a person has spent on the website (in other words the time spent on the website). Google Analytics calculates the duration of a session, from the moment the user enters the website, until they leave the website, or the user has not been active for 30 minutes (Google Analytics Help, n.d.). In cases where the user leaves the website open, the default settings determine that the session is ended after 30 minutes of inactivity.

Average\_session\_duration =

Measuring the average time spent on a page together with the average session duration, would be the best way to get a proper view of what is happening with the user that entered your website. A good average session duration might be about three minutes (Albright, 2021) and the techniques used to achieve it include: more engaging images or videos, a user-friendly interface, more engaging contents etc., depending on the type of website.

### Ein Bild, das Text, Zahl, Screenshot enthält. Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Google Analytics (2021f)

### Bounce Rate

Bounce rate is defined by the percentage of users that leave our website without doing any action (Kaushik, 2010). It is important to calculate this metric, since when the value of the bounce rate is high in your website, it may indicate that you have issues with the interface or the content of your website.

In Google Analytics, a bounce is linked with sessions that have only one single request to the server (Google Analytics Help, n.d.). The single page sessions divided by the total number of sessions define the boounce rate (Kaushik, 2010). These single page sessions have a time duration of zero seconds since the metric ‘Time spent on Website’ can not be calculated.

Bounce rate (%) =

The best value for the bounce rate depends on the type of website you are working with. If one of the the main focus points of your website is to have users that engage and visit more than one page of the website, than it is good to have a low level of bounce rate. In this case, if the value of the bounce rate is high, it means that the users are only visiting the home page and not viewing content on the other pages of the website, which is not good for us.

Meanwhile, in blogs or websites with a single main page, a high bounce rate can not be considered a bad outcome, since in these cases the bounce rate is not important.

### Ein Bild, das Text, Schrift, Reihe, Screenshot enthält. Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Google Analytics (2021c)

### Exit Rate

The exit rate metric calculates the number of people that have left a website from a certain page (Kaushik, 2010). It may show the point or the page of our website where the user lost interest and left. Even though, this metric can be thought of as the leakage point of our website, it doesn’t always show an appropriate result. There will always be a point from which the users will leave our website for different reasons: they will search again later, they found what they were looking for etc.

The question in this case is: How can we separate which users entered the website and never made an action and which users came, found what they wanted and left? To answer this question the exit rate metric should be calculated and evaluated together with other important metrics such as bounce rate and the conversion rate. The main difference between exit rate and bounce rate is that bounce rate is based on the sessions that start and end in a specific page. Meanwhile, exit rate is the percentage of pages that were the last page accessed in a session.

### Conversion Rate

To better undersand what conversion rate is about, we can describe it’s usage in an e-commerce website. The conversion rate is the Outcomes (orders made in ecommerce) divided by the number of visits/ unique visitors (Kaushik, 2010). We can choose visits or unique visitors depending on our final objective.

conversion rate =

The conversion rate metric is helpful to understand whether the audience of the website is keen on purchasing a product or making an order. We may choose Visits in the case when we suppose that a user places an order for each visit they make, and unique visitors when we suppose that a user may visit the website several times before making an order.

### Engagement

The engagement of users may be one of the most difficult metrics to interpret. The level in which each user engage in our website can be estimated by have a clear calculation of some other metrics, such as: time spent on site, subscribing to the newsletter, registering on the website (Kaushik, 2010).

The engagement is both a quantitative and qualitative value. The quantity of engagement can be in a range from a low involvement in the website, to a high involvement. Seen as a quantitative metric, it is easy to give a value to the user’s engagement. The user either visits the site and spend time searching and reading the content, or do not spent much time and is not engaged in the website. This can be easily measured by web analytics tools. The important thing to do is understand how the value measured may provide a positive outcome to the company.

What about the qualitative dimension? A user can have a positive or negative experience in a website. For example, if a user is searching for a specific article and they are not finding it due to the misorganization of the information in the website, they may seem engaged but are not happy at all. Seeing engagement as an emotional attachment of the user’s experience is quite difficult since no analytic tools can measure the thoughts and emotions of the users while navigating a website. Potential methods to give a qualitative measure to engagement are: conducting inline surveys to customers, use the data collected from market research or calculate customer retention in a long time interval (Kaushik, 2010).

### Ein Bild, das Text, Screenshot, Zahl, Schrift enthält. Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Google Analytics (2021d)

### Self-Check Questions

1. Please list 6 of the main metrics in Web Analytics.

*number of visits*

*time spent on the website*

*bounce rate*

*exit rate*

*conversion rate*

*engagement*

1. Please mark the correct statement.

* A session ID is linked with each unique visitor entering the website.
* *A cookie is used to identify unique visitors in a website.*
* A good average session duration is about eight minutes.

1. Please define the sentence below as True/False.

Typically, a low bounce rate indicates a possible issue with the website content or user interface True/*False*

## 3.2 Clickstream Analytics

**Clickstream**

The path each visitor follows through their visit in a website.

The term **clickstream** defines the path each visitor follows through a website. The analysis of clickstream data gives a perspective about the way a site is being navigated by the different users (Kaushik, 2010). Clickstream data include views, time spent on a page, page load time and other measurables that provide a better understanding of the way a website is being used. By conducting clickstream analysis, we can understand the strength and the weaknesses of a website and take appropriate measures to fix weak points or to develop new strategies for the future. Companies can benefit in areas such as: traffic analysis, marketing campaigns, segmentation of customers, and service experience.

Combined with information about different sessions, clickstream analysis can help in distinguishing the users who reached the website from different channels, may it be from social media, search engine etc. By analyzing and identifying the engagement of users, also knowing the source they came from, a webmaster can find which are the channels that provide the most interested users. This way, the company may focus on optimizing and delivering more resources on the channel that provide the most engaged users.

In the main web metrics mentioned before, exit rate was the one that gave a detailed information about the point in which the users lose interest. Knowing the exit page can help in improving targeted content so that to still keep the users engaged in their website. Web traffic analysis also provides important information related to the most popular keywords, product flows etc. (Uttamchandani, 2020). The data linked with clickstream can be used by: marketing specialists to understand customer experience, data analyst to segment users, experimenters to run different scenarios, data scientist to predict future improvements etc. (Uttamchandani, 2020).

### Clickstream Analytics Steps

There are four main steps that can be followed in each use case related to clickstream data.

1. Understand: As mentioned in the previous section, there are several reasons why clickstream analytics can be performed. The first step a company should undertake is deeply understand their need and the outcome they want to achieve after analyzing clickstream data, rather it be to review the traffic channels, improve content etc.
2. Collect: Knowing the final objective makes the second step easier, since we need to collect only the data that is linked with the final outcome that the company wants to achieve. One of the best analytical tools for clickstream data is Google Analytics. This tool offers a variety of free options such as the number of visitors, visitors interacting with your website etc. For each session, the relevant web metrics are shown and can be easily compared (Bush, 2020).
3. Analyze: In the analysis step, there is a need to compare and contrast the information collected in the second step. It is important to find patterns or exceptions in the way different users interact with a website. It is also important to make the best use of the large amount of information that can be collected in a web analytic case.
4. Improve: After analyzing the data collected and the patterns seen, some conclusions can be drawn so that to further improve the user experience in a website and increase user’s engagement.

### Common Methods for Performing Clickstream Analysis

Clickstream analysis includes different analytical strategies, depending on the type of business: B2B (business to business) or B2C (business to customer). One of the most important concepts that must be mentioned is segmentation. There are different types of users that may access a website for different purposes. This mandates a good segmentation strategy to offer an effective and useful experience to each type of user. Segmentation helps in focusing on the specific level, providing ideas that can be practically applied (Kaushik, 2010). Creating reports and graphical representation of the revenues based on the different sources is the best way to show management the effect of web analytics and the necessary steps to be considered in the future.

Ein Bild, das Text, Zahl, Schrift, Screenshot enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Google Analytics (2021e)

Other important analyses which can give a good overview of a website’s performance are SEO (search engine optimization), paid search, and email campaigns (Kaushik, 2010).

#### Internal site search

Similar to big search engines such as Google and Yahoo, the majority of websites offer an internal search, which provides the possibility to search for a given keyword in the whole website. By controlling what people search in the internal site search, it is possible to estimate the intent of the site’s visitors (Kaushik, 2010). Using analytical tools, such as Google Analytics, a percentage of visitors that have used the site search in given time intervals can be calculated and used to show searching trends. Related to this type of analysis, a report containing metrics such as: unique searches, search exits, result of search etc. provides information about the interests of users and the material they are looking for. Internal site search accuracy can be determined by controlling it together with bounce rate. Since the bounce rate calculates the number of people who leave the website in a specific page, linking it with internal site search provides information about which results of the internal search are working.

#### SEO

Search engine optimization is the process of improvement of a website so that to increase its visibility in search engines (Kaushik, 2010). SEO related data is found in both web analytics tools and external tools provided by search engines. People use search engines to find information related to a keyword or group of words, which determine the importance of optimizing website to show up in search engines for related keywords.

Some common measures to improve a website’s visibility are having a clean URL structure, provide relevant content in the website’s pages, utilizing the necessary keywords etc.

A website can get traffic from a variety of sources, where the most important to be mentioned are as follows (Mackenzie, 2019):

* Direct traffic: Traffic that is not coming from another referring site.
* Paid search traffic: Traffic that comes from search campaigns.
* Organic traffic: Traffic that comes from search engines (not paid for).
* Social traffic: Traffic that comes from social media platforms (Facebook, Instagram, LinkedIn etc.).
* Email Traffic: Traffic that comes from e-mail campaigns.

Ein Bild, das Screenshot, Text, Reihe enthält.

Automatisch generierte Beschreibung

Source: Haveriku,based on Google Analytics (2021e)

Google Analytics offers the possibility to segment the traffic into different sources, such as paid traffic, organic traffic etc. (Kaushik, 2010). By defining different time intervals, it is possible to understand how different sources contribute to the website’s visibility and see how paid campaigns are working for your website.

A company with a well-defined strategy related to search optimization, can determine its strong points and choose the best source which provides them more clients. The goal of a company’s search strategy is to have a proper indexed website in search engines (Kaushik, 2010). Popular search engines (Google, Yahoo etc.) offer the Webmaster Tools service, which provides statistics related to the progress of your SEO efforts. To use these Webmaster tools, it is important to sign up with the search engine and authenticate yourself by adding a HTML file in the root folders of the website you are owning (Kaushik, 2010). Each search engine categorizes pages in a different way, so it is important to check the Webmaster tools, so that to understand what is happening with your website. To check whether your SEO campaign is working or not, you need to control if the number of pages accessed by direct traffic have increased from the start of the SEO efforts.

After the analyzes of reports and the calculation of the number of visitors from different traffic sources, an analyst has to take measures that may seem trivial at first, but that will give a good outcome for the company. The analyst, after checking the traffic sources, have to define which of the traffic incomes is relevant and which not. Slowing down the traffic from not relevant sources, may seem counterproductive at first, but the measurement of outcomes such as Conversion Rates later on can show that in fact it has provided a larger percentage of people who successfully accomplish their goals.

#### Email campaigns

Email marketing, if it is done in the right manner, proves to be a good and cost-efficient acquisition channel (Kaushik, 2010). It is possible to create a targeted traffic and improve the awareness of an organization’s services and products. If an organization considers the usage of an email campaign, it needs to carefully track this campaign, so that to understand if it is converting into more customer’s engagement and generating outcomes.

Google Analytics, for example, provides a good way to track email campaigns and to check the traffic segmentation. The first step towards tracking is the creation of trackable URLs, which provides the best way to check which specific emails are bringing customers to your website and which emails are not working efficiently. It is necessary to add UTM (urchin tracking module) parameters in the end of an URL which are tags that save information about the person that clicked the link and entered a website (DeLane, n.d.).

An URL builder helps in adding these tags in the end of the URL and provides the possibility to create custom email campaigns. The important UTM parameters to create a trackable URL for further analyses in Google Analytics are as follows (DeLane, n.d.):

* UTM\_source: describes the source or referrer (e.g., a newsletter)
* UTM\_medium: defines the marketing activity (daily, weekly, monthly)
* UTM\_name: details or name of the campaign

Google Analytics Campaign URL Builder helps in the automatic creation of an URL with the specific UTM parameters (Google Analytics, 2021a). The main information that needs to be filled are as mentioned: the website main page URL, campaign source, campaign medium and campaign name. The URL is automatically generated and can be used to send trackable URLs in email campaigns.

An example URL generated can be:

<https://www.example.com?utm_source=newsletter&utm_medium=weekly&utm_campaign=autumn_sales>

where the information saved in the URL is:

* website URL: https://www.example.com
* campaign source: newsletter
* campaign medium: weekly
* campaign name: autumn sales

After sending these URLs in email campaigns, they can be tracked in Google Analytics. In the Google Analytics dashboard, in Acquisition, Campaigns the information provided by tracking the URLs build can be found (DeLane, n.d.).

To analyze what is happening with the users that accessed your website via the built URLs, the following key metrics can be used:

1. delivery rate. Checking if the email sent was successful at getting users to visit our website. Did the user engage in the website, or did they just visit the main page and left?
2. open rate. The total number of emails opened, even though usually images and scripts sometimes are blocked due to viruses’ concern.
3. click rate. This measure is key to understand how efficiently the email sent to the different users.
4. bounce rate. Shows the necessity of improvement to specific pages. An email campaign with 100% response rate and 99% bounce rate, is not effective at all, since the users don’t properly engage with your website.

### Self-Check Questions

1. Please list some of the main types of traffic on a website:

*direct traffic*

*paid search traffic*

*organic traffic*

*social traffic*

*email traffic*

1. Please mark the incorrect statements.

* *SEO related data is found only in web analytics tools and not on external tools offered by search engines.*
* UTM parameters create a trackable URL for further analysis in Google Analytics.
* *The click rate in an email campaign can be calculated as the number of opened emails divided by the total number of emails sent.*

1. Given an example URL as below, please find the value of the parameters: frequency and name of the campaign.

<https://www.audio.com?utm_source=google&utm_medium=daily&utm_campaign=spring_sales>

*frequency: daily*

*name of campaign: spring\_sales*

## 3.3 Recommender Systems

**Recommender system**

An automated procedure to offer the best service/product based on user’s preferences.

A **recommender system** is an information-filtering set of techniques which are used to present item of interests (books, images, video etc.) to a user (Negre, 2015). It seeks to predict the preference of users based on previous information known about that user. Each time that we are trying to purchase items on an e-commerce platform, a recommendation system is helping towards the best product that fits your interests. We may think of recommender systems as an automated procedure which substitutes salesmen and introduce the bests service/ product based on the user’s history and preferences.

Some of the most common recommender systems, that we use every day without even thinking are product recommendations on Amazon, videos recommended on YouTube, newsfeed on Facebook, music on Spotify etc. (Shetty, 2021). The data provided to recommender systems comes from explicit or implicit ratings (Shetty, 2021), where explicit ratings include: star ratings, likes, follows, reviews and implicit rating includes user’s interaction (clicks, views, purchase etc.).

The basis of recommender systems stands in understanding the following relationships between users and products:

* 1. User-User: Recommending based on mutual friends, similar age etc.
  2. User-Product: Recommending products that are needed based on user’s interests/jobs.
  3. Product-Product: Recommending similar music, books, videos etc.

Organizations utilize recommender systems for increasing sales and offering a personalized experience to their customers. Companies can retain or gain new clients by targeting customers with new offers that fit their interests.

### Main Approaches for Recommender Systems

The two main approaches used in recommender systems are: collaborative filtering and content-based (Negre, 2015). Content-based recommenders are based on the attributes of each of the items while collaborative filtering recommendation are based on the past activity of the user. Following each of these methods will be described in more details.

#### Content-based approach

A recommender system which is content-based, finds a match between an item and the profile of a user so that to determine the best recommendation (Negre, 2015). Each item is linked with a set of attributes which describe all their main features. For example, a book is linked with attributes such as: author, publisher, genre etc., while for a user we may take into account the age, the job, hobbies etc.

Content-based methods, based on the information (attributes of an item or features of a user), try to create a model that explain the interaction user-product. For example, we may try to model the behavior of teens who prefer certain music genres and of adults who usually prefer other kind of music genres.

In this approach, new users or new products can easily be categorized based on their characteristic, which does not interfere in the suggestions process. The only problem may be when new types of users or products, that have not been seen before, enter the system, making it difficult to find the perfect fit for them. Even in this case, the older the recommender systems get, the fewer problems arise in the designation of user-product interaction.

The main advantages and disadvantages of content-based recommender systems are described in the following table.

Advantages and Disadvantages of Content-Based Recommender Systems

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Recommendations are based on the user’s past activity. | Not all content can be described by keywords (videos, images). |
| It is not necessary to be able to access data related to other users. | There is a risk of over-specialization occurring. |
| Adding new users and products is simple. | Based on item and interests scores: the fewer the score, the more limited the recommendation. |
|  | There is no known history for new users in the system. |

Source: Alba Haveriku (2021)

Example: Consider a website where users can share their passion about reading

books. The users register in the website, giving personal interests and entering/ rating the book they have read. The website wants to recommend books to the users depending on the user’s preferences, defining a proper matching process.

**User-Product Interaction: Attributes of Different Books**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Genre** | **Author** | **Keywords** | **Number of pages** |
| 1984 | dystopian fiction | George Orwell | big brother, propaganda, censorship, technology | 328 |
| Anna Karenina | Realism | Leo Tolstoy | marriage,  society,  fidelity | 864 |
| Murder on the Orient Express | Crime | Agatha Christie | mystery,  murder,  detective | 256 |

Source: Alba Haveriku (2021)

**User-Product Interaction: Attributes of a User**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Titles** | **Genres** | **Authors** | **Keywords** | **Number of pages** |
| A study in Scarlet,  War and Peace,  Einstein: His life and Universe | realism, crime | Arthur Conan Doyle, Leo Tolstoy, Walter Isaacson | detectives,  society,  mystery | 300 |

Source: Alba Haveriku (2021)

**User-Product Interaction: Matches User-Book**

|  |  |  |  |
| --- | --- | --- | --- |
| **User’s interests** | **1984** | **Anna Karenina** | **Murder on the Orient Express** |
| Genre | No | Yes | Yes |
| No. pages | Yes | No | Yes |
| Keywords | No | Yes | Yes |

Source: Alba Haveriku (2021)

The three tables above show how we can derive the perfect books for a user, depending on the book’s attribute and the user preferences. From the third table, where the interaction user-product is better seen, we can derive that the best book to recommend for this user is *Murder on the Orient Express*.

#### Collaborative filtering approaches

Collaborative filtering recommender systems produce recommendations by checking the similarity between the preferences that the users of a platform have (Negre, 2015). In this approach, the recommended items are based on the opinions shared by other users with similar interests. The interactions between users and products are saved in a user-product interaction matrix (Rocca, 2019). The algorithms related to collaborative filtering are divided into two groups: memory based (work with past data without a model) and model based (explains user-product interaction based on a model).

In this approach the more users interact with product, the more effective the recommender system becomes. The problem in this approach stands in the interaction with new users or new products which is known as cold start or the lack of data to make good recommendations (Milankovich, 2015). Since in the beginning they have a small number of interactions it is difficult to make proper suggestions.

Example: We consider four users who have rated six movies. The users provide a rating for each movie: 1 if they liked the movie and 0 if they did not like the movie. A blank space in the table below means the user has not yet seen that movie. We are going to analyze two collaborative filtering approaches: item to item and user to item.

Collaborative Filtering: Movie Recommendation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Lord of the rings | Harry Potter | Back to the future | Star Wars | La La Land | Jurassic Park |
| Frank |  | 1 | 1 |  | 0 |  |
| Steve | 1 | 1 |  | 1 |  |  |
| Alice |  | 1 | 1 |  | 0 | 1 |
| James | 1 | 1 |  | 0 |  |  |

Source: Alba Haveriku (2021)

Having in disposal the data in the above table, for each approach we can determine:

Item to item: Frank and Alice like the movies: “Harry Potter” and “Back to the Future”. Since these two users like both of these movies, we may conclude that even James who likes “Harry Potter”, could like “Back to the Future”, so this movie is recommended to him.

User to user: Frank and Alice both like “Harry Potter” and “Back to the Future” and dislike “La La Land”. Since these two users seem to have similar interests, we can arrive in the conclusion that Frank would like the movie “Jurassic Park” since Alice likes it.

#### Other approaches

The following are some other approaches that can be used to create recommender systems.

* 1. Knowledge-based approaches: A type of recommender systems which is based on explicit knowledge gathered by combining products, the preferences of a user and recommendation criteria (Burke, 2002).
  2. Hybrid approaches: A system which uses the combination of different approaches: collaborative filtering and content-based (Burke, 2002).
  3. Graph-based approaches: A recommender system where data are represented in a graph form, where nodes can be users or items, and the edges are the interaction between users or items (Ricci et al., 2015).
  4. Context-Aware recommender systems: Systems that adapt to different contexts in which the user is found (location, nearby people etc.).

### Techniques to Implement Recommender Systems

In the context of recommender systems, different techniques and similarity calculations can be used, to provide the best recommendations in different cases. In this section, we are going to mention some of the most used methods, but the list of possible methods that can be used in different context is larger.

For the purpose of explaining the following methods, a matrix representing the linkage between users, books, and the genres with which they correspond will be taken as an example case.

**Vector Representation: Book Preferences**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Crime | Realism | Romance | Fiction |
| User 1 (u1): Anne | 1 | 0 | 1 | 0 |
| User 2 (u2): Ronald | 1 | 0 | 0 | 1 |
| Book 1 (b1): War and Peace | 0 | 1 | 1 | 0 |
| Book 2 (b2): A walk to remember | 0 | 0 | 1 | 0 |

Source: Alba Haveriku (2021)

#### Similarity methods

1. The Euclidian distance calculates the distance between two vector representations. For the above table it is calculated by using the following formula:

where n is estimated as the total number of terms (Negre, 2015).

Using the following Python code, we can automatically calculate the Euclidian distance between the two users u1 and u2.

from math import sqrt

u1 = [1, 0, 1, 0]

u2 = [1, 0, 0, 1]

e\_dist = sqrtsum((e1-e2\*\*2 for e1, e2 in zip(u1,u2)))

print(e\_dist)

The similarity between the two users is 1.414. Meanwhile, we can compare the Euclidian distance between u1/b1 and u2/b1 to see which is the most relevant user for the first book.

Euclidian\_distance\_u1\_b1=1.4

Euclidian\_distance\_u2\_b1=2.0

From the result achieved, the first book (b1) is more relevant to the first user (u1) than the second user (u2).

1. Cosine similarity evaluates the similarity between two documents (expressed in vector representation). It is calculated by the cosine of the angle between two vectors and checks if two vectors are pointing in the same direction (Han, Kamber, & Pei, 2012).

The formula to calculate the cosine similarity between two books is

Using the following Python code, we can automatically calculate the Euclidian distance between u1/b1 and u2/b1.

import numpy as np

from numpy import dot

from numpy.linalg import norm

u2 = np.array([1,0,0,1])

b1 = np.array([0,1,1,0])

result = dot(u2, b1)/(norm(u2)\*norm(b1))

print(result)

The results achieved are

Therefore, cosine similarity also shows that the first book (b1) is more relevant to the first user (u1) than the second user (u2).

#### Classification methods

The classification methods provide the possibility to group users or items by using a categorization system (Negre, 2015). The main purpose of classification is to determine a set of rules which will help in the categorization of items or users based on their characteristics. In the classification process, a training sample is necessary so that the model learns the classification rules and a test sample to validate the accuracy of the classification.

One of the most famous classification techniques is the k Nearest Neighbors method (KNN) (Negre, 2015). The KNN algorithm can be best described by the quote: “Show me who your friends are, and I’ll tell you who you are” in the famous work of Miguel de Cervantes, *Don Quixote* (de Cervantes Saavedra, 1755/2010).

The KNN algorithms is a supervised algorithm which works by

* determining the k nearest users/products to X (new user/new product) by calculating the distance between X (new user/product) and X’ (users/products from training set),
* counting the number of occurrences of each category for the nearest k user/products, and
* linking the new user X with the most frequent category.

### Self-Check Questions

1. Please list the main approaches used to build recommender systems:

*content-based*

*collaborative-filtering*

*knowledge-based*

*hybrid*

*graph-based*

*context-aware*

1. Please mark the correct statements.

* *Content-based recommender systems do not need to access data related to other users.*
* *A hybrid approach uses the combination of collaborative filtering and content-based approaches.*
* Collaborative filtering approaches do not need to have access to information about the user’s past activity.

1. Please define the sentence below as true or false.

The main problem in the collaborative filtering approaches is known as the cold start. *True*/False

Summary

In this chapter the main approaches and techniques linked with web analytics and recommender systems are introduced. Web analytics includes the process of collecting and analyzing data from a website and provide results which improve user’s engagement and determine future goals. A large number of web metrics can be constantly monitored to have a proper view of the activity in a website and the efficiency of different campaigns.

Six of the most important metrics which were described included: visitors, time spent on the website, bounce rate, exit rate, conversion rate and engagement of users. Each user that visits a website is linked with a unique session ID which is associated with the user’s activity, while the definition of unique visitors is made possible by the usage of unique cookies which are saved in each user’s web browser. The time spent by users in a website, is a key metric which usually determines their positive or negative experience in the website. The bounce rate calculates the percentage of users that leave a website without doing any action, and analyzed together with other metrices such as visitors and time spent provides the best possibility to understand user’s engagement in a website. Meanwhile, exit rate and conversion rate are other metrics that give a full view of the user’s experience. Other important analyses that can be made to evaluate the performance of a website are SEO, paid search and email campaigns.

Recommender systems provide and automated procedure to offer the best product or service to a user based on their preferences and interests. The basis of recommender systems stands in understanding the interaction user-user, user-product and product-product, to increase sales, target customers and gain new clients. The two main approaches used in recommender systems are content-based approaches which are based on the attributes of the items and collaborative-filtering which are based on the past activity of the users. Other approaches include knowledge based, hybrid approaches, graph based and context-aware approaches. The main techniques to implement recommender systems that were described are similarity methods (Euclidian distance and Cosine similarity) and classification methods such as KNN algorithm.

# Unit 4 – Social Network Mining

**Study Goals**

On completion of this unit, you will be able to …

… define what social media analytics is.

… understand the steps towards social media mining.

… review the best tools to mine common social media platforms.

… assess the importance of social media mining.

… implement Python code to mine Twitter and Facebook.

# Social Network Mining

## Case Study

The worldwide spread of SARS-CoV-2 since late 2019 has led researchers from all around the world conduct advanced analysis of the data generated daily so that to create the best tracing plans and minimize the spread of the virus. Some governments have taken in consideration the usage of mobile applications to identify possible infected contacts and limit the spread of the virus (Fraser Group, 2020). Since these kinds of applications need constant tracking of the user and may intrude on personal privacy, they may not be well-accepted by the general public.

On the other hand, almost half of the world uses social media platforms where users publicly share personal information on a daily basis (Statista, 2021). Furthermore, the user’s engagement activity has significantly increased during the pandemic (Cresswell, et al., 2020).

A study conducted with more than 107,990 tweets related to SARS-CoV-2 on Twitter, in the period December 2019 – March 2020 (Boon-Itt & Skunkan, 2020) showed that by mining social media it is possible to better understand COVID-19 trends and understand the public sentiment. By using analytics techniques such as word frequency, topic modeling and NLP approaches such as LDA algorithm, the most common topics and the categorization of tweets was possible in (Boon-Itt & Skunkan, 2020). The results achieved showed meaningful aspects related to the negative outlook of people towards COVID-19, the trend of COVID-19 symptoms and the automatic categorization of tweet’s collected was made possible. The three main groups into which the tweets were categorized are: pandemic emergency, how to control Covid, and Covid reports (Boon-Itt & Skunkan, 2020).

The final results achieved in this study showed that Twitter, as much as other social media platforms, can be beneficially used to understand public concerns related to COVID-19 but not only (Boon-Itt & Skunkan, 2020). The findings from the analysis of this public data, can help health officials to address the most common concerns and questions people are having. Also, the governments can benefit by analyzing how policies and restrictive measures are affecting people in their country.

## 4.1 Introduction to Social Media Analytics

The usage of social media platforms is increasing rapidly each year. More and more people are sharing personal opinions and information across different platforms. Social media can be defined as a group of applications that are built based on the foundation of Web 2.0, allowing the exchange of content generated by each user (Bekmamedova & Shanks, n.d.). The user’s engagement and communication in different social media is the reason why a large number of organizations are focusing on online marketing, and the promotion of their products and services in their respective social platforms. Having a proper online presence has become a must for each organization, so that to increase reputability and broaden their current audience.

Facebook stands as the leader in the market, with more than 2.98 billion active users (Statista, 2021). Since the company now owns other large social media platforms such as WhatsApp and Instagram, the approximated number of total users for these three platforms (Facebook, WhatsApp, Instagram) in the first quarter of 2021 was over 3.5 billion monthly active users (Statista, 2021).

In the graphic below, you can find a representation of the average number of monthly active users worldwide for some of the most famous social networks (Statista, 2021).

Ein Bild, das Text, Screenshot, parallel, Zahl enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku based on Statista (2021)

**Social media analytics** is related to the conversion all of the data generated on social media platforms into meaningful business insights (Khan, 2015). This analytical process can be thought as related to both science and creativity, since it involves the extraction and analysis of information with sophisticated techniques, and the interpretation of these analytics towards reaching the goals of an organization (Khan, 2015).

**Social Media Analytics**

The conversion of data generated on social media platforms into business insights.

Social media analytics (SMA) continuously uses the content generated in social media platforms to provide a real time analysis of how trends are created and shared in social media. SMA can supply organizations with information related to customer’s opinions and the sentiment linked with specific products.

The development of data mining applications is directly linked with the level of data accessibility, data which is commonly not provided to everyone. In the case of social media platforms companies can easily access information, since most of the platforms provide an API (Application Programming Interface), which provides a structured access to the data. The availability of the information depends on the way users share their information (public or friends network).

One of the main challenges that comes with “social media analytics” is the large amount of information to be processed, or what is commonly known as “Big Data,” which requires new efficient methods and technologies to properly access and process the available data. In 2001 the term “Big Data” was associated with the three V’s: volume, velocity, and variety, by the analyst Doug Laney (Laney, 2001).

The data generated in common social media platform can be seen in the following three aspects:

* volume: The large number of social media users and their activity in these networks has created a large amount of data on social platforms.
* velocity: Is related to the speed in which the data is generated. For example, only Twitter users are estimated to generate 500 million tweets in a day (Internet Live Stats, n.d.). Mining tools should offer the possibility to analyze data while it is generated, so that to create real-time reports.
* variety: The data generated in social media platforms comes in an unstructured format. It can either be textual, video files, image etc.

But, with large amounts of data comes great responsibility. Since the Cambridge Analytica scandal (Confessore, 2018) there has been a lot of controversy, dividing people in two groups: avid social media users (who argue that people choose what they share) and people who think there is a larger need to defend user’s data privacy. The European Union has passed the GDPR (General Data Protection Regulation), and it is being applied since 25th May 2018 in all the states that are members of EU (General Data Protection Regulation- GDPR, n.d.). This regulation has expanded privacy rights and enabled stricter rules towards companies that hold users’ personal data.

### The Difference Between Social Media and Business Analytics

The use of social media analytics for businesses and business analytics itself are closely related to each other, but yet there exist some distinguishment between them. While SMA is focused in the interpretation of data collected from social media platforms to be able to accomplish the company’s main goals and strategies, business analytics is related with the interpretation of the organizational data which is being periodically stored in the organizations management systems. So, one of the contrasts that can be mentioned, is the fact that SMA usually is a set of techniques and tools that analyses and interprets ever changing data in real time and business analytics is more focused on historical data which a company has in disposal.

While the traditional forms of media consist in the flow of information from the publisher to the user (consumer), social media platforms have provided a new model where the user is both a publisher and consumer. Social media platforms are based on the concept of ‘networking’, or the interaction between users and their posts. The importance or value of social data is usually determined by the dimensions of the network in which it is shared. The more an information is shared the greater its value becomes (Khan, 2015). The value of information can be calculated in terms of the number of views of a post/ page or the number of likes/ followers etc.

Meanwhile, in technical aspects, what can be mentioned as the most important difference between social media analytics and business analytics is the form of the data in disposal. In business analytics, we usually work with well-structured data, which are stored in organizational databases, or data that come from other sources in a pre-known format. On the other side, in social media analytics the data collected comes in different forms (text, emoticons, images, videos, hyperlinks etc.). This leads to the necessity of cleaning processes, which can transform the data in the best format to be saved and processed.

Another change that is important to mention is the fact that business analytics is tightly linked with private data that is controlled by an organization and is stored in the organization’s databases. This leads to more bureaucratic procedures and less flexibility in the processing of this data.

A general overview of the differences between these two disciplines is given in the table below.

Social Media Analytics and Business Analytics

|  |  |
| --- | --- |
| **Social Media Analytics** | **Business Analytics** |
| The data is collected from social media platforms. | The data is collected from the organizations management systems. |
| Both historical and real-time data management is possible. | Usually only historical data management is possible. |
| The data comes in structured and unstructured format. | The data is mostly in a structured format. |
| The data is mostly accessible by everyone. | The data is usually private and controlled by the company. |
| The value of the data is linked with the proportion of sharing. | The value of the data is only for the company who owns it. |

Source: Alba Haveriku (2021)

### Social Media Analytics Steps

Before starting the social media analytics journey, it is important for an organization to answer the following questions, so that to create a proper view of the procedure and final results that need to be achieved.

1. Which steps need to be followed during the social media mining process in our organization?
2. How to access the right information from social media platforms?
3. Which are the main data mining techniques and tools that can be used?
4. How can the organization profit from this process?

The SMA procedure includes 5 main steps, starting from the identification of the information that should be collected to the generation of meaningful insights and visualization. These steps are as follows:

1. Identification: The first step includes the identification of the most relevant information. Having an exact plan of what you want to achieve in the end, helps in the identification of the main sources of information. Data linked with an organization’s social media platforms (official Facebook, Twitter, or other social media account) is the most important one to understand the user’s engagement. It is important to do an alignment between the business objectives of your organization and what you want to benefit from social media analytics, rather it be brand awareness, increasing profits, targeted marketing campaigns etc.
2. Data collection:After having a proper thought of the necessary data to be collected to achieve the expected results, the collection process can begin. Depending on the type of data that is going to be collected (text, image, video) different tools can be used to extract it. In most cases, API (application programming interface) calls are used to extract data from social media platforms. APIs can be seen as a set of protocols that social media platforms have previously set up to access specific parts of information. Despite, programming or developing yourself the analytical process, each social platform has an analytical tool linked with it. Twitter for example uses Twitter Analytics, Facebook with the Insights tab provides detailed information linked with a Facebook Page, YouTube offers YouTube analytics etc. (Bonzanini, 2016).
3. Data cleaning and pre-processing:The data collected should be cleaned to remove unwanted information. In the case of textual data, natural language processing techniques may be used to remove unimportant information from the extracted text. Data cleaning and pre-processing are dependent on the social platform that we are working with. It is linked with the removal of unimportant words, removal of punctuation etc.
4. Data analysis:The pre-processed data can be analyzed in this phase to generate possible insights. In this chapter, we will mostly be focused on text mining or text analytics, by analyzing the posts and comments made by users in a social platform. Using Python libraries and modules it is rather easy to perform hashtag analysis in Twitter, finding trending topics in Facebook, find friends connections etc.
5. Presenting the results and interpretation: After having analyzed the collected data, it may be presented in forms that are more easily understood by general audiences. Visualizations can be used to effectively transmit and explain the generated insights to interested parties. This visualization may come in the form of graphs, word clouds, charts, infographics, plots etc. The final step of the whole social media analytics process would be to interpret the results achieved and use them to improve future business decisions.

Ein Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Bozanini (2016).

The main Python libraries that will be used in each step of SMA are

* NumPy (Numerical Python): The possibility to process data structures in an array like form,
* Scikit-learn: A machine learning library in Python which supports different classification algorithms: support vector machines, k-means etc.,
* Pandas: A Python library for data analyses and manipulation, and
* NLTK: The main Python library used to process textual information.

Meanwhile, Jupyter Notebook can be used as an open-source web app to code live and directly create visualizations in graphical or textual form (Jupyter, n.d.).

### Self-Check Questions

1. Please list the 5 main steps included in Social Media Analytics:

*identification*

*data collection*

*data cleaning and pre-processing*

*data analysis*

*presenting the results and interpretation*

1. Please mark the correct statement.

* *Business analytic consists of analyzing the private data of a company.*
* The largest proportion of the data generated on social media is not publicly accessible.
* The data saved in an organization’s management systems is not well-structured.

1. Please define the sentence below as true or false.

The data generated on social media platforms comes mostly in an unstructured format. *True*/False

## 4.2 Mining Common Social Media Platforms

In this section, we will properly investigate the techniques and tools that can be used to mine some of the most common social media platforms. Specifically, two platforms will be taken in consideration: Twitter and Facebook. For each of them, the creation of a development environment with Python and the proper analytical approaches to be followed, will be thoroughly explained.

### Mining Twitter

Twitter is one of the most used social networking services nowadays. It was firstly introduced in 2006, reached more than 100 million users by 2012 (Twitter Inc., 2012) and now it counts more than 330 million active users (Statista, 2021). This platform offers what is called a microblogging service, a form of blogging, where the content is shorter (Bonzanini, 2016), specifically in Twitter a limit of 140 characters per tweet was allowed, and from 2017 this limit was expanded to 280 characters (Boot, Sang, Dijkstra, & Zwaan, 2019). The network topology in Twitter is one directional, which means that a connection may not be mutual, you can follow someone who doesn’t follow you back. The Twitter API includes all the necessary tools to engage with or analyze everything that is happening on Twitter.

The majority of tweets is linked with personal experiences, but tweets may also be associated with different events happening all around the world. Information related to political events, heavy traffic, fires, earthquakes or other natural/ human disaster. One of the strongest points in the nature of this information in the Twitter network is the real-time sharing and the geolocation information linked with tweets. For example, in 2011, when the strong tsunami wave hit Japan, Twitter was one of the main communication tools used, since the traditional communication media where impossible to be used (Wallop, 2011).

The main topics of interest linked with Twitter mining are as follows:

* Twitter terminology and the structure of a tweet
* interacting with Twitter API by using Tweepy
* entities such as hashtags, URLs, mentioning of users
* histograms of the data collected in Twitter

#### Twitter terminology

To better understand the use of the Twitter’s API, we need to have a clear view of the main concepts and terminology linked with Twitter. In Twitter users post tweets, which appear in their timelines. A tweet, as we mentioned has a limit of 280 characters, and may contain hashtags, user mentions, URLs etc. Even though a tweet may seem simply a short string, it is linked with two pieces of metadata: entities and places (Rusell & Klassen , 2019). Entities include hashtags, URLs and other media that may be linked with a specific tweet, while places may show the location from where a tweet was generated.

All Twitter APIs returning tweets provide the encoding of data using JSON (JavaScript Object Notation). A tweet considered as a JSON object is expressed as key-value pairs.

If we consider a simple tweet format with some of the main attributes it can be represented in the following JSON format (Twitter Developer Platform, n.d.-a):

{

"created\_at": "Thu Apr 06 15:24:15 +0000 2017",

"id\_str": "850006245121695744",

"text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps:\/\/t.co\/XweGngmxlP",

"user": {

"id": 2244994945,

"name": "Twitter Dev",

"screen\_name": "TwitterDev",

"location": "Internet",

"url": "https:\/\/dev.twitter.com\/",

"description": "Your official source for Twitter Platform news, updates & events. Need technical help? Visit https:\/\/twittercommunity.com\/ \u2328\ufe0f #TapIntoTwitter"

},

"place": {

},

"entities": {

"hashtags": [

],

"urls": [

{

"url": "https:\/\/t.co\/XweGngmxlP",

"unwound": {

"url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",

"title": "Building the Future of the Twitter API Platform"

}

}

],

"user\_mentions": [

]

}

}

In the above representation we may notice some main characteristics which are found in each tweet such as: the author, the text message, an ID; user related data, retweet count, date of creation etc. Despite these main characteristics a tweet itself can contain more than 150 attributes: coordinates, favorites, place, retweet\_count etc. (Twitter Developer Platform, n.d.-a).

The Structure of a Tweet

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| created\_at | String | UTC time when the tweet was created |
| id | Int64 | Unique identifier of a tweet. |
| text | String | UTF text of the post |
| in\_reply\_to\_status\_id | Int64 | If the tweet is a reply, this attribute contains the ID of the original tweet. |
| in\_reply\_to\_user\_id\_str | String | If the tweet is a reply, this attribute contains the ID of the user. |
| User | User object | This field contains information related to the attributes of the user ( id, location, friends\_count, description et.). |
| Coordinates | Coordinates | Geographic location where the tweet was posted. |
| Retweet\_count | Int | The number of times the tweet was retweeted. |

Source: Alba Haveriku (2021)

#### Types of Twitter APIs

An API is a group of definitions for the integration of application software, and for making easier the interaction between you and the computer/ system that you are trying to access/ perform an instruction (Red Hat, 2020). Access to Twitter’s data can be achieved by using a series of APIs. Two of the main categories are: REST API and Streaming API.

REST stands for Representational State Transfer, a term defined in 2000 by Roy Fielding (Gupta, 2021). The REST API is based on REST architectural style and allows the interaction with RESTful web services (RedHat, 2020). This API is beneficial when we want to access previously published tweets, collect data from a user’s timeline, generate new tweets etc.

All the connections request an authentication procedure, which in the Twitter’s case is Open Authorization (OAuth). OAuth is an open authorization framework for REST APIs. It provides applications the possibility to gain limited access to a user’s data without entering the user’s password. Using this kind of authentication, the user’s password is never exchanged with third parties, guaranteeing the trust of the users (Raible, 2017).

The Streaming API provides a completely different way of accessing Twitter data. When making a connection, a data stream is opened, and as new tweets are generated, they enter the connection and are read by the application. This type of API has limited capacity depending on the amount of information that is expected to be extracted. The three main types of Streaming API are as follows (Wang, Callan, & Zheng, 2015):

1. Public streams: Streams that contain public tweets
2. User streams: Streams related to the tweets of a specific user
3. Site streams: Streams related to more than one user

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

The REST API helps us collect old tweets, while the Streaming API looks at the future and opens a stream which collects specific upcoming tweets. Depending on the type of analysis that we want to conduct we can choose either of these two types of APIs.

#### Rate Limits

Twitter has a rate limit for the number of requests a specific application can make to the API accessed in a defined time frame. These limits are defined so that to provide a reliable API, since thousands of developers make requests to the Twitter API (Twitter Developer Platform, n.d.-d). Rate limits depend on the type of authentication method that is being used, and are set based on the access token given. To become familiar with the concept of rate limit the official documentation in the Twitter developer account has detailed information related to each of the Twitter APIs (Twitter Developer Platform, n.d.-b). The only problem when hitting the API limits is that an error message is shown and if multiple requests are made, the regular access may be revoked, flagging you as a potential abuser (Bonzanini, 2016). A Python module named time(), can be used to specify a delay between two requests, to avoid potential passing of the allowed rate limit.

For example, the commands below, create a 20 second delay between two different API requests:

request\_1()

time.sleep(20)

request\_2()

#### Interacting with the Twitter APIs

The first step towards interacting with a Twitter API, is creating an account on Twitter and registering an application. An application linked with your account can be created on the log-in page (Twitter, n.d.), where each developer should apply for a Twitter developer account (Russell & Klassen, 2019). By creating an application, a set of authentication tokens (Access Token and Access Token Secret) will be created, to later programmatically access the platform (Russell & Klassen, 2019).

Ein Bild, das Text, Webseite, Website, Screenshot enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Twitter Developer Platform (n.d.-c.)

In the tab “Keys and access tokens,” you can find the consumer key, consumer secret, access token and access token secret necessary for the authorization process (Bonzanini, 2016). The consumer key and secret represent the application itself, while the access tokens represent the developer’s account. In the ‘Permissions’ tab the access level can be set, determining the level of interaction (read-only; read and write; read, write and direct messages), where read-only is the safest option since it allows only the collection of the tweets without editing options.

Now, an authenticated connection can be made. In the example code below, you can substitute your own credentials and then create a simple API call.

import tweepy

API\_KEY = ''

API\_SECRET = ''

ACCESS\_TOKEN = ''

ACCESS\_TOKEN\_SECRET = ''

a = tweepy.OAuthHandler(API\_KEY,API\_SECRET)

a.set\_access\_token(ACCESS\_TOKEN, ACCESS\_TOKEN\_SECRET)

API = tweepy.API(a)

home\_tweets = API.home\_timeline()

for x in home\_tweets:

print(x.text)

The tweepy library was imported to handle the interaction with the Twitter API. By using the OAuthHandler and set\_access\_token classes and passing the respective tokens the authentication process is complete. Meanwhile, the API class gives access to the REST API methods and the home\_timeline() method allows the collection of the most recent tweets from the accounts we follow. After executing this example, an instance of the Twitter API is created, by using the OAuth credentials of the user and the most recent tweets from the people you follow will be printed on the screen.

#### Collecting data

After the authentication process, the gathering of data for different purposes is possible. As was mentioned, depending on the final result we want to achieve we may either use the REST or Streaming API.

If we want to search for a trending topic, such as the hashtag #covid, we can create a query to search and gather information related to this hashtag in a csv file.

import tweepy

import csv

API\_KEY = ''

API\_SECRET = ''

ACCESS\_TOKEN = ''

ACCESS\_TOKEN\_SECRET = ''

a = tweepy.OAuthHandler(API\_KEY,API\_SECRET)

a.set\_access\_token(ACCESS\_TOKEN, ACCESS\_TOKEN\_SECRET)

API = tweepy.API(a)

topic="#covid"

tweets = tweepy.Cursor(API.search, q = topic, lang='en').items(20)

data=[[tweet.id\_str,tweet.created\_at,tweet.retweet\_count, tweet.text.encode("utf-8")] for tweet in tweets]

with open('%s\_tweets.csv' % topic, 'w') as f:

writer = csv.writer(f)

writer.writerow(["Id","Created\_at","Retweet\_count","Text"])

writer.writerows(data)

The method tweepy.Cursor is used to go through the first 20 items in our home timeline. The object tweets, which was created, is iterable, which means we can select each of the tweets collected easily in the next lines of code. For each tweet, the attributes saved in the csv file are: id, created\_at, retweet\_count and tweet’s text.

Another option is to collect the same tweets and store them in a file in a JSON format, which also provides a good format for next analytical processes.

import tweepy

import json

API\_KEY = ''

API\_SECRET = ''

ACCESS\_TOKEN = ''

ACCESS\_TOKEN\_SECRET = ''

a = tweepy.OAuthHandler(API\_KEY,API\_SECRET)

a.set\_access\_token(ACCESS\_TOKEN, ACCESS\_TOKEN\_SECRET)

API = tweepy.API(a)

with open('home.jsonl', 'w') as file:

for page in Cursor(API.home\_timeline,

count=200).pages(2):

for status in page:

file.write(json.dumps(status.\_json)+"\n")

After executing this example code, a file home.jsonl will be created. The extension .jsonl follows a JSON Lines file, in which a JSON document is saved in each line of the file.

This code in the end will create a file called timeline.jsonl in which the information about tweets will be saved. A file with the extension .jsonl is a JSON Lines files, where each line is a JSON document (JSON Lines, 2021). In this case, we saved the tweets from our home timeline, but if we want to collect tweets from a user timeline there is just a simple change needed:

with open('user\_t.jsonl', 'w') as file:

for page in Cursor(API.user\_timeline,

screen\_name='nytimes',

count=200).pages(2):

for status in page:

file.write(json.dumps(status.\_json)+"\n")

The above block of code saves 400 tweets from the user with screen\_name “nytimes”.

#### Entity analysis

Despite simply collecting, printing or storing tweets related to different search queries, we may go deeper and start with some simple statistical manipulations. The module collectionsfrom Python can be used to count the occurrences of a term and compute a frequency distribution (Russell & Klassen, 2019). Analyzing hashtags for example would provide a good information related to the most famous topics found in a user’s timeline.

import sys

import collections

import json

with open('user\_timeline.jsonl', 'r') as file:

hashtags = Counter()

for i in file:

tweet = json.loads(i)

entities = tweet.get('entities', {})

h = entities.get('hashtags', [])

hashtags\_tweet = get\_hashtags(tweet)

[tag['text'].lower()for tag in h]

hashtags.update(hashtags\_tweet)

for t,x in hashtags.most\_common(10):

print("{}: {}".format(t, x))

In the above code, we read line by line the .json file created before (in which we saved the tweets collected from a user timeline). The function collections.Counter calculates the count of strings, where strings ( in this case hashtags) are the keys in a dictionary and the value linked with them is their respective frequency.

The tweet itself is saved in a dictionary, and all the entities of this tweet are saved in a variable called entities. Since we are particularly interested in the hashtags mentioned, from these entities using the get() function we can access the hashtags from the entity. The hashtags are all transformed in lower case with the function lower(). The method most\_common() is used to order the keys from the largest to the smallest one, providing a list with ten of the most mentioned hashtags.

If we execute the code above for tweets collected for Barack Obama, we may get the following output:

obamaleaders: 3

wearorange: 2

getcovered: 2

collegesigningday: 2

marchmadness: 1

otd: 1

inauguration2021: 1

mlkday: 1

apromisedland: 1

voteearlyday: 1

Similar to the hashtag analysis, we can collect information related to the other users mentioned in a user\_timeline. Below there is a demonstration of the collection of user mentions in a specific user timeline:

import sys

from collections import Counter

import json

with open('user\_timeline.jsonl', 'r') as file:

mention = Counter()

for line in file:

tweet = json.loads(line)

entities = tweet.get('entities', {})

h = entities.get('user\_mentions', [])

users\_in\_tweet = [tag['screen\_name'].lower() for tag in h]

mention.update(users\_in\_tweet)

for t,x in mention.most\_common(5):

print("{}: {}".format(t, x))

The code above would print 5 of the most mentioned users in Barack Obama’s Twitter profile recently:

obamafoundation: 36

joebiden: 36

michelleobama: 28

barackobama: 15

kamalaharris: 11

#### Tweet’s text analysis

Previously, we analyzed the entities linked with a tweet. Another important step is the processing of the unstructured data that we collect from Twitter. We are going to have a look at the NLTK (Natural Language Toolkit) library, with its TweetTokenizer class to tokenize, remove stop words and punctuation. Differently than stop words in simple text analytics (articles, adverbs etc.), in this case we also have to remove common terms found in tweets, such as RT (retweet) and via (when mentioning the original author of a post).

from nltk.tokenize import TweetTokenizer

from nltk.corpus import stopwords

from collections import Counter

import sys

import string

import json

def preprocess(text, tokenizer=TweetTokenizer(),stopwords=[]):

text = text.lower()

tokens = tokenizer.tokenize(text)

return [t for t in tokens if t not in stopwords and not t.isdigit()]

file='user\_timeline.jsonl'

tokenizer = TweetTokenizer()

punctuation = list(string.punctuation)

stopwords = stopwords.words('english') + punct + ['rt','via', '...']

tf = Counter()

with open(file, 'r') as file:

for i in file:

tweet = json.loads(i)

tokens = preprocess(text=tweet['text'],tokenizer=tokenizer, stopwords=stopwords)

tf.update(tokens)

for t, c in tf.most\_common(10):

print("{}: {}".format(t, c))

The function process()determines the preprocessing phase by calling the functions lower(), tokenize() and isdigit(). In the stopwords group, despite the common stop words in English language, the ‘rt’ and ‘via’ are included since we are processing tweets. Depending on the list of words we want to exclude from the tweets we are collecting, we can expand the stopword\_list.

import matplotlib.pyplot as plt

y = [count for t, c in tf.most\_common(20)]

x = range(1, len(y)+1)

plt.bar(x, y)

plt.ylabel("Frequency")

plt.savefig('distribution.png')

Ein Bild, das Text, Screenshot, Zahl, Schrift enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

The matplotlib library is used to visualize the frequency of the terms in a graphical way. If we have a better look at the curve created in the above chart, it is an approximation of the law of power. The distribution of the frequencies of terms end in a long tail, which shows that only a few terms are dominating the distribution. This is also known as the Pareto principle, which defines that 80% of the term occurrences come from 20% of the unique terms (Glen, n.d.). Later Zipf’s Law was made popular by George Zipf, an American linguist. He stated that in a given group of documents, the most frequent word will be found twice as often as the second one, and so on (Hosch, 2018).

### Mining Facebook

Facebook is the most used social media platforms today, where more than half of its users are daily active by posting, chatting, commenting or reacting to posts . Facebook has one of the largest pools of information that can be accessed through its API, but it comes with a detailed set of privacy controls (Russell & Klassen, 2019).

Even though Facebook is defined as a social graph (a graph where people are connected by their relationships), it has slowly transformed into an interest graph. Interest graphs are a representation of all the things a person is interested into (Nayal, 2011), so they are based on an individual interests rather than an individual friends’ network. Facebook and Twitter can be considered both a social and interest graph since they fulfill both these definitions (Russell & Klassen, 2019).

The main topics that will be considered during Facebook mining would be:

1. The Facebook Graph API
2. The open graph protocols
3. The Graph API Explorer

#### The Facebook Graph API

To extract data from Facebook, similarly with Twitter we need to register an application in the Facebook developer platform, specifying the level of access that we want to give to the user of the application.

**Facebook Graph API**

A HTTP based API, used to manipulate data, post text, photo, stories etc. in Facebook.

There are two ways to access this platform, by developing an application or with the help of the Graph API Explorer (Russell & Klassen, 2019). The **Facebook Graph API** is a HTTP based API, which can be used to manipulate data, post photos, stories etc. (Facebook for Developers, n.d.-b). This API’s name is based on the fact that Facebook is largely known as a social graph. Facebook itself is a graph data structure, which is composed of nodes (peoples or users) and the connection between these nodes (friendships) (Russell & Klassen, 2019).

When creating an application, we can generate an access token, which will request permission to access the user’s Facebook data (such as name, profile picture etc.). After giving the permission the user is logged into the app with their Facebook data. The access tokens (OAuth tokens, as mentioned in the Twitter’s case) provide a secure way to access the user’s data without requiring their password.

Nodes in the Graph API are unique objects with a unique ID (Facebook for Developers, n.d.-b). Some of the nodes in the Facebook graph are: user nodes (each unique individual logged in the network), groups, posts etc. Each node is linked with fields such as id, name, description etc. A special node is the */me* node which determines the current person who is making the API calls. The result in this API, as you can see in the figure below, is given in a JSON format.

Ein Bild, das Screenshot, Text, Software enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Facebook for Developer.

The initial query, which shows just the name and the id of the user may be modified to include other fields such as friends’ connections, like connections etc. However, there exists some restriction on these fields, and your app needs to go under review by Facebook, to check whether the data is used safely by the app, following all the privacy guidelines ([Facebook](https://developers.facebook.com/docs/permissions/reference) for Developers, n.d.-d). The code below programmatically explores and prints in a JSON format the same information that we could automatically get from the Graph API.

import json

FACEBOOK\_TEMP\_TOKEN=" "

graph=facebook.GraphAPI(FACEBOOK\_TEMP\_TOKEN)

profile=graph.get\_object('me',fields='id,name,email')

print(json.dumps(profile,indent=4))

#### The Open Graph Protocol

The Open Graph protocol (OGP) is a protocol created by Facebook in April 2010, to integrate web pages into the Facebook social graph (Russell & Klassen, 2019). This protocol, from the name itself, determines a way of opening the social graph, by including lots of new objects (web pages). To better understand this protocol let’s take an example. Consider a web page where in the end you see the option: *Like on Facebook*. Facebook incorporates RDFa metadata in this page, so that to include these pages in the Facebook’s activity stream (Russell & Klassen, 2019). RDFa (Resource Description Framework in Attributes) is a HTML5 extensions, which provides the possibility to add attribute related extensions to HTML/ XML documents (RDFa, n.d.). This facilitates searching and provide a better connection.

Facebook has continuously provided tools to enhance searching and connectivity. Two projects that may be mentioned are: the Facebook Graph Search, which dates in 2013, and Facebook Local in November 2017 (Russell & Klassen, 2019). The Facebook Graph Search used natural language processing to make search queries, such as entering in the searching bar the sentence “Friends who live in Italy and ride a motorbike” would provide relevant information to what your connections are interested into. This tool has been dropped off, favoriting the searched based on keywords (Russell & Klassen, 2019). As for the Facebook Local, it is a mobile application which gives you information about local places that are famous or well-appreciated by your friends group. Both these tools are based on the Facebook social graph and the connection between different nodes in this graph (Russell & Klassen, 2019).

#### Using the Graph API Explorer tool

The Graph API Explorer can be commonly used to easily generate access tokens, test API queries, get examples of codes linked with the queries we have executed etc.

Some of the main components of this tool are:

1. Facebook App dropdown: A dropdown menu where you can choose the application you want to test.
2. User or Page: A menu where you can choose between application, user or page tokens for the application you are working with.
3. Add a permission: A menu where you can add permissions to the current application. The default permission is public\_profile. Meanwhile, three other groups of permission exists: User Data Permissions (email,photos etc.\_, Event Groups Pages (page\_read\_engagement, page\_events etc.) and Other (Instagram\_basic, publish\_video etc.
4. Access Token: After choosing the right options and permissions the access token can be generated and shown in this field.

Ein Bild, das Text, Screenshot, Zahl, Software enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021), based on Facebook for Developers (2021a).

1. The Query field: Here we can define the different queries that we want to execute.

Ein Bild, das Text, Schrift, Reihe, Screenshot enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on Facebook for Developers (2021a)

1. Get Code/ Debug Information: Here we can get the code which is automatically created after we execute a query in the query field. The Debug Information option can be used when one of our queries is failing. With the generated debug information, we can later contact the Developer Support to better understand where the problem lies (Facebook for Developers, n.d.-d).

### Self-Check Questions

1. Please list the three main types of Twitter Streaming APIs.

*public streams: streams that contain public tweets*

*user streams: streams related to the tweets of a specific user*

*site streams: streams related to more than one user*

1. Please mark the correct statement.

* The Pareto principle defines that 20% of the term occurrences come from 80% of the unique terms in a document corpus.
* *The REST API is beneficial when we want to access previously published tweets.*
* The attribute in\_reply\_to\_status\_id contains the ID of the user who originally create the tweet.

1. Please define the sentence below as true or false.

Facebook and Twitter can be considered as both social and interest graphs. *True*/False

Summary

In this chapter the usage and importance of social media analytics was introduced. Nowadays, social media platforms are a vital part of a large number of people, be it for connecting with friends or for business purposes (marketing, product sharing etc.).

Social media analytics is linked with the conversion of data into meaningful insights, to generate profit or gather important information. Social media platforms are a good way to gather publicly shared data, which is easily accessed through the APIs used by each social platform. The social media analytic process includes five main steps which are: identification, data collection, data cleaning and pre-processing, data analysis and the presentation of the results. When taking in consideration the mining process of social media platforms, two are the main platforms that usually are taken in consideration: Twitter and Facebook. Twitter is mostly taken in consideration because of the high accessibility through different REST or Streaming APIs, and the detailed documentation. Meanwhile, Facebook is one of the most used platforms with the largest user’s engagement, so the audience is larger than other platforms.

When trying to access tweets, all the connections need to go through an authentication procedure, which in this case is OAuth. Each tweet may include entities such as hashtags, URLs and other media (photo, video). The tweepy library is used to handle the interaction with the Twitter API.

Facebook on the other side, provides the Facebook Graph API which can be used to manipulate data, post photos, stories etc. (Russell & Klassen, 2019). The Graph API Explorer provides some basic options to generate access tokens, add permission to the application, execute queries etc. Facebook, by using the Open Graph protocol (OGP) is able to integrate different web pages into their own social graph, with the incorporation of RDFa metadata (Russell & Klassen, 2019).

# Unit 5 – Testing and Experimentation

**Study Goals**

On completion of this unit, you will be able to …

… determine the importance of testing and experimentation.

… review methods such as A/B, multivariate, and multi-armed bandit testing.

… understand the advantages and disadvantages of each testing method.

… define the best testing method to use in different case scenarios.

… experiment with multi-armed bandit testing.

# 5. Testing and Experimentation

## Case Study

Lumosity is an application composed of more than 50 games, which was developed to improve users’ memory and enhance their problem-solving skills. Users pay a monthly subscription to access the games daily, via either the web or a mobile application, and track their everyday journey (Siroker & Koomen, 2013). Lumosity now has more than 85 million members (McGowan, 2020) showing its popularity among users of all ages. A study involving 4,715 participants was conducted to analyze the improvement of the participants’ cognitive performance after playing either Lumosity games or crossword puzzles. After 10 weeks of training, the Lumosity group showed better improvement in cognitive assessments (Hardy et al., 2015).

The main metric measured for each Lumosity user is regular training, with the best option being to train for 15 to 20 minutes per day, four to five times per week (Siroker & Koomen, 2013). Despite this being the recommended usage, it was not mandatory for the participants to do so. The data collected showed that users initially played for hours, but over time their login sessions reduced drastically. The team thought that the best solution to keep their customers engaged daily would be to limit the length of time for which they could play each day. The risk was, of course, that users may be frustrated by the restricted use of an application that they had paid for.

The company decided to try this model as part of an A/B test. Two models where provided—the original model, which allowed unlimited training and the new model, which limited the amount of time the application could be used per day (Siroker & Koomen, 2013). These models were offered to different users. Over time, the results showed that the second model (with limited training) was more successful, with users training more over time. Testing two models provided the best option to check which was the most successful option. After the success of this first initiative with A/B testing, the company used further testing methods to find the best options, which provided users with the best experience while using their application (Siroker & Koomen, 2013).

## 5.1 Practical A/B Testing

People run experiments in almost every scientific field, from biology to physics, chemistry and beyond. Experimentation can also be seen in social fields such as marketing, where marketers experiment by providing different services to different audiences. Testing and experimenting in the online world was, at first, considered something optional and associated with large and well-structured organizations. Today, online experimentation has become a must for almost every company, from small to large-scale enterprises. Some of the most well-known tools available that assist with experimentation are VWO, Optimizely, A/B tasty, and Oracle Maxymiser. (VWO, n.d.).

**A/B testing**

A simple experimentation method to determine the best version of a page on a website.

The simplest type of experimenting is known as **A/B testing** (also commonly referred to as split testing). The A stands for the initial testing variable and the B refers to the new version of the initial variable. A simple example illustrates this concept. A company wishes to determine which variant of a page on their website will bring in the larger number of new customers. Two groups of people are shown the two different variations and insights can be drawn from the results (Siroker & Koomen, 2013). A/B testing can be seen as an experiment, the results of which determine the best option for achieving the main goals of a website. The impact of the A and B variables can be measured in terms of the web metrics, i.e., the successful sign-up rate achieved by each variant, and the web metrics can then be used to decide whether A or B should be implemented permanently.

A/B testing can only produce meaningful results if the goal is clear. The perfect time to try A/B testing is in the moment one of our website metrics (such as the conversion rate) is not at the level it should be. By experimenting with images, buttons, or other design features, it is possible to check which version is providing the best value for the chosen metric. It is important to make a proper plan before starting to experiment with the website components.

Ein Bild, das Text, Screenshot, Schrift, Rechteck enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

### The A/B Testing Process

Four main steps can be followed to conduct a proper A/B testing process in a website (Siroker & Koomen, 2013).

#### Step 1: Understand the purpose of the website

Before starting the experimentation phase, we need to identify the main purpose of our website and the elements that could be improved. The purpose of the website and the possible optimization processes need to be turned into quantifiable metrics. We need to define the metrics that are going to be improved by performing different tests, in order to make it easier to check the final results and determine the best variant of our website. For example, fundraising websites define their success by calculating the average number of donations for each visitor, Google’s engineers measure the number of users that leave the search page without clicking anything (also called abandonment) (Siroker & Koomen, 2013).

To sum it up, the first step in the testing strategy is defining what is and what is not a conversion for our website. A conversion is the point at which the users take the desired action on the website (Siroker & Koomen, 2013). By identifying the main focus points in a website, we know where to start creating impactful tests.

#### Step 2: Collect data and identify possible bottlenecks

In addition to identifying the points at which conversion occurs, it is also important to identify the main points of the website where the customers or users are lost (possible bottlenecks).

Barack Obama’s presidential campaign was strongly associated with data analysis and experimentation (Siroker & Koomen, 2013) and exemplifies the potential benefits of removing bottlenecks. In November 2007, one of the main bottlenecks of the campaign’s website was convincing visitors to the website to sign up for the email list. The email signup step could be optimized to generate a higher rate of email subscription. In 2008, the optimized page generated an additional 2.8 million subscriptions (Siroker & Koomen, 2013). Further experimentation in the Obama campaign example showed that testing methods assist with making objective decisions. For example, despite the majority of the team thinking that having a video on the campaign’s landing page was the best idea, testing showed that the Obama’s family photo provided the most clicks (Siroker & Koomen, 2013).

#### Step 3: Define a hypothesis

Having identified the success metrics and possible bottlenecks in our website, it is important to formulate hypotheses that identify possible ways to achieve the final goal of the website. Before starting to run an experiment, it is important to form a hypothesis, which can subsequently be tested to determine whether we have guessed the correct path towards achieving the predefined success metrics. Even though a test can fail and give negative results, the results can still provide valuable information about the incorrect approach and may lead to the formulation of new hypotheses.

#### Step 4: Start testing

After determining the possible hypotheses, we need to start experimenting based on the importance of the experiments. Visitors are selected randomly and offered variations of the website. Their user experience is tracked, to check which variation is linked with better results of the defined success metrics. The results of one test may reveal that further experimentation is required, creating a cycle which eventually leads to the best possible version of the website.

### Common Test Statistics Used in A/B Testing

After running an A/B test, the statistical significance of the test results should be checked. Statistical significance is a tool that can provide a measurement of the uncertainty in the collected data. A null hypothesis is defined, which assumes that there is no relationship between the data (Frost, 2020). Statistical significance measures the likelihood of an observation assuming that the null hypothesis is true. Statistical tests define the degree of difference between the tested variables and the null hypothesis (Frost, 2020).

To better link these statistical concepts with A/B testing, we can consider an example. After running an A/B test, one of the following statements can be true:

* There is a true improvement in the new variant we have created.
* The test results show an improvement, but it is just a rare outcome.

A p-value (probability value) is commonly used to quantify the statistical significance of the result. The probability of getting a result, assuming the null hypothesis is true, depends on the p-value. The lower the p-value is, the lower the probability of achieving this result is. The result achieved would be assumed to be statistically significant if it allowed the rejection of the null hypothesis. For example, if a result has a p-value of 0.04, or a 96% significance, there is a 1/25 chance that the observed result has occurred due to a natural variation.

It is important to have a good understanding of probability distributions before analyzing statistical tests in more detail. Depending on the type of metrics that we need to analyze, different types of distributions can be used to model the user’s behavior in A/B testing. Some of the most used commonly used distributions are binomial distribution, normal distribution (gaussian distribution), T-distribution, and Poisson distribution (Frost, 2020).

Binomial distribution is used where there are only two possible outcomes. It is a discrete distribution where the variables considered are the probability of success (p), the total number of trials (n), and the total number of successes (x). The formula to calculate binomial distribution is (Thomopoulos, 2017)

Statistical significance tests that can be used when we have a binomial distribution of the data are Fisher’s exact test and Barnard’s test. These tests can be applied in example cases with discrete metrics such as click-through rate, conversion rate, and bounce rate.

Normal distribution is one of the most widely used distributions, because the data are distributed normally in a wide range of real-world cases. It is also commonly known as Gaussian distribution and is a continuous distribution. The formula to calculate the normal distribution is (Thomopoulos, 2017)

where μ is the mean and σ is the standard deviation.

The statistical significance tests that can be used when studying a dataset with a normal distribution are Welch’s t-test and Student’s t-test. These tests can be applied in example cases with discrete metrics such as average revenue per user and average session duration.

Other statistical tests that are important to mention are Pearson’s chi-squared test, the Z- test, and the Mann-Whitney U test.

### Pros and Cons of A/B testing

A/B testing is one of the simplest and cheapest experimentation techniques, in part because it only involves comparing two variations. It is the best solution to use when a company is just starting their testing efforts. A/B testing does not usually require extensive resources and the results achieved are relatively easy to explain. It is a good way to find possible weaknesses in the marketing strategy of an institution or even simply to confirm that an existing strategy is working effectively to achieve the desired goal.

In contrast, it is difficult to explain the results when multiple variations are tested. We can easily identify the winning test page, but it is impossible to understand which elements caused that variation to be successful.

In the graphic below, two versions of the same page are shown, the only difference being the color and content of the buttons. Even though the change may seem small, it can make a difference to the click rate. The picture on the left shows that the original version of the website had a click rate of 52 percent; while the new version of the website has a higher click rate of 72 percent.

### Ein Bild, das Text, Screenshot, Website, Webseite enthält. Automatisch generierte Beschreibung

Source: Maxime Lorant (2015). CC-BY-SA-4.0

### Content That Can Be A/B Tested

Each content component on a website needs to be optimized, in order to achieve good conversion rates. The most important components are as follows:

* headlines. As headlines are one of the first things that the users of a website read, it is important to determine the best combination of words (which is not too short and not too long) and to test different text sizes and fonts.
* content. Visitors fall into two main groups—those who like to read long pieces of text and those who prefer reading short summaries about a topic. Using A/B testing, we may create one variation with long content and another with shorter content. The results of A/B testing will show which version is favored by the users, and thus identifies the version that achieves the best metrics, such as a higher conversion rate and a smaller bounce rate.
* design and layout. The design and layout of a page should provide clear information and avoid confusing the visitors. The content should be simple and well organized. As the home page is the first page to be visited, it needs to be optimized perfectly and should be the first page which undergoes A/B testing. Different pictures, videos, and text styles can be tested to find the variation that provides the highest conversion rate.
* navigation bars. Navigation should be structured and simple to understand, so that users do not have problems finding what they are looking for.
* reviews. Including reviews, recommendations, and testimonials on a website may provide direct benefits to the website’s owner. A/B testing can determine whether including reviews in the website’s content will increase the conversion rate for the specific website owner.

### Self-Check Questions

1. Please list the four main steps that should be followed to conduct an A/B test.

*understand the purpose of the website*

*collect data and identify possible bottlenecks*

*define a hypothesis*

*start testing*

1. Please mark the correct statements.

* *A/B testing is one of the cheapest experimentation techniques.*
* *A/B testing is a good way to find weaknesses in the marketing strategy of an institution.*
* A failed test does not provide any important information to the tester.

1. Please define the sentence below as true or false.

The results of A/B testing can identify the specific component on a page that provides a higher conversion rate. True/*False*

**Multivariate testing**

A form of A/B testing where several combinations of elements in a page are evaluated.

## 5.2 Multivariate Tests

**Multivariate testing** (MVT) is a form of A/B testing whereby several components are modified at the same time and the performance of each combination of the modified components is evaluated (Kaushik, 2010). The main goal is to determine which of the created variations performs best. Both A/B testing and MVT can be used in different fields. Although these testing techniques are most commonly used to evaluate users’ engagement online, they can also be used offline to evaluate how users influence each other and to what extent products meet the users’ needs.

Data about each of the possible combinations are collected and used to determine which combination of elements performs better. This testing technique can be used in various situations, including when we want to make radical changes to a website and thus changes to more than one component are likely to be made.

The following example illustrates how multivariate testing can be used. Let’s suppose that we have a website for our online business. The home page of this website has three important elements: images, buttons, and headlines. Three versions of each of the three elements will be tested to determine which combination performs best. The formula used to determine the total number of variations for the MVT is

Number of variations = No. of versions of first element \* No. of versions of second elements \* No. of versions of third element

Number of variations = 3 \* 3 \* 3 = 27 combinations

Applying the above formula, there are 27 different combinations that need to be tested. Each of the combinations will be tested to find the one that provides a better conversion rate.

The graphic below shows another example where two components are being tested: a headline and a button. Two versions of the headline and two versions of the button are considered, which produce four different combinations in total. Each variation of the page is linked with its conversion rate, estimating version three to be the winner, on the basis that it produced the highest conversion rate.

Ein Bild, das Text, Screenshot, Reihe, Diagramm enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

### Multivariate Testing Techniques

Two of the most commonly used multivariate tests are full factorial testing and fractional factorial testing (Kaushik, 2010), each of which utilizes a different approach to directing traffic.

* Full factorial testing distributes the incoming website traffic equally among the different testing combinations.
* Fractional factorial testing (also known as partial factorial testing) experiments with a smaller number of combinations and tries to guess what is going to happen with the other combinations.

Both techniques have their advantages and disadvantages. On the one hand, full factorial testing provides detailed information about each possible combination, but the time required to test a large number of combinations would be considerable. On the other hand, fractional or fractional factorial testing loses some data about the combinations but provides faster results. Fractional or fractional factorial testing also requires less traffic, and therefore is a good option for testing for websites with low traffic volumes.

### Advantages and Disadvantages of Multivariate Tests

MVT is a simple method whereby the data collected directly provide a better understanding of the preferences of the majority of users, which can help determine the optimal variation to offer to the users. The main advantages of MVT are as follows (Kaushik, 2010):

#### Advantage one: The ability to test multiple components within a page

Let’s suppose we want to trial some changes in our website and want to start testing, in particular to compare the use of an image versus a video on the home page and to test two variations of a button—“Try now” or “Buy now.” If A/B testing were to be used, two consecutive tests would be required; the first test compares the conversion rates for the page with an image and the page with the video, resulting in either the image or the video being chosen as the more effective element and the same is done for the second test, with the more effective button being identified.

Using MVT instead of A/B testing provides the possibility to test all possible combinations at once and to find out which combination works best. MVT offers the possibility to test the best combination of various elements on a page. In other words, MVT saves time by avoiding the need to conduct multiple, consecutive A/B tests.

#### Advantage two: The ability to identify elements that do not contribute positively

In addition to identifying the best combination of elements in a page, MVT also spots the elements that do not contribute to the website’s conversion rate. Inappropriate textual content, images, videos and similar may be relocated within the page or may be removed entirely, in order to attract and focus the attention of possible visitors.

#### Advantage three: Determine the structure of a page

By testing different types of elements in different positions, MVT can help make decisions about the structure of the page. The first elements that a visitor sees usually determine their willingness to further explore the website; therefore, it is important to display the most relevant components first to attract their attention. MVT testing is the correct technique to use when there is uncertainty about the structure of the home page or landing page.

The MVT also has a number of disadvantages, the most significant of which are discussed below (VWO, n.d.).

#### Disadvantage one: Requires a higher traffic

MVT should generally be used for sites that have high traffic volume; only in this case is it possible to generate meaningful results. If MVT needs to be applied to a site with low traffic, a compromise may be reached by testing a smaller number of combinations.

#### Disadvantage two: MVT is more complex than A/B testing

MVT, as mentioned previously, requires an extensive effort on behalf of the organization. It is a complex procedure, which needs to be conducted by experienced people who know what they are doing. A/B testing is easily implemented, even by non-technical persons; in contrast, MVT requires experts in the field to achieve good results. MVT requires significant effort particularly from the IT department, the marketing department, managers, and other key players. The experiments in MVT provide a useful result only if the customers’ preferences have been properly considered.

#### Disadvantage three: Time cost

A/B tests can be set up quickly and provide fast answers. Meanwhile, because MVT is more complex, it is slower to set up and to run. Each organization needs to determine whether the time required to run multivariate tests is acceptable and whether the testing will provide profitable results.

#### Disadvantage four: High risk of failure

A/B testing may result in success or failure, but in either case, new knowledge about the website components is gained. In comparison, the costs (including the time cost and the cost of other resources) of MVT are prohibitively high. Therefore, it is risky to use this model, unless a well-thought-out plan is in place.

### Self-Check Questions

1. Please name the two main multivariate testing techniques.

*Full factorial testing*

*Fractional or factorial testing (also known as partial factorial testing)*

1. Please mark the correct statements.

* Multivariate testing does not require a page to have a high volume of traffic.
* *Multivariate testing is more complex than A/B testing.*
* *Fractional factorial testing provides faster results than full factorial testing.*
* Multivariate testing is faster than A/B testing.

1. Please fill in the blank space.

The home page of a website has three important elements. If we decide to test two versions of these elements, the number of possible variations in a multivariate test would be *eight* combinations.

## 5.3 Multi-Armed Bandit Testing

**Multi-armed bandit testing**

A form of experimentation, which dynamically allocates traffic to the best performing version of the page.

In traditional testing methodologies, such as A/B testing and MVT, the user traffic is evenly split between the different versions of the website. **Multi-armed bandit testing** is a new form of experimentation, which allows traffic to be dynamically allocated, based on which version is performing better, leaving underperforming versions with less traffic.

### Exploration and Exploitation

The terms exploration and exploitation are the two essential concepts directly linked with multi-armed bandit algorithms. To better understand these concepts, think of a scientist who continuously tries to experiment (exploration) and a business owner who wants to maximize their profits (exploitation).

Classical A/B testing uses exploration to determine results that are statistically significant. In A/B tests, the conversion rate for each variant is calculated and the best performing version is declared the winner (Gupta, 2021). In contrast, multi-armed bandit testing works as an explore-exploit trade-off. A multi-armed bandit algorithm explores and dynamically allocates the majority of traffic to the version with a higher likelihood of winning (exploit).

The figure below shows the difference between A/B testing and multi-armed bandit algorithms, including what happens throughout time and how the best version is chosen. We can see that, in the case of A/B testing, the traffic is equally divided between the two different versions, while in multi-armed bandit testing, the traffic changes depending on the success rate of each variation being tested.

Ein Bild, das Text, Screenshot, Schrift, Diagramm enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku (2021)

Whether A/B testing or multi-armed bandit testing is the appropriate choice in any given situation will depend on the scenario. A/B testing is the perfect choice when an organization’s objective is to obtain the statistical results for each variation of the website and to apply the resulting knowledge to the development of other products. Conversely, multi-armed bandit algorithms are the best fit when the organization does not have time to gather statistical findings but wants to optimize a metric in a short timeframe (White, 2012).

**Greedy algorithm**

An algorithm which chooses the best action to take in any given moment, without considering future events.

Bandit algorithms slowly reduce the necessary exploration time and focus the time and resources towards the best performing version of the website (White, 2012). To describe this process, it can be said that a good multi-armed bandit algorithm will converge at some point.

### The Epsilon-Greedy Algorithm

This section analyzes a simple algorithm that can determine the exploration-exploitation trade off. A **greedy algorithm** is an algorithm that chooses the best action to take in a specific moment, without thinking about the consequences (White, 2012). The epsilon-greedy algorithm is an almost greedy algorithm; it randomly switches between the two options: explore and exploit. The logic behind it is quite simple—we can think of flipping a coin every time we need to make a decision between the two available options.

Suppose a business wants to make some changes to their website, but it is unclear whether the changes are going to be well accepted by the users. The first issue is whether image A or image B has the higher conversion rate and thus should appear as a banner on the home page.

The steps followed by the epsilon-greedy algorithm are as follows (White, 2012):

* A new user visits the website. The epsilon-greedy algorithm randomly chooses between the two options: exploration and exploitation. We consider that the algorithm determines the decision depending on a coin flip. Suppose the coin lands on tails with an epsilon probability.
* If the coin lands on heads, exploitation will occur. The algorithm checks the previously recorded information to find out which image has the highest conversion rate thus far. After identifying this image, it displays the image for the current visitor to the website.
* However, if the coin lands on tails, exploration will happen, meaning that the choice between images A and B will be made randomly. In this case, the probability for both images is 50 percent. If the coin comes lands on tails, image A will be shown to the visitor, if it lands on heads, image B will be shown.

Ein Bild, das Text, Screenshot, Schrift, Diagramm enthält.

Automatisch generierte Beschreibung

Source: Alba Haveriku, based on White (2012)

This algorithm should run for a relatively a long period of time, in order to direct traffic to the apparently better version of the home page.

The main elements linked with the epsilon greedy algorithm are as follows:

* The algorithm exploits with a probability 1-epsilon.
* The algorithm explores the best option with a probability epsilon/2.
* The algorithm explores the other possible version with a probability epsilon/2.

#### Main concepts: one-armed bandit and reward

The name of the multi-armed bandit algorithms is based on slot machines in a casino (White, 2012)., which are called one-armed bandits, due to the possibility of earning a lot of money with just one lever. The multi-armed bandit method is something like choosing the slot machine which has a higher chance of winning, which therefore gave the name to this algorithm.

In multi-armed bandit algorithms, the term “reward” is used to determine the successful achievement of something, whether it be a click, a new customer sign up, or some other positive result. It needs to be something that can be calculated mathematically, so that the reward from different versions can be calculated and compared to find the best fitting version.

#### Python implementation

The process and code required to implement the epsilon-greedy algorithm in Python are explained below.

The main values important to specify in the implementation are the (White, 2012)

* epsilon—the frequency with which one of the options will be chosen (explore/exploit),
* count—an array in which we save the number of times each version (arm) has been chosen. In an example with two arms, where the first one has been chosen zero times and the second has been chosen two times, the count array would appear as [0,2], and
* values—an array in which the number of rewards linked with a given arm are saved.

These values will be incorporated into the definition of a class. The values (epsilon, count, and values) will be used as parameters to create objects of the epsilonGreedy() class (White, 2012).

class EpsilonGreedy():

def \_\_init\_\_ (self, epsilon, count,values):

self.epsilon=epsilon

self.count=count

self.values=values

return

The above code represents the definition of an epsilon greedy class. Next, the main actions that can be performed must be defined. These actions are the selection of an arm and the update of previously known information. The two methods to be implemented are (White, 2012)

* define\_arm. Each time that we enter exploration mode, the define\_arm method (without arguments) will be called to return the index of the arm chosen.
* update\_information. The arm pulled returns a reward to the system. The update function takes three arguments: an object, a numeric index of the arm chosen previously, and the reward of this arm. This function updates the previous information and make the necessary changes for the next actions (White, 2012).

def ind\_max(x):

m = max(x)

return x.index(m)

def select\_arm(self):

if random.random() > self.epsilon:

return ind\_max(self.values)

else:

return random.randrange(len(self.values))

In the above code a condition is first checked. The randomly generated number is compared to the value of epsilon. If the random number is greater than epsilon, the largest value in the values array defines the arm that will be selected. Conversely, if the value of the randomly generated number is smaller than epsilon, the arm is chosen randomly. So, in this part of the code, the algorithm choses to either explore or exploit.

The update function specified below defines the way in which the best arm is chosen. The first step is to increment the value of count (which records the number of times an arm has been visited). If this is the first time visiting the arm, the value of the reward is considered. If the arm has been visited before, the new value for that arm is calculated using the following formula (White, 2012):

new\_value = ((n - 1) / float(n)) \* value + (1 / float(n)) \* reward

The value of a chosen arm is dependent on the weighted average of the previous values and the reward received in that moment (White, 2012).

def update(self, chosen\_arm, reward):

self.counts[chosen\_arm] = self.counts[chosen\_arm]+1

n = self.counts[chosen\_arm]

value = self.values[chosen\_arm]

new\_value=((n-1)/float(n))\*value+(1/float(n))\* reward

self.values[chosen\_arm] = new\_value

return

The epsilon-greedy algorithm also has some weaknesses. For example, even once it is known that one component has a lower conversion rate, some time will be invested in exploring this option, which is also known as over-exploring. The epsilon-greedy algorithm is one of the simplest algorithms to study and implement. There are other multi-armed bandit algorithms which provide improvements to the epsilon-greedy algorithm, such as the Softmax algorithm (or Boltzmann Exploration) and the UCB (Upper Confidence Bound Algorithm)., which provides a more efficient method to identify the best variation of a page than the epsilon-greedy algorithm. However, the epsilon-greedy algorithm is the simplest algorithm to study and implement and it forms the basis for other important algorithms, which can be applied in this field.

### Self-Check Questions

1. Please define exploration and exploitation in the context of multi-armed bandit algorithms.

*Exploration means continuously experimenting to find the best version of a page in a website.*

*Exploitation means choosing the version which maximizes the profits of an organization.*

1. Please mark the correct statement.

* *A multi-armed bandit algorithm dynamically allocates traffic.*
* A/B testing is the best option to use when a metric needs to be optimized in a short space of time.
* A good multi-armed bandit algorithm never converges.

1. Please indicate whether the following statement is true or false.

Bandit algorithms slowly reduce exploration time and focus the resources on one version of the website.

*True*/False

Summary

Experimentation and testing is one of the most important steps for both small and large scale businesses, as it enables the business to gain better insight into customers’ perceptions of the website. Testing with different components provides valuable information and can improve metrics such as the bounce rate and conversion rate.

The simplest method is A/B testing, where two or more variations of a website can be tested among different users. The final results of A/B testing identify which variation performed better, leading to further testing decisions. To conduct A/B testing it is essential to follow some important steps, starting with understanding the purpose of the website, identifying possible bottlenecks, defining a hypothesis, and starting testing. Each component on a website, including headlines, content, design, navigation bars, and reviews, can be optimized.

Multivariate testing is a special for of A/B testing where several components are changed in a website, and a number of combinations of these components is evaluated to see which of them performs better. Data is collected for each of these combinations, determining how each component affect the website’s performance. There are two main multivariate testing techniques: full factorial and fractional factorial. Multivariate testing is more complex than A/B testing and usually requires websites with a high traffic.

Multi-armed bandit testing is a new form of experimentation, which allows the dynamic allocation of traffic to the versions that are having a better performance. In this case the versions that are not performing well receive a lower traffic. The main terms linked with multi-armed bandit algorithms are exploration (continuous experimentation) and exploitation (finding the component which maximizes profits). A/B testing and multi-armed bandit testing can be used depending on the scenario, where the A/B test usually are used when statistical results are needed and multi-armed test are used when we want to optimize metrics in a short timeframe. The epsilon-greedy algorithm is a simple algorithm using mathematical formulas to determine the trade-off between exploration and exploitation.