MASTER MEIM 2021-2022

BIG DATA ANALYTICS

Master 2021-2022

TEXT MINING

Giorgia Riviecchio

Professor of Economic Statistics at Parthenope University

BIG DATA ANALYTICS

BIG DATA ERA

In the data-driven era, providing value means being able to translate data rapidly into value-add information.

In today's fast growly and highly complex world, the main challenge is to find meaning in data, so that the derived knowledge can be used to make informed decisions.

The Big Data era is characterized by:

- A large and complex quantity of data
- The development of IT

Examples: financial services, fraud detection, implementation of algorithms for trading, risk analysis, retail, CRM.

BIG DATA ANALYTICS

BIG DATA ANALYTICS

Big data analytics includes all the activities to help managers in coping with strategic decisions, in achieving major goals and in solving complex problems, by collecting, analyzing and reporting the most relevant information.

Information could be about the causes of the current situation, the most likely trends to occur, and what should be done as a result.

The purpose of the collection and processing of large volumes of complex data is to understand the trends of the phenomena of interest, uncover hidden trends, detect anomalies, in order to take data-driven decisions.

BIG DATA ANALYTICS

ARTIFICIAL INTELLIGENCE

Inundated with information and cognitive stimuli in a world where about 80% of data consists into free text, written in natural language, not classified or structured, treating and analyzing such a large amount of data can simultaneously represent a opportunity and a challenge.

Artificial intelligence techniques such as text mining, a technique aimed at extracting useful information hidden in a text, respond to this challenge.

This technique uses natural language processing (NLP) to transform free text, also known as "unstructured", into structured and normalized data.

A brief introduction is therefore required to distinguish structured data from unstructured ones.

BIG DATA ANALYTICS

STRUCTURED DATA VS UNSTRUCTURED DATA

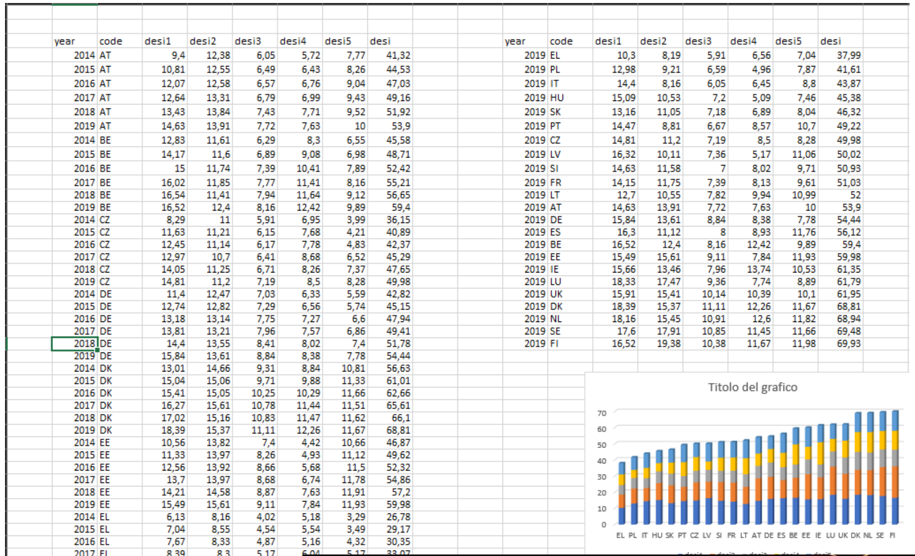
Structured data are classical data, which are organized and well formatted, and conform to the formal system of relational databases and traditional spreadsheets. In summary, they are all those quantitative data that can be stored in a SQL (Structured Query Language) database.

Structured data are analyzed by means of data mining.

Unstructured data, not being organized or correctly formatted, implies many difficulties in their collection, processing and analysis.

They are qualitative data processed by text mining.

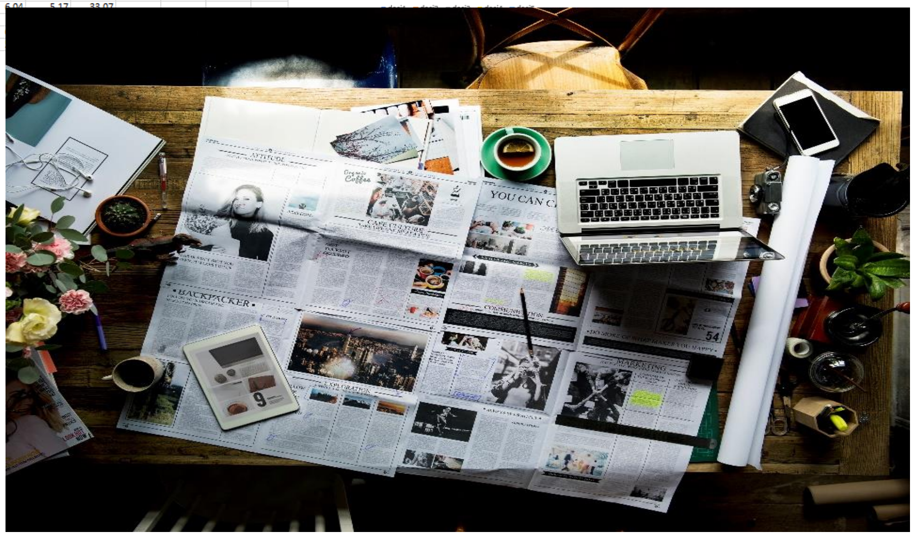
Examples: Newspaper articles, market place reviews, web pages, e-mails, messaging are just a few examples.



year	code	desi1	desi2	desi3	desi4	desi5	desi
2014	AT	9,4	12,38	6,05	5,72	7,77	41,32
2015	AT	10,81	12,55	6,49	6,43	8,26	44,53
2016	AT	12,07	12,58	6,57	6,76	9,04	47,03
2017	AT	12,64	13,31	6,79	6,99	9,43	49,16
2018	AT	13,43	13,84	7,43	7,71	9,52	51,92
2019	AT	14,63	13,91	7,72	7,63	10	53,9
2014	BE	12,83	11,61	6,29	8,3	6,55	45,58
2015	BE	14,17	11,6	6,89	9,08	6,98	48,71
2016	BE	15	11,74	7,39	10,41	7,89	52,42
2017	BE	16,02	11,85	7,77	11,41	8,16	55,21
2018	BE	16,54	11,41	7,94	11,64	9,12	56,65
2019	BE	16,52	12,4	8,16	12,42	9,89	59,4
2014	CZ	8,29	11	5,91	6,95	3,99	36,15
2015	CZ	11,63	11,21	6,15	7,68	4,21	40,89
2016	CZ	12,45	11,14	6,17	7,78	4,83	42,37
2017	CZ	12,97	10,7	6,41	8,68	6,52	45,29
2018	CZ	14,05	11,25	6,71	8,26	7,37	47,65
2019	CZ	14,81	11,2	7,19	8,5	8,28	49,98
2014	DE	11,4	12,47	7,03	6,33	5,59	42,82
2015	DE	12,74	12,82	7,29	6,56	5,74	45,15
2016	DE	13,18	13,14	7,75	7,27	6,6	47,94
2017	DE	13,81	13,21	7,96	7,57	6,86	49,41
2018	DE	14,4	13,55	8,41	8,02	7,4	51,78
2019	DE	15,84	13,61	8,84	8,38	7,78	54,44
2014	DK	13,01	14,66	9,31	8,84	10,81	56,63
2015	DK	15,04	15,06	9,71	9,88	11,33	61,01
2016	DK	15,41	15,05	10,25	10,29	11,66	62,66
2017	DK	16,27	15,61	10,78	11,44	11,51	65,61
2018	DK	17,02	15,16	10,83	11,47	11,62	66,1
2019	DK	18,39	15,37	11,11	12,26	11,67	68,81
2014	EE	10,56	13,82	7,4	4,42	10,66	46,87
2015	EE	11,33	13,97	8,26	4,93	11,12	49,62
2016	EE	12,56	13,92	8,66	5,68	11,5	52,32
2017	EE	13,7	13,97	8,68	6,74	11,78	54,86
2018	EE	14,21	14,58	8,87	7,63	11,91	57,2
2019	EE	15,49	15,61	9,11	7,84	11,93	59,98
2014	EL	6,13	8,16	4,02	5,18	3,29	26,78
2015	EL	7,04	8,55	4,54	5,54	3,49	29,17
2016	EL	7,67	8,33	4,87	5,16	4,32	30,35
2017	EL	8,39	8,3	5,17	6,04	5,17	33,07
2018	EL	9,34	7,99	5,38	[illegible]	[illegible]	[illegible]
2019	EL	10,3	8,19	5,91	[illegible]	[illegible]	[illegible]
2014	ES	9,07	10,29	5,88	[illegible]	[illegible]	[illegible]

year	code	desi1	desi2	desi3	desi4	desi5	desi
2019	EL	10,3	8,19	5,91	6,56	7,04	37,99
2019	PL	12,98	9,21	6,59	4,96	7,87	41,61
2019	IT	14,4	8,16	6,05	6,45	8,8	43,87
2019	HU	15,09	10,53	7,2	5,09	7,46	45,38
2019	SK	13,16	11,05	7,18	6,89	8,04	46,32
2019	PT	14,47	8,81	6,67	8,57	10,7	49,22
2019	CZ	14,81	11,2	7,19	8,5	8,28	49,98
2019	LV	16,32	10,11	7,36	5,17	11,06	50,02
2019	SI	14,63	11,58	7	8,02	9,71	50,93
2019	FR	14,15	11,75	7,39	8,13	9,61	51,03
2019	LT	12,7	10,55	7,82	9,94	10,99	52
2019	AT	14,63	13,91	7,72	7,63	10	53,9
2019	DE	15,84	13,61	8,84	8,38	7,78	54,44
2019	ES	16,3	11,12	8	8,93	11,76	56,12
2019	BE	16,52	12,4	8,16	12,42	9,89	59,4
2019	EE	15,49	15,61	9,11	7,84	11,93	59,98
2019	IE	15,66	13,46	7,96	13,74	10,53	61,35
2019	LU	18,33	17,47	9,36	7,74	8,89	61,79
2019	UK	15,91	15,41	10,14	10,39	10,1	61,95
2019	DK	18,39	15,37	11,11	12,26	11,67	68,81
2019	NL	18,16	15,45	10,91	12,6	11,82	68,94
2019	SE	17,6	17,91	10,85	11,45	11,66	69,48
2019	FI	16,52	19,38	10,38	11,67	11,98	69,93

Titolo del grafico
70
60
50
40
30
20
10
0
EL PL IT HU SK PT CZ LV SI FR LT AT DE ES BE EE IE LU UK DK NL SE FI



BIG DATA ANALYTICS

DATA MINING VS TEXT MINING

The challenge, therefore, with text mining consists precisely in the creation of algorithms (image), and computer and automated procedures capable of preparing this data source for an analytical use, accessible, in QUANTITATIVE form, to the various data mining algorithms.

Text mining and data mining are considered complementary techniques.

Specifically, text mining extracts, starting from NON-STRUCTURED textual information, STRUCTURED numerical data to be processed later with DATA MINING techniques.

To transform unstructured textual resources into structured information text mining adopts linguistic and statistical techniques capable of both analyzing free text formats and combining each document with metadata.

BIG DATA ANALYTICS

NATURAL LANGUAGE PROCESSING (NLP)

Therefore, Text Mining is thus an Artificial Intelligence (AI) technique that uses the Natural Language Processing (NLP) to transform the free, unstructured text of documents / databases such as web pages, newspaper articles, e-mails, press, post / comment on social media etc. in structured and normalized data.

With the expression Natural language processing (NLP), we mean that part of computer science and artificial intelligence that deals with the processing of human languages, capable of performing a linguistic analysis, which, essentially, helps a machine to "read" the text.

BIG DATA ANALYTICS

TEXT MINING

Typically, text mining is used for two main functions:

classifying data by segment and predicting its behavior.

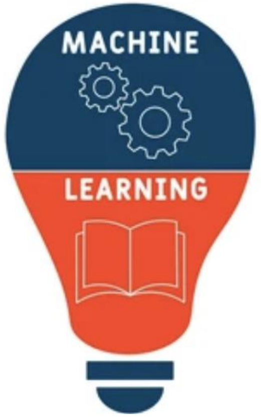
The most commonly used features are:

- Cleaning, literally the process of cleaning the text aimed at organizing the material for analysis.
- Extraction, the extraction of keywords that involves the choice of words and phrases that best identify the nature of the text.
- Categorization, the classification of the text into one or more categories.
- Modeling, which combines the essential parts of the text with structured data attributes and is useful for optimizing the forecast.

BIG DATA ANALYTICS

OBJECTIVES OF TEXT MINING

- Identify thematic groups
- Classify documents into predefined categories
- Discover hidden associations (links between topics, or between authors, temporal trends, …)
- Extract specific information (ex: names of geniuses, names of companies, …)
- Train search engines
- Extract concepts for the ontology learning.
- Identify attitudes and opinions on opposite emotional states (sentiment analysis)



BIG DATA ANALYTICS

Text mining applications

Among all the possible applications, the most frequent are:

- Digital marketing for contextual retargeting. Compared to the traditional cookie-based approach, contextual advertising offers greater accuracy, fully preserving the user's privacy.

- Customer care, as it allows to monitor and improve the customer experience using various sources of information such as surveys, trouble tickets and customer call notes, to optimize problem resolution

- Engagement analysis and brand reputation: to analyze or predict the needs of customers and understand the perception of their brand, it is possible to analyze both large volumes of unstructured data, and by extrapolating opinions, emotions and relationships with brands and products and feelings.

BIG DATA ANALYTICS

STEPS OF TEXT MINING

1. Pre – Processing:

- Corpus
- Tokenization
- Stemming

2. Document term matrix

3. Exploratory Analysis

- Word frequency e word cloud
- Word association
- Word clustering

```
      Share rate                 3746.23h/s              319
      Hash rate                       8                  318
      Submitted                       8                    1
      Accepted                        0
      Stale
      Hi/Lo Share Diff  0.041204 /  2.1791e-05
      Diff: Net 5.2902, Stratum 1.3483, Target
      Net hash rate (est) 189.34 Mh/s
01 02:58:01] 320 Accepted 319 S1 R0 B0, 45.626 sec (35
01 02:58:19] New Stratum Diff 1.35461, Block 217826, J
01 02:58:23] 321 Accepted 320 S1 R0 B0, 21.239 sec (35
1 02:58:41] 322 Accepted 321 S1 R0 B0, 18.646 sec (34
1 02:59:19] New Stratum Diff 1.36724, Block 217826, J
 02:59:21] 323 Accepted 322 S1 R0 B0, 40.121 sec (34
 03:00:19] New Stratum Diff 1.37363, Block 217826, J
 03:00:31] 324 Accepted 323 S1 R0 B0, 69.044 sec (35
03:00:32] New Stratum Diff 1.36993, Block 217827, J
03:00:32] Block Algos: Dark Lite Turtle (Rot. 8.1,
          Hashrate: 2009.5 h/s | Average: 2563.4 h/
```

BIG DATA ANALYTICS

PRE-PROCESSING STEPS:

- Data selection (creating corpus)
- Data manipolation (tokenization)
- Reduction of the variants associated with each word (stemming)
- Generation of characteristics and vectors
- Classification
- It requires solving numerous difficulties in data processing
- Ambiguity of language: the same word can take on different meanings depending on the context and different words can mean the same thing (synonyms)
- Language sensitivity (sensitive topics)
- Numerous dimensions involved in the extraction of concepts / words
- Difficulties due to: spelling errors, abbreviations, language variants, etc…

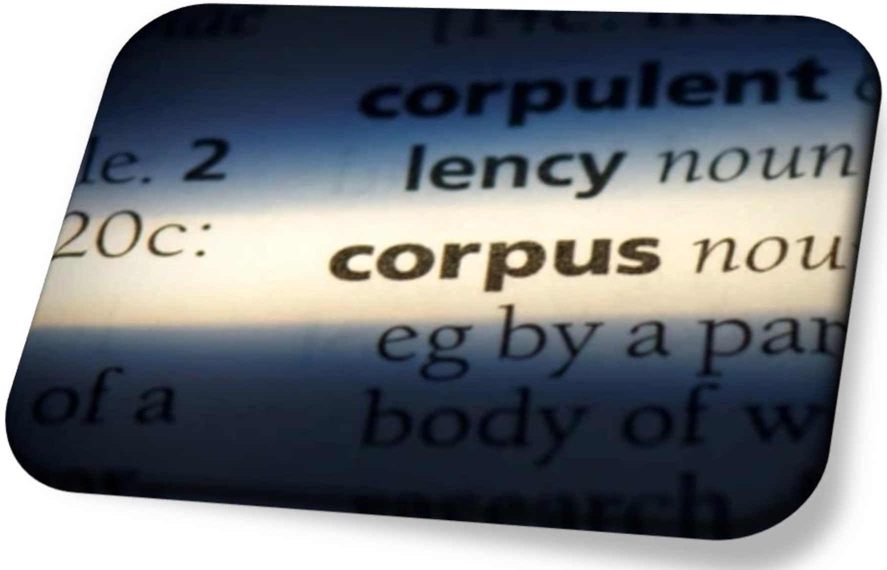
BIG DATA ANALYTICS

PRE-PROCESSING - CORPUS

It transforms a group of separate text documents into a single text that merges them all.

- □ Client 1: Good morning, I would like to have information on product X.
- □ Client 2: I would like to know where product X can be purchased.

Good morning, I would like to have information on product X. I would like to know where product X can be purchased.

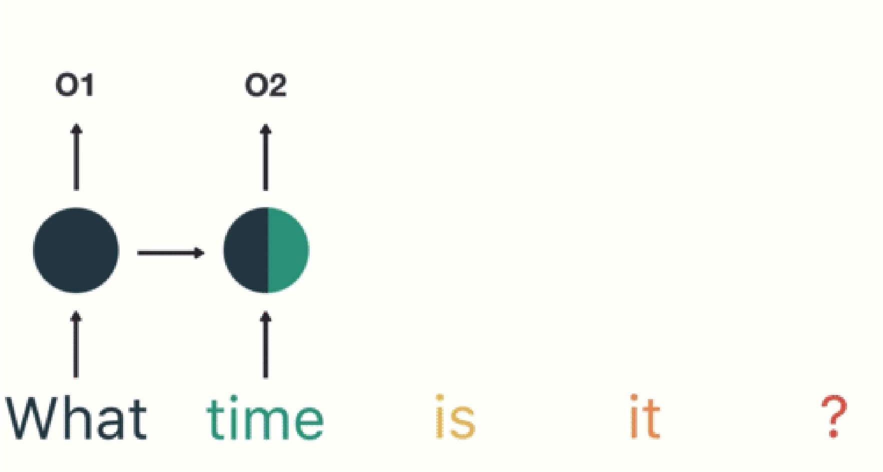


BIG DATA ANALYTICS

PRE-PROCESSING - TOKENIZATION

It is the manipulation of data to extract the relevant elements, namely words, phrases or even letters.

The text is divided into tokens, which are blocks of atomic text, made up of indivisible characters, as a sequence of characters surrounded by delimiters



There is great demand on the market for this product

There | is | great |demand |on | the | market| for | this | product

BIG DATA ANALYTICS

PRE-PROCESSING - TOKENIZATION

DELETE STOPWORDS

Stopwords are non-informative words, like articles, prepositions.

In the text processing, it is therefore necessary to load, for each language, a list of "stopwords" that the system will eliminate before proceeding with the processing of the text. Example on a facebook comment.

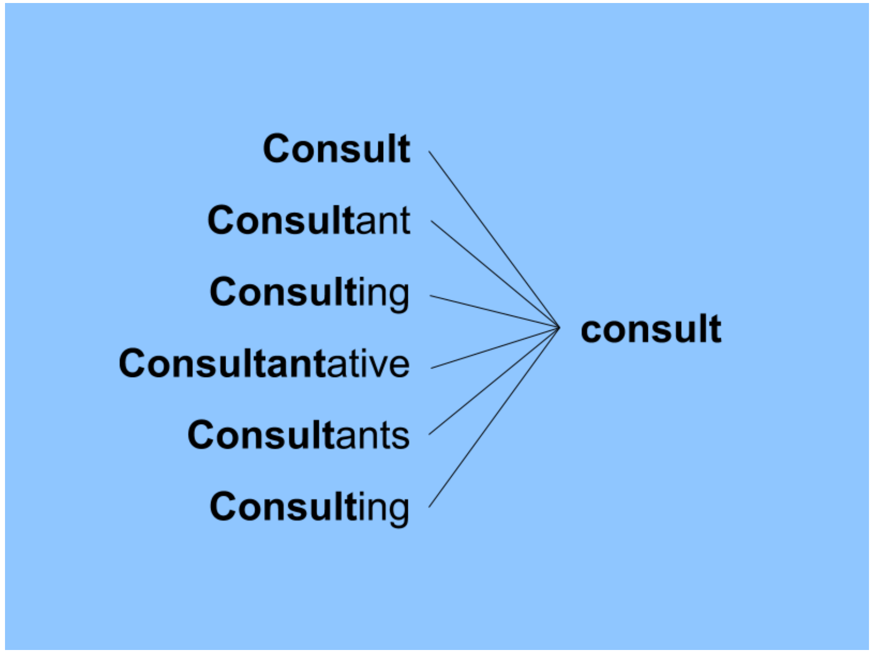
Description	Before	After
Remove capital letter	The product is fantastic, the NUMBER 1!	the product is fantastic, the number 1!
Remove punctuation	the product is fantastic, the number 1!	the product is fantastic the number 1
Remove numbers	the product is fantastic the number 1	the product is fantastic the number
Remove spaces	the product is fantastic the number	the product is fantastic the number
Remove specific terms	the product is fantastic the number	product fantastic number

BIG DATA ANALYTICS

PRE-PROCESSING - STEMMING

Process of reducing the inflected form (any morphological variation) of words to the basic form, called root or theme.

Extraction of the root of a word, removing affixes and endings (e.g. playing, playing games, players, play…).



BIG DATA ANALYTICS

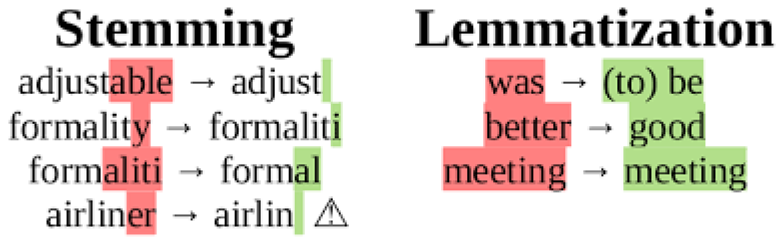
PRE-PROCESSING - LEMMATIZATION

Identification of the NLP vocabulary (lemma) starting from a word with ending. Determining the part of speech of a word, and then applying the different normalization rules.

Given a wordform, stemming is a simpler way to get to its root form. Stemming simply removes prefixes and suffixes.

Lemmatization on the other hand does morphological analysis, uses dictionaries and often requires part of speech information.

Thus, lemmatization is a more complex process



BIG DATA ANALYTICS

2. DOCUMENT-TERM MATRIX (DTM)

It consists of one of the most common formats for representing a corpus of text in a bag-of-words format.

The Document matrix is a table containing the frequency of each word occurring in the text.

	word 1	word 2	word 3	…	word k
document 1					
document 2			n_{23}		
document 3					
…					
document n					

n23= Number of times which the word 3 occurs in the document 2

The resulting data frame is

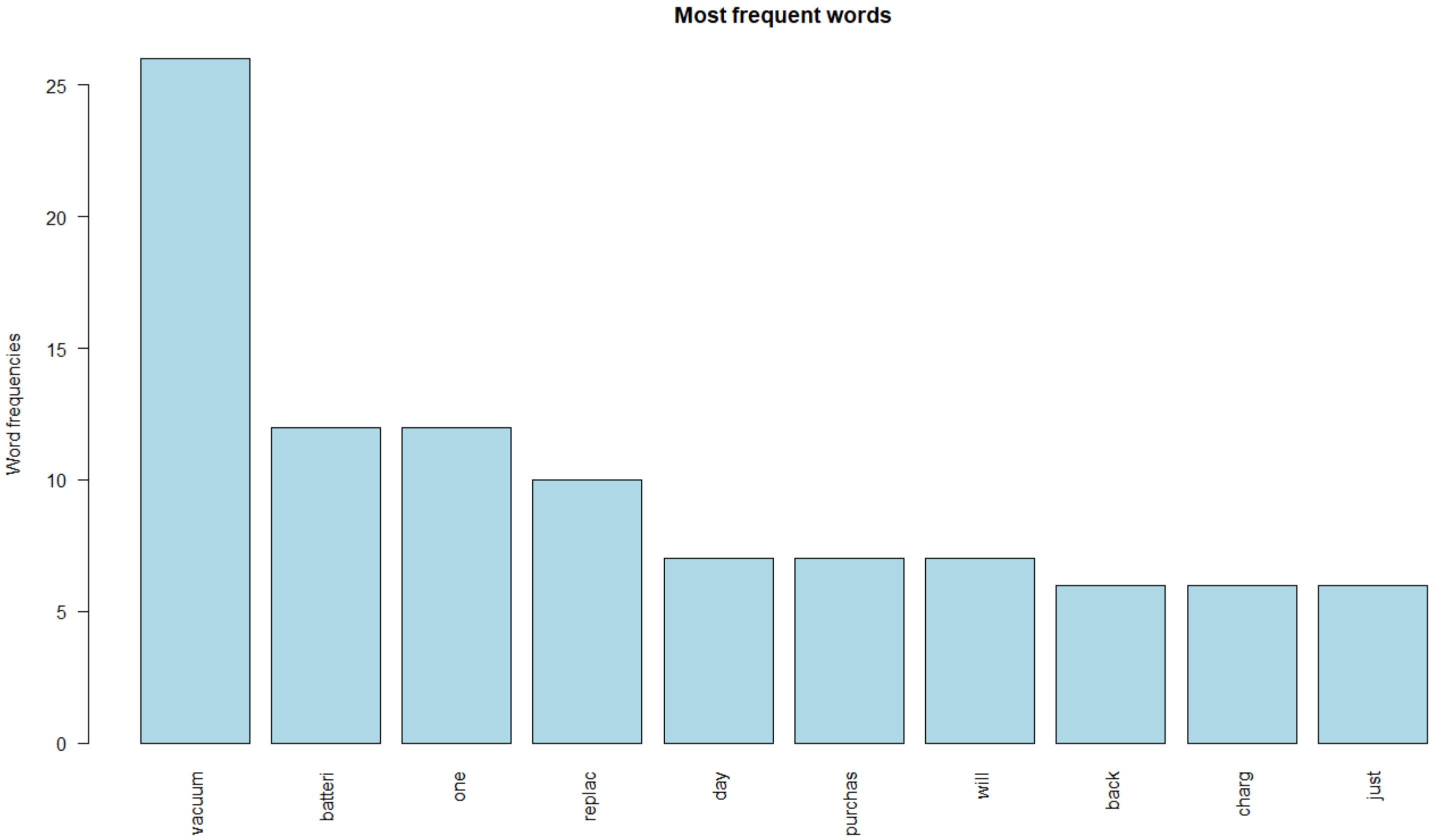
Term	Frequency
Bad	77
Beautiful	58
Need	30
Assistance	26
break	24
small	20
negative	18
failure	10
problem	7
price	5

BIG DATA ANALYTICS

EXPLORATORY ANALYSIS -

WORD FREQUENCY

The bar-plot allows you to compare the frequency of the most used words in a text



BIG DATA ANALYTICS

EXPLORATORY ANALYSIS -

WORD CLOUD

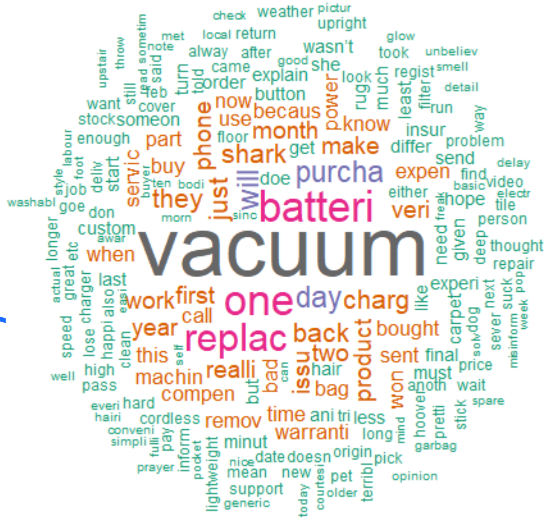
Once the frequency associated with each word detected in the text has been identified, the so-called word cloud can be presented, showing the most frequent terms of the text.

Word clouds are a powerful communication tool. They're easy to understand, easy to share, and they're impactful.

Word clouds add simplicity and clarity. The most used keywords stand out best in a word cloud. By size and color.

Word clouds are more visually appealing than table data

fun
easy
cool
presentations
inclusive
share
software
live
beautiful
reflection
exciting
anonymous
thoughts
interactive
brainstorm
knowledge
ideas
ice breaker

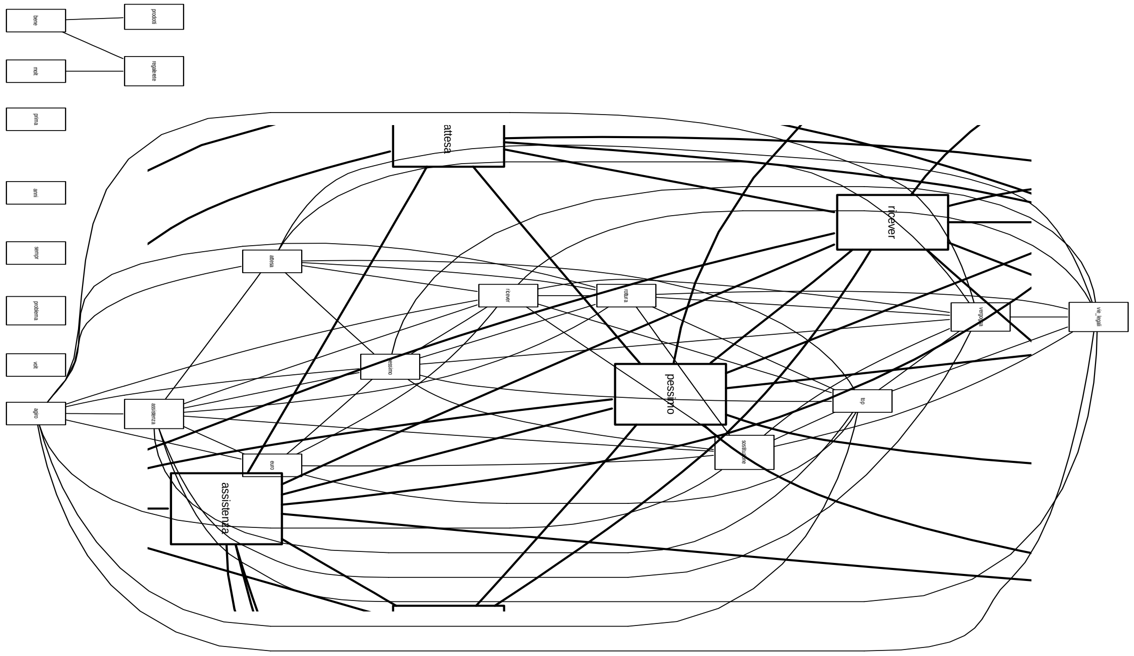


bread
breakfast
fare
tea
food
supper
break
walk
refreshment
evening
drink
beer
holiday
coffee
weekend
meal
menu
cocktail
afternoon
entertainment
sandwich
wedding
soup
dinner
salad
restaurant
picnic
pizza
trip
shopping
snack
luncheon
dessert
buffet
dish
cake
ride
treat
recipe
reception

BIG DATA ANALYTICS

EXPLORATORY ANALYSIS - WORD ASSOCIATION

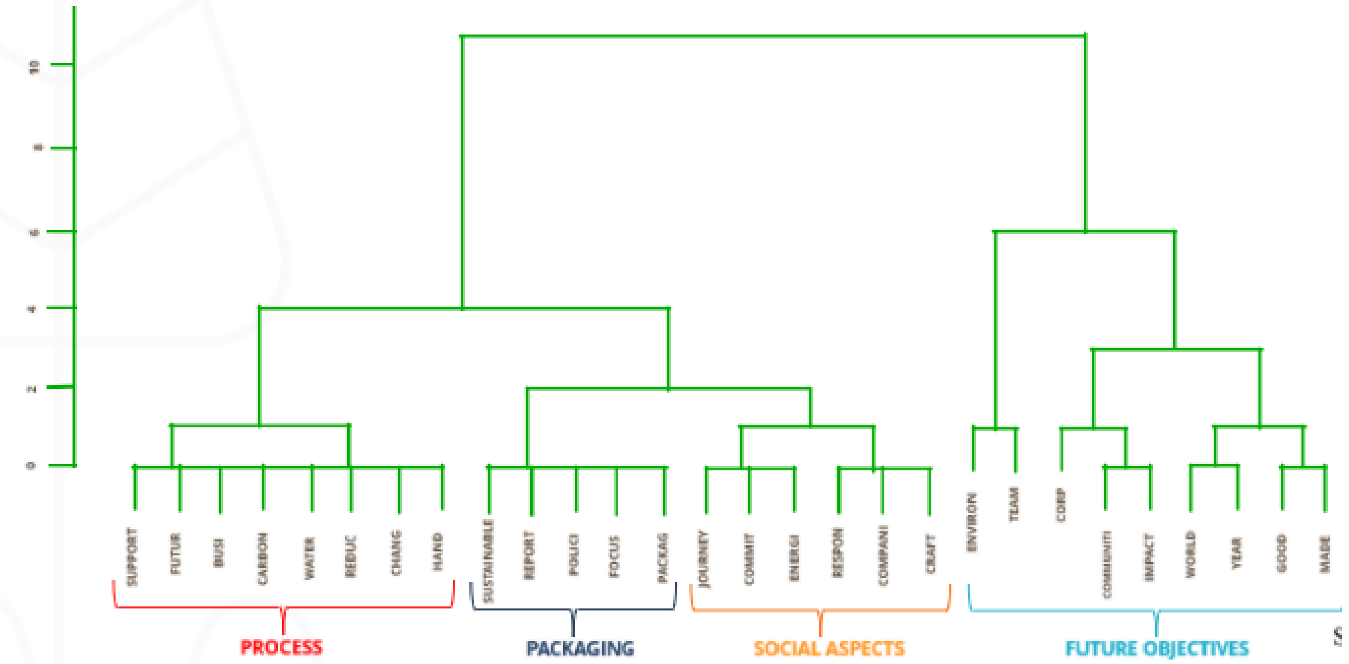
The link between the words can be summarized through a plot of the network that can be created between them.



BIG DATA ANALYTICS

EXPLORATORY ANALYSIS - WORD CLUSTERING

THE HIERARCHICAL CLUSTER OF WORDS IS REPRESENTED BY THE DENDROGRAM



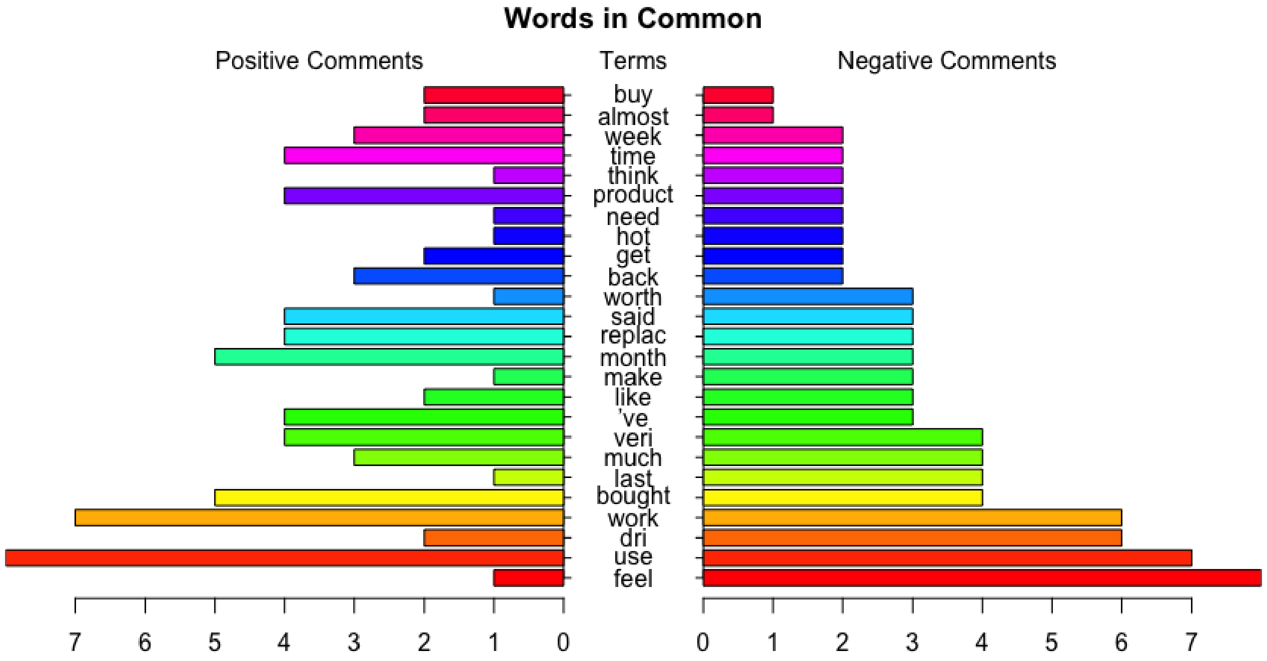
BIG DATA ANALYTICS

EXPLORATORY ANALYSIS - POLARYZED TAG PLOT

The polarized tag-plot (pyramid plot) allows to identify the frequency of a term used into two different documents

It is created starting from a data matrix with all common words occurring in both corpora.

Adding another matrix for the absolute difference between both corpus for each word and the graph is created.



BIG DATA ANALYTICS

EXPLORATORY ANALYSIS – SENTIMENT ANALYSIS

Sentiment Analysis is a technique aimed at identifying the opinions expressed in online texts on a product or service, on a company, on a brand or on an event. This type of analysis allows to understand the nature of the interactions carried out between users, in a precise context and in a given period of time.

It is a multifunctional tool, it can be used in a different but functional way in various fields, even very different from each other.

EXAMPLES: Companies that want to know directly the opinion of their users, political parties, sociologists, museums, research bodies, computer scientists, even seismologists, epidemiologists, …

BIG DATA ANALYTICS

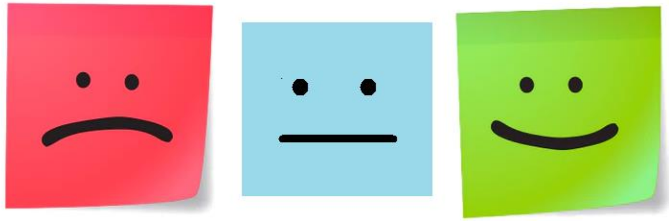
EXPLORATORY ANALYSIS – SENTIMENT ANALYSIS

Sentiment Analysis is a useful tool for market research, such as understanding the opinion of the chosen target on a specific product or topic of interest or it can facilitate the work of market segmentation in order to get to know your customers or main stakeholders better.

Furthermore, it can be a useful tool to analyze Brand Reputation, through Social Networks, to understand the general opinion of stakeholders and in particular of customers regarding their brand.

It still allows you to monitor marketing campaigns by evaluating the effectiveness of a specific marketing activity.

Evaluating digital ***word of mouth*** thus becomes an important aspect in evaluating the brand reputation of a brand or customer satisfaction, in relation to a service or product.

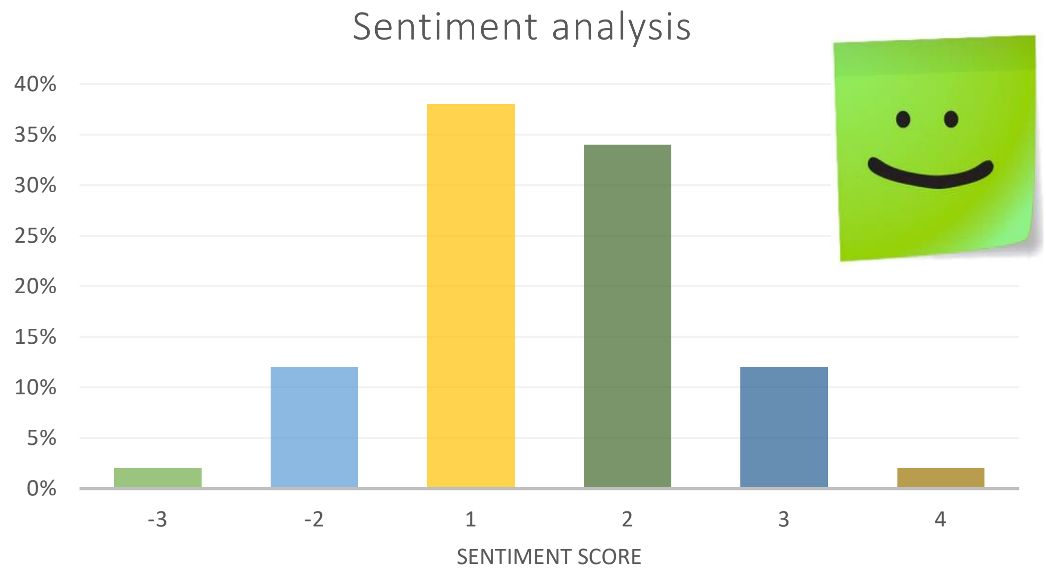


BIG DATA ANALYTICS

EXPLORATORY ANALYSIS – SENTIMENT ANALYSIS

Furthermore, given that the discussion on social media can take place at the same time as the choice process is matured or even coinciding with the purchase process, the monitoring of social channels has an important impact on the success of new products, and on the effectiveness of communication or marketing campaigns (Jansen et al., 2009).

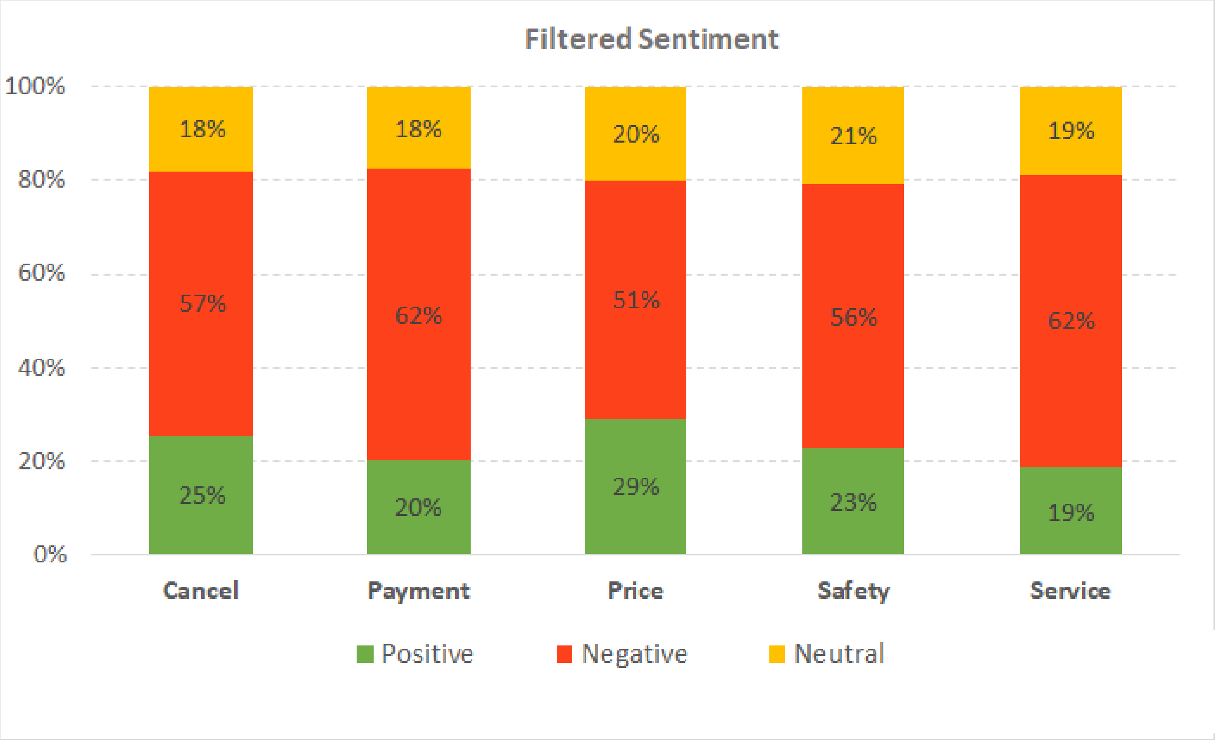
However, the judgment can also be expressed subsequently with respect to the consumption of the same, thus allowing the satisfaction of the buyer to be assessed. Social media becomes an important brand management tool. With this approach, therefore, consumer sentiments and opinions become the fulcrum of analysis. There are two approaches used by companies to monitor the mood of their market: Top Down and Bottom Up.



BIG DATA ANALYTICS

EXPLORATORY ANALYSIS – SENTIMENT ANALYSIS

Identify a sort of polarity that people show towards a topic by creating an index that associates numerical values from "completely positive" to "completely negative", passing through the neutral position.



MASTER MEIM 2021-2022

THANKS FOR YOUR ATTENTION AND LET’S START TO APPLY TEXT MINING!