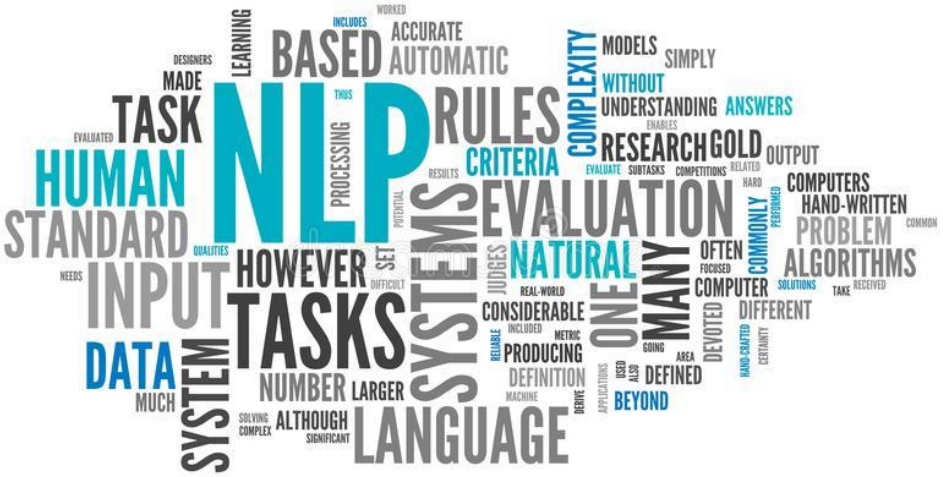
University of Pisa

Topic Detection in Strategic Management

An overview on actual AI techniques applied to business documents



Drafted by: Arancio Febbo Salvatore, Di Donato Mattia, Giorgi Matteo

Summary

Abstract

Nowadays Artificial Intelligence applications are innumerable and many of these could have important impacts on activities of different type of companies. It is also possible to apply AI techniques to natural text, in this case we talk about NLP, Natural Language Processing.
Like many computational tasks of a certain complexity, NLP has found particular benefit from advancements in the field of Machine Learning, an approach which is powered and made possible by big data. Data in ML is important for at least three reasons:

- it is used to train the model, in this way the model is able to perform a task
- the more the quality of data the more the quality of the model
- it is used to test the model, which is essential to evaluate it and to understand how good it will be in the execution of the task on similar data

In the NLP case, data are linguistic data: words, sentences, texts each of them that belong to the human language. The data is collected in resources, *corpus* or *datasets*. Although both terms are often with the same meaning, indicating a collection of linguistic data, there is a subtle difference between the two: a corpus becomes a dataset when all or part of it is chosen as the data source for an NLP project (in the rest of the article we will use the term dataset to refer to data collections).

Introduction

Our interest is to analyse MIT Sloanmanagement review papers to find the main topics related to strategic management. The data that we processed comes from a scraping of thirty years of MIT online papers, from 1990 to 2022, for a total of 1753 documents.
The Branch of NLP that aims to discover hidden semantic structures in a text body is called Topic Modeling. The intuition behind this technique is that if a given document is about a particular topic, we should expect to find a high-medium repetition of specific set of words in it.
Usually, a document covers multiple topics in different proportions, indeed the way these algorithms work is by assuming that each document is composed of a mixture of topics, and then trying to find out how strong a presence of each topic has in each document. This is done by grouping together the documents based on the words they contain and noticing correlations between them.
In other words, it is possible to find what are the topics and how much relation there is between them and the documents analysed. In the following paragraphs we will describe and compare some of the most popular NLP techniques with our solutions.

Topic modeling algorithms

The general idea behind the following two algorithms is that documents with the same topic will use similar words. It's also assumed that every document is composed of a mixture of topics, and every word has a probability of belonging to a certain topic.

Latent Semantic Analysis (LSA)

LSA is the traditional method for topic modeling, it is based on a principle called the distributional hypothesis: “words and expressions that occur in similar pieces of text will have similar meanings”. Therefore, for every word in each document, we can calculate the frequency of them and group together documents that have high frequencies of the same words. The frequency can be calculated simply by counting the times a word appears in a document but this approach proves to be a bit limited, so tf-idf (term frequency–inverse document frequency) is normally used. Tf-idf considers how common a word is overall (in all documents) in comparison to how common it is in a specific document, so more common words are ranked higher since they are considered to be a better “representation” of a document, even if they are not the most numerous.
These algorithms ignore syntax and semantics of the words and treat every document as an unsorted BoW (bag of words). A BoW is a representation of the text that involves a vocabulary of the words and a measure of the presence of them. With this BoW we can obtain a representation of our data that is a document-term matrix that has a row for every word and a column for every document. Each cell is the calculated frequency for that word in that document.
Hidden inside it is what we want: a document-topic matrix and a term-topic matrix, which relate documents to topics and terms to topics. These matrices are the ones that show information about the topics of the texts.

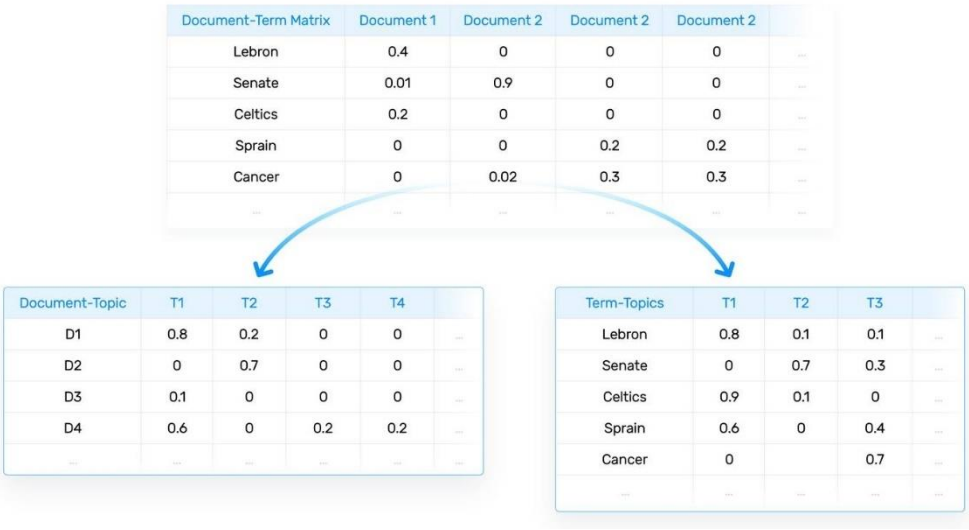


Figure 1 : splitting of document-topic matrix (on the left) and a term-topic matrix (on the right) from document-term matrix (on top)

The vectors that compose these matrices represent documents expressed with topics, and terms expressed with topics; they can be measured with techniques such as cosine similarity to evaluate.

Latent Dirichlet Allocation (LDA)

LDA is another Topic Modeling technique which is based on the Dirichlet probabilistic distribution. As LSA we need to provide the number of clusters (or in this case topics) that are believed to exist and calculate the dictionary from the document that we pass and with this also the BoW which is essential for these algorithms. Differently form LSA, LDA needs as input also 2 other hyperparameters, called α and β.

- Alpha controls the similarity of documents. A low value will represent documents as a mixture of few topics, while a high value will output document representations of more topics -- making all the documents appear more similar to each other.
- Beta controls the similarity of topics. A low value will represent topics as more distinct by making fewer, more unique words belong to each topic. A high value will have the opposite effect, resulting in topics containing more words in common.

Top2Vec

Top2Vec is an innovative method which do not rely on bag-of-words representation and do not ignore the ordering and semantics of words. Instead, it creates a distributed representations of documents and words which can capture semantics of words and documents. Thus, this algorithm leverages joint document and word semantic embedding to find topic vectors. This model does not require stop-word lists, stemming or lemmatization, and it automatically finds the number of topics. At a high level, the algorithm performs the following steps to discover topics in a list of documents:

1. Generate embedding vectors for documents and words, in this way we will have that document will be near to other similar documents and near to the most distinguishing word of them.

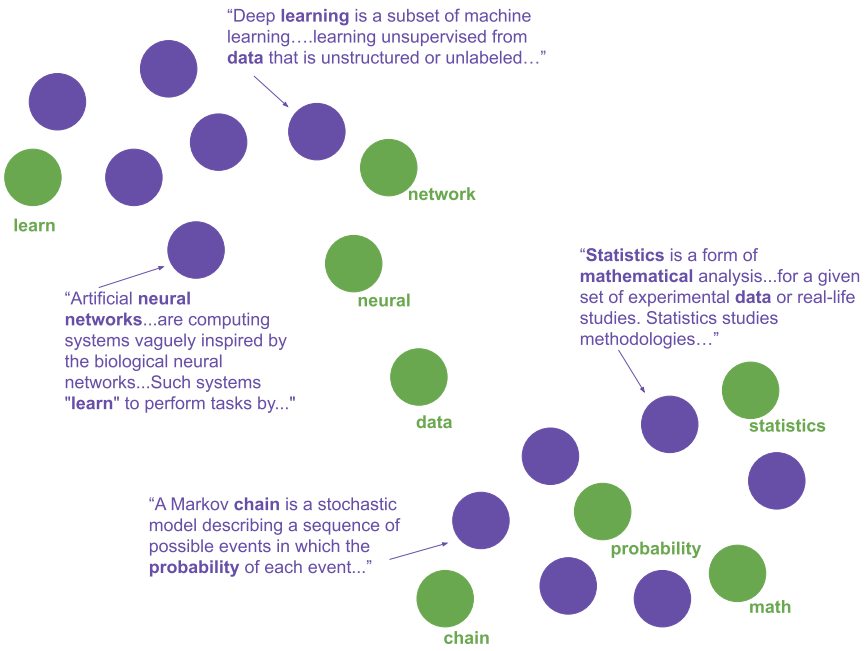


Figure 2: document and word embedding in the same space

2. Perform dimensionality reduction on the vectors using an algorithm such as UMAP. Similar to other dimensional-reduction algorithms like PCA, this phase allows to reduce the dimension of the vector in order to reduce the sparsity and find dense areas
3. Cluster the vectors using a clustering algorithm such as HDBSCAN, an extension of DBSCAN based on hierarchical clustering
4. For each dense area discovered we calculate the centroid of document vectors, this will be the topic vector.
5. Find the n-closest word vectors in the same cluster of the topic vector, these will be the word vectors.

Custom models

The following algorithms have been implemented by us to obtain a custom solution to our problem. We called them Specific Topic Discovery (STD) and General Topic Discovery (GTD).
In both we start by carrying out these preliminary operations on documents:

1. Pre-processing:
    1. Remove multiple spaces between sentences
    2. Tokenization by word: an array has been created containing all the words of the document
    3. Relevant stop-words removal: we removed from the array all the common English words importing the NLTK library and extending it with numbers and names of people that were not relevant for our analysis.
    4. POS-Tagging: we attributed to all the words in the array a grammatical tag and we extracted only the nouns
    5. Lemmatization: we brought each word to the basic form
2. Embedding: we transform each word of each document into a vector of 300 elements using a pretrained Neural Network of Spacy, the “en_core_web_lg”. In this space, words with similar meaning will be near to each other.

STD

STD is a Python algorithm that we realized for analysing the single document. All vector that we obtain will be treated as a single cluster. The more the similarity in meaning the more the words will be nearer. The searched topic will be identified with high probability in dense area therefore calculating the centroid will bring us closer to this one.
Using the similarity cosine between centroid and all the words in the cluster, will allow us to obtain a list of word sorted by the distance from the centroid. We choose words near to the

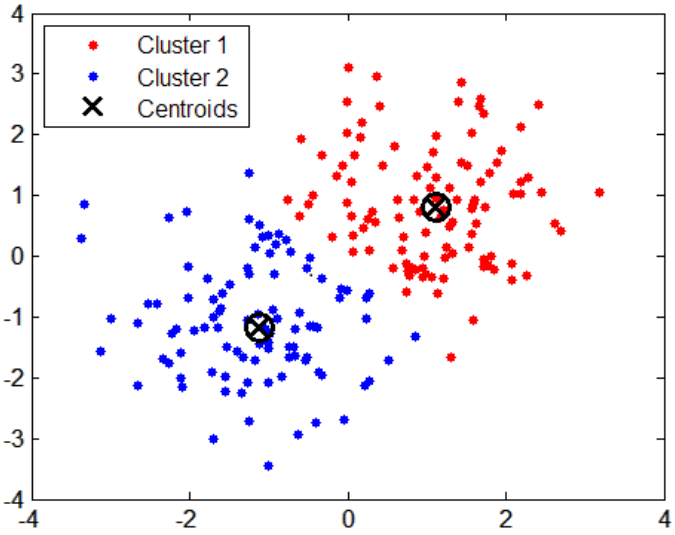


Figure 3: example of centroids of clusters

centroid with a similarity value that goes from 1.0 to 0.3 and then we compute the frequency of this words within the text in the end we output a visual representation of the result.

GTD

GTD is a Python algorithm that we realized for analysing multiple documents within a year.
Once all the vectors related to all the words of all documents have been generated, the DBSCAN density algorithm will be used for data analysis, manipulating the min_samples, eps and metric parameters we will find the best clustering.
The *min_samples* allow us to define a minimum number of elements that must populate a cluster. A value of 5 was chosen as a compromise between many unsignificant clusters (for values lower than 5) and clusters too specific (for values greater than 5).

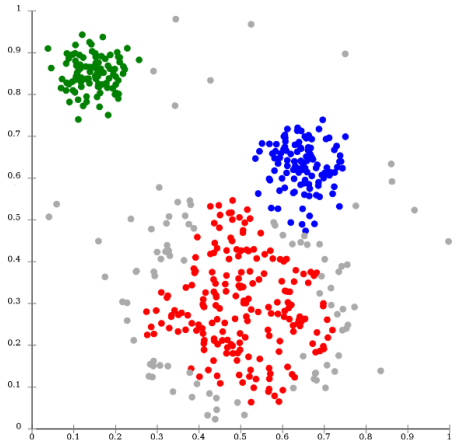


Figure 4: DBSCAN example

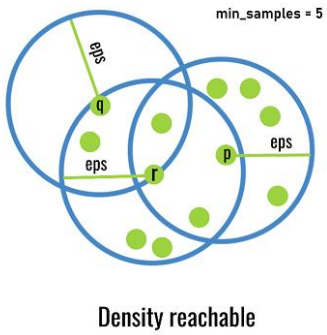


Figure 5:min samples and eps

Eps is the maximum distance between two samples for one to be considered as in the neighbourhood of the other. This is the most important DBSCAN parameter to choose appropriately for our dataset and distance function. To ensure a good result, Eps was dynamically chosen through a loop with a variation from 0.3 to 1 with a step of 0.1, where the range chosen depends on the metric taken into account. In this case the radius will correspond to the similarity of the words. The value chosen will be the one that will create more clusters and therefore topics.

The *metric* used to calculate the distance between the vectors of word is the cosine similarity which values goes from 1 to -1. Two vectors will have a similarity of 1 if they will be proportional, a value of 0 if they will be orthogonal and a value of -1 if they are opposite.

$$cosine\ similarity = S_c(A,B) = \frac{A * B}{||A|| * ||B||} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\ \sqrt{\sum_{i=1}^{N} B_i^2}}$$

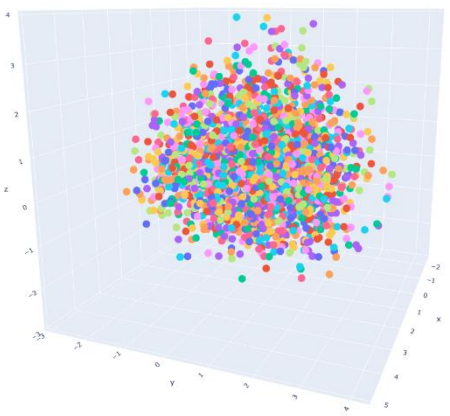


Figure 6: cluster of all embedded words

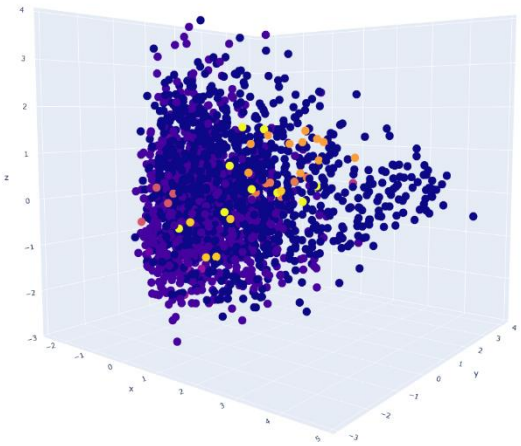


Figure 7: plotting of clusters

Results

The extraction of the topics carried out was developed using a Top-Down approach as well as making a distinction for decades, single years and single documents present in each year.
Using the Top2Vec algorithm, the topics of a decade have been identified.
Going into detail and then focusing on a single year, the LDA and LSA were used and compared with our solution, the GTD, which gave us sometimes a more detailed result. Finally, to go into even more detail, an algorithm (STD) has been created which allows to know the theme of the single document.
The analysis brought to light several evolutions at the company level. Analysing the various decades, from 1990 to 2020, we see a much more digitized approach to business entrusted to artificial intelligence. Even the internal management of companies has changed a lot, managers create a real strategic plan by giving more importance to the employee and their feedback.
Below are shown the results obtained on all the decades in the following intervals (1990-1999, 2000-2009, 2010-2019, 2020-2022) and for one year of each decades its results will be analysed and published.

Top2Vec from 1990 to 1999

Corporate strategies to remain competitive in the market and a corporate restructuring to increase profitability by also joining partnerships. Particular attention to customer needs.

Below the list of topics:
[*managers, brainstorming, ceo, organizing, hierarchical, cio, redesign, productivity, outsourcing, competitor, consultants, consultant]*
[managerial, profitability, negotiators, customer, salespeople, combinations, ibm, innovators, provider, concurrent, standardization, partnership, consortium]

Word frequency of 1990

Word frequency contained in all (12) documents of 1990:

Analysing the frequency of words of all 1990 documents with the Tag Cloud method, is clear that, the main topic is to improve the company through business innovation where the customer's

interests play a pivotal role in improving processes and approaches to the market. Listed below are the words that made people understand what was just said.

Below the list of words that represent the topic:
[Making e demand, improve process knowledge, Make better process, Upgrade Approach, Idea for innovation, Market opportunity and make better use of It, Desire / need of individual / interest focus on clients Attention]

Topics in LDA model of 1990

Through the LDA model we confirm the particular strategic attention to customer and sales process.

Below the list of topics:
Topic: 0
Words: 0.001"company" + 0.001*"process" + 0.001*"sale" + 0.001*"product" + 0.001*"service" + 0.001*"management" + 0.001*"system" + 0.001*"technology" + 0.001*"compensation" + 0.001*"manager"*

Topic: 1
Words: 0.028"technology" + 0.017*"management" + 0.016*"problem" + 0.016*"company" + 0.015*"program" + 0.014*"client" + 0.013*"system" + 0.012*"service" + 0.010*"product" + 0.010*"group"*

Topic: 2
Words: 0.045"process" + 0.026*"organization" + 0.023*"partnership" + 0.016*"business" + 0.015*"system" + 0.014*"executive" + 0.013*"information" + 0.012*"relationship" + 0.011*"manager" + 0.011*"company"*

Topic: 3
Words: 0.001"service" + 0.001*"company" + 0.001*"organization" + 0.001*"process" + 0.001*"customer" + 0.001*"system" + 0.001*"problem" + 0.001*"management" + 0.001*"manager" + 0.001*"technology"*

Topic: 4
Words: 0.044"service" + 0.022*"company" + 0.021*"sale" + 0.021*"customer" + 0.019*"product" + 0.013*"name" + 0.012*"compensation" + 0.011*"association" + 0.011*"manager" + 0.011*"extension"*

Coherence Score: 0.26435471506963826

Topics in LSA model of 1990

In this case the LSA returns the same information as the LDA, with only substantial difference that the score is much higher, confirming the previous deduction.

Below the list of topics:
Coherence: 0.39375169878764094
[
Topic: 0
(0, '0.517"process" + 0.305*"service" + 0.253*"company" + 0.205*"organization" + 0.192*"customer" + 0.186*"system" + 0.184*"business" + 0.161*"manager" + 0.151*"management" + 0.150*"information"'*
Topic: 1
(1, '-0.675"service" + 0.519*"process" + -0.290*"customer" + -0.161*"employee" + 0.135*"business" + -0.134*"company" + 0.110*"redesign" + -0.097*"manager" + 0.085*"information" + -0.084*"standard"'*
Topic: 2
(2, '-0.638"partnership" + -0.359*"executive" + 0.341*"process" + -0.290*"relationship" + 0.145*"service" + -0.141*"organization" + 0.121*"company" + -0.111*"line" + -0.103*"technology" + -0.100*"information"'*
Topic: 3
(3, '-0.453"sale" + -0.368*"product" + -0.293*"name" + -0.276*"compensation" + -0.247*"association" + -0.232*"extension" + 0.223*"service" + 0.219*"process" + 0.178*"partnership" + -0.115*"effort"'*
Topic: 4
(4, '0.432"sale" + -0.382*"name" + -0.325*"association" + 0.313*"compensation" + -0.306*"extension" + -0.288*"product" + 0.133*"plan" + -0.109*"line" + 0.107*"management" + -0.101*"quality"'*
]

Topics in GTD model of 1990

The GTD analysis not only gives further confirmation of the analyses carried out with the other algorithms but returns a pool of topics that allow us to extrapolate the main topic but also the secondary ones that allow us to interpret the year in a more detailed and meticulous way. For example, analysing a secondary topic (Topic 6), we can deduce that the strategies discussed so far will also have been adapted in the automotive market.

This is the output of the annual custom model:

Topic 0 :
len: 987
cluster words:
therefore, concern, consider, understanding, fact, reason, way, sense, matter, understand, hence, extent, possibility, assume, need, depend, importance, purpose, lack, well, responsibility, kind, consideration, explain, meant, sufficient, one, determine, appropriate, individual, change, make,

attention, basis, nothing, specific, potential, situation, knowledge, future, focus, doubt, opportunity, result, affect, approach, mind, latter, idea, others, better-paying, thats, bring, sort, effort, some-one, something, ability, benefit, depends, interest, desire, suggest, recognize, maintain, consequence, continue, anything, intention, take, provide, identify, time, making, process, thinking, impact, theyre, example, demand, require, avoid, work, point, manner, perspective, imperative, present, examine, whats, higher-leverage, being, respect, thing, someone, develop, encourage, value-creation, trust, risk,

Topic 1:
selected year: 1990
len: 4
cluster words:
building, brick, house, construction,

Topic 2:
selected year: 1990
len: 6
cluster words:
director, executive, president, chief, officer, vice,

Topic 3:
selected year: 1990
len: 4
cluster words:
inefficiency, mismanagement, unreliability, inefficient,

Topic 4:
selected year: 1990
len: 4
cluster words:
room, bathroom, desk, floor,

Topic 5:
selected year: 1990
len: 7
cluster words:
detroit, chicago, maryland, carolina, minneapolis, fargo, portland,

Topic 6:
selected year: 1990
len: 11
cluster words:
miata, cadillac, jeep, pontiac, chevrolet, audi, volkswagen, honda, ford, ferrari, chrysler,

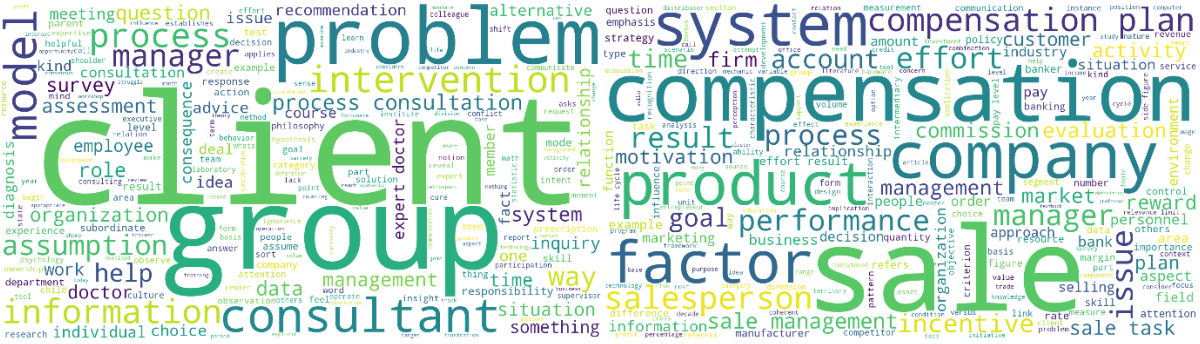
Topic 7:
selected year: 1990
len: 5
cluster words:

skirt, hat, shirt, clothing, pant,

Topic 8:
selected year: 1990
len: 9
cluster words:
hair, shampoo, conditioner, shampoo/conditioner, mouthwash, detergent, deodorant, soap, toothpaste,

Topics in STD model of 1990

These are tag-clouds of two articles from 1990 that confirm the topics covered in that year.



Top2Vec from 2000 to 2009

The common topic is always business organization and outsourcing, but there are more words that identify the birth of start-ups. So together with them we talk about investments, patents, entrepreneurs, and bankruptcies. Related to start-ups we also talk about multinational companies (MNCS) such as 'Sony' 'Nike' 'Nokia' 'Benetton' 'IBM'. It is also mentioned the use of 'erp' for business organization, shareholders, analysts, merging between companies and regulations such as antitrust. Another topic is sustainability, respect for the environment and globalization.

Below the list of topics:
[entrepreneurial, entrepreneurs, startups, patent, multinationals, innovations]
[mergers, antitrust, shareholders, regulators, analysts, erp, ericsson, failures]
[brands, benetton, nokias, advertising, xerox, sony, nike]
[sustainability, environmental, mncs, sustained, ibm, globalization, telecom]

Word frequency of 2003

From the frequency of the words of all the 76 documents of the year it is not possible to precisely identify the topics, but we can identify some words of interest such as innovation, change, strategy, competitor, firm, market that make us think of innovation and change in firm's strategy to compete in the market.

Topic in LDA model of 2003

Main topics are service and product innovation, therefore an organization of processes in order to obtain an advantage in the market. Becoming sustainable and maintaining high performance. Product differentiation. Approach to technology with the development of new products and services. Search for new employees through HR. Brand innovation and strategic decision.

From this model we extract the following words related to topics:

[brand, Community, innovation, market, strategy, decision, hr, shareholder, profit, approach, technology, venture, product, differentiator, performance, sustainability, innovation, organization, research, process, market, service, product, cost]

Coherence Score: 0.33949067820001244

Topic in LSA model of 2003

The LSA score is lower than the LDA, so the goodness of the results is inferred. Despite this, it is still possible to identify topics relating the innovation of companies to be able to compete in the market. Research of customer satisfaction in relation to the products and services offered. Identify the customers. Strategic changes related to energy. Technological developments to be able to compete. Below the list of words that represent the topic:
Coherence : 0.2583372093414241

[company, customer, business, product, innovation, market, customer, satisfaction, product, innovation, identity, customer, venture, theory, shareholder, stakeholder, experience, energy, venture, change, development, extension, experience, technology, competitor, change, energy, strategy]

Topic in GTD model of 2003

With the GTD method we have this results.

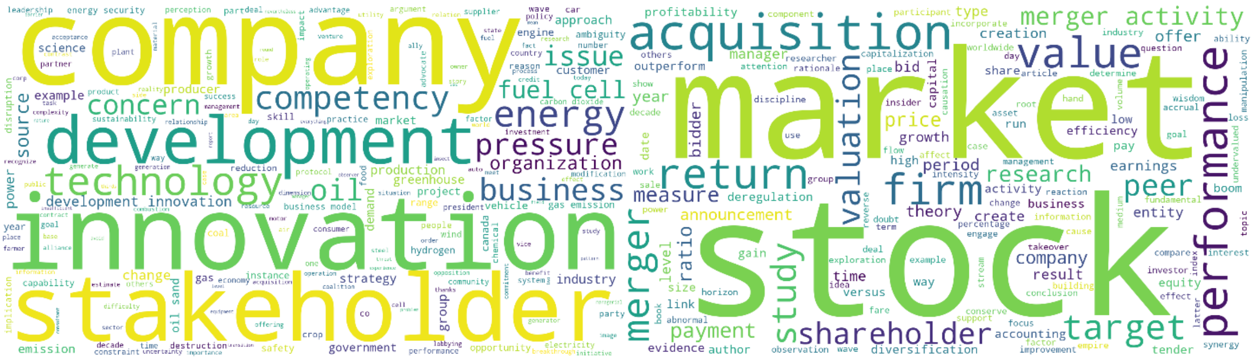
Big-cluster:

knowledge-enhancing, business-process, human-resources, employee-opinion, experience-innovation

Mini-Cluster of :

Polluting gases
Computer companies
Car companies
Media (TV, radio...)
Casinos, Gamblers and Betting
Apple and Ipod
Printers, scanners and ink
Microscope and electrons
Arteries, cardiology, veins (medical field)

Topic in STD model of 2003



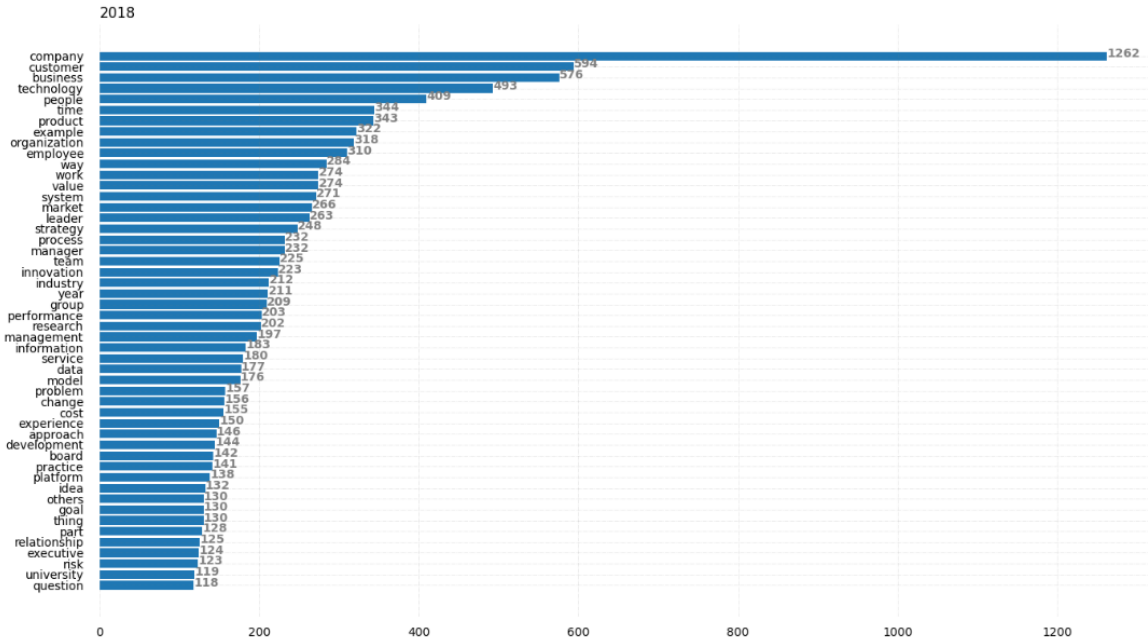
STD confirms our analysis highlighting the sustainable development innovation (SDI) of business processes, and therefore the reduction of energy and oil consumption, while as regards the other file, we can infer that it talks about market, stock, firms and therefore acquisitions and mergers of companies.

Top2Vec from 2010 to 2019

In this decade is still present the topic of competitors, globalization, innovation. Important multinationals such as huawei, telecom, unilever, 'toyota', 'googles' are mentioned. Consortia, founders and collaborations are keywords for these years.
We start talking about the “Agile” method and algorithm development. We are moving more and more towards technology, and therefore the salient topics are robots, AI, automation, cybersecurity, computers, smartphones, digitization, and predictive analysis ('forecasting'), benchmarking.
Together with social networks, including the main Twitter, we talk about 'influencers', market, followers, 'interviewees', collaborations between companies and 'influencers' to do 'advertising' towards 'audience'.

Below the list of topics:
[competitors, innovating, firms, huawei, globalization, telecom]
[coworkers, unilever, innovate, ethical]
[innovate, collaboration, agile, consortium, founders, algorithms, wireless, executives, develop, technologies, sponsors]
[ecosystem, toyota, innovate]
[robots, ai, automated, algorithms, cybersecurity, computers, googles, machines, smartphones, agile, digitization, predictive, analytics]
[analytics metrics, forecasting data analyzed, digitization, benchmarking, scientists]
[influencers marketers, consumer followers, interviewees, advertising twitter collaborative, audience]
[agile, developers]

Word frequency of 2018



From the histogram we get some of the most important words such as technology, product, innovation, information, words closely related to what is the topic of the decade, as well as other words such as manager, team, relationship so we can think about the encouragement of employees and the trust within the team.

Topics in LDA model of 2018

Analysing the files of 2018 with the LDA we can found that the main topic mainly covered is business and corporate management. It can be deduced that there is a focus on internal communication in the teams and to the latter are added different topics, including the issue of citizen mobility, the recruitment and career opportunities of students, the technology concerning cybersecurity and finally the gender discrimination in the workplace, suffered by women.

Below the list of topics:
[communication, feedback, strategy, employee, manager, policy, cybersecurity, service, relationship, mobility, city, car, connectivity, vehicleopportunity, career, ecosystem, cio, partner, assumption, woman, student, position, mensituation, ethicbehavior]
Coherence Score: 0.331403249472908

Topics in LSA model of 2018

Through the LSA algorithm, results are similar to LDA despite the fact that the Coherence parameter is lower. In fact, we deduce that among the main topics there is the corporate reorganization of leaders and groups. In addition, we find arguments regarding technology, security, and the world of blockchain and finally the role of women.

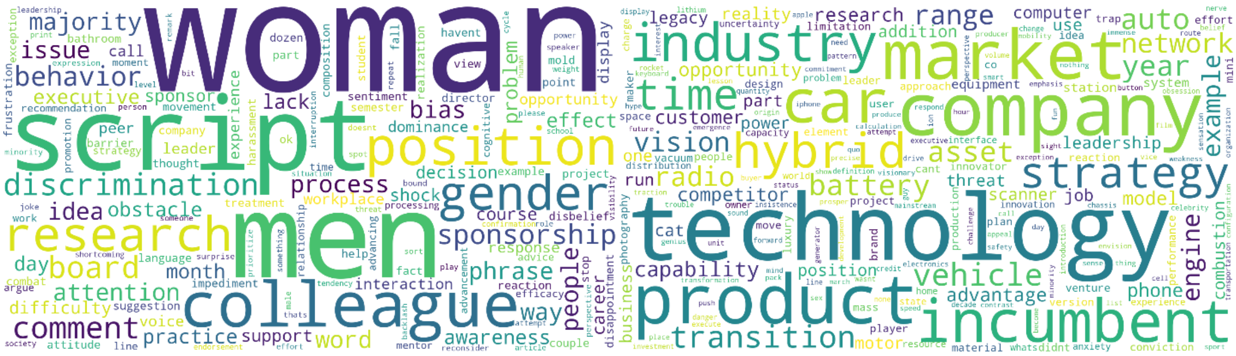
Below the list of topics:
Coherence : 0.2997813660410923
[employee, team, woman, feedback, slack, mobility, city, car-connectivity, vehicle, technology, showroom, woman, cybersecurity, security, network, student, blockchain]

Topics in GTD model of 2018

Analysing the year 2018 through a GTD clustering algorithm reveals different results. In fact, we find a main cluster in which the focus is once again aimed at business improvement by looking for new ideas and marketing strategies. In addition to it, we also find associated other small clusters such as telephony, metals, cars, giants in the IT field and electronic components.

Below the list of words that represent the topic:
[understand, attention, change, decision-making, idea, opportunity desire,complex-need, human-level, interest, know-how]

Topics in STD model of 2018



Sampling two files among those present in 2018 with the tag cloud method, we note that the deductions made previously in the annual analyses are confirmed. The tag cloud in fact highlights all those terms that make us think about topics concerning women, technology, companies, business and its strategies.

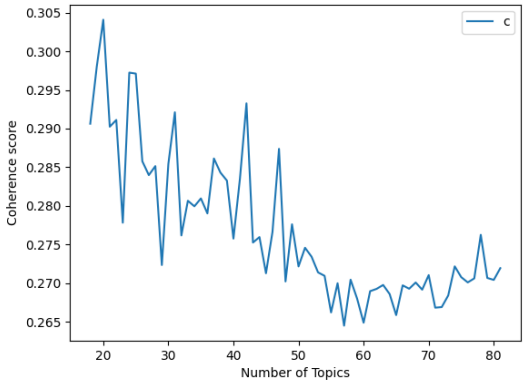
Top2Vec from 2020 to 2022

Artificial intelligence, research, volunteering, 'cybersecurity', 'blockchain', pandemic and sustainability become the most important topics that have led to a further revolution for companies. Other two interesting words are KPIs ESG. The first one is the performance indicator and by setting it the company can make smart business decisions about the direction of all current projects.
The other one stands for Environmental, Social and Governance and refers to three central factors in measuring the sustainability of an investment.

Below the list of topics:
[kpis, cybersecurity, blockchain, agileanalytics, algorithms, consumer, managing, generate, suppliers, develop, automation, developedtechnological, pandemic, esg, sustainable, predictive]
[ecosystem, startups, ai, volunteering, researchers]

Final consideration and Future development

Analysing the results, we see that Top2Vec gave us a great overview over the decade but we can't ask it more than this, because it currently not working with a low number of documents.
LSA and LDA gave us very similar results that are strictly correlated to the number of topics asked to the model to be found. This parameter was set by training the model with different values and selecting the best one. Like in the picture below, the best result for coherence is achieved with number of topics equal to 20.



The GTD, instead performs better when the number of documents is not too large while the STD even if it gave us a clear result, it is limited to the single document analysis. For this reason, we opted out to use together these algorithms.

Further improvements on our works could be two possible changes.
The first observation is that by comparing all the documents of a single year at the same time, the most frequent words are always the same, as well as those dealing with business, but since this information that represents the macro topic is already in our possession, we can think of eliminating them giving, perhaps, more priority to the subtopics we want to find.
The second observation concerns our GTD algorithm, as it will always find a main topic made up of hundreds of words and peripheral topics made up of a few units of words. When the main topic tends to be too large, one might think of recursively applying the GTD algorithm within the most important topic. This would allow us to make the topic smaller and probably find something more detailed and easier to interpret.

Code

Below we will find snippets of our software code.

```
if __name__ == "__main__":

    if (len(sys.argv) == 4):
        arg_from_command_line = True
    else:
        arg_from_command_line = False

    while 1:

        if (arg_from_command_line == False):
            choose = input('Insert:\n\n'
                           '- LDA -> Digit "lda" to execute this text mining method.\n\n'
                           '- LSA -> Digit "lsa" to execute this text mining method.\n\n'
                           '- TOP_2_VEC -> Digit "top2vec" to execute this text mining method.\n\n'
                           '- WORD_CLOUD -> Digit "wordcloud" to execute this text mining method.\n\n'
                           '- GTD -> Digit "gtd-cluster" to execute this text mining method in a year.\n\n'
                           '- STD -> Digit "std" to execute this text mining method for each file in a year. \n\n'
                           '- EXIT -> Digit "exit" to exit the program.\n\n')
        else:
            choose = str(sys.argv[1])
```

The main program, here we run the menu of our application where we can choose which algorithm run for the documents analysis.

```
def parallelized_function(file):  # use to clear and prepare source text of each file
    if file.endswith(".txt"):
        input_file = open(f"data/{file}", encoding="utf8")
        file_text = input_file.read()
        file_text = tp.remove_whitespace(file_text)  # remove double space
        file_text = tp.tokenization(file_text)  # tokenization
        file_text = tp.stopword_removing(file_text)  # remove stopwords
        file_text = tp.pos_tagging(file_text)  # add tags to words
        file_text = tp.lemmatization(file_text)  # lemmatize words

        logger.info("Subprocess for file -> [%s]", file)

        return file_text
```

Here we have a method that allows us to do pre-processing of documents in parallel, optimizing computation times.

```
def densityArea(docs, title, year):
    for i in range(0, len(docs)):

        clear_results = [list(dict.fromkeys(docs[i]))]
        tot_vectors = {}
        for word in clear_results[0]:
            tot_vectors[str(word)] = ew.get_embedding(str(word))
        if not os.path.exists(f"output/{year}/STD/{title[i][:-4]}"):
            os.makedirs(f"output/{year}/STD/{title[i][:-4]}")
        pca.pca_clustering_3D(list(tot_vectors.values()), list(tot_vectors.keys()),
                              f"output/{year}/STD/{title[i][:-4]}/InitialCluster__nWords_{len(tot_vectors)}")
        transformer = RobustScaler(quantile_range=(0, 75.0))
        transformer.fit(list(tot_vectors.values()))
        centroid_ = transformer.center_
        centroid_ = np.array([centroid_])
        distance_vector = {}
        for j in range(0, len(tot_vectors) - 1):
            dist = cosine_similarity(centroid_, np.array([list(tot_vectors.values())[j]]))
            distance_vector[list(tot_vectors.keys())[j]] = dist[0][0]
        distance_vector = sorted(distance_vector.items(), key=operator.itemgetter(1),
                                 reverse=True)
```

This is one of the main features of the std algorithm which analyses documents individually by plotting a cluster of words.

```
def choice_clustering_method(tot_vectors, year, file_text):
    bigClusters = db.DBSCAN_Topic(tot_vectors, year)
    words = []
    for i in range(0, len(file_text)):
        for t in range(0, len(bigClusters)):
            for j in range(0, len(bigClusters[t])):
                if bigClusters[t][j] == file_text[i]:
                    words.append(bigClusters[t][j])
    return
```

Here an analysis is made using a clustering method in which all the documents of a given year are analysed.

```
def top_2_vec(documents):
    # training model with all documents.  speed : fast-learn / learn / deep-learn
    model = Top2Vec(documents, embedding_model='universal-sentence-encoder', speed='deep-learn', workers=4) # with pre-trained embedding model
    #model = Top2Vec(documents, speed='deep-learn', workers=4) # without pre-trained embedding model
    numberOfTopis = model.get_num_topics()
    print(f"Number of topics : {numberOfTopis}")
    topic_words, word_scores, topic_nums = model.get_topics(numberOfTopis)
    print(f"Topic words : {topic_words}\nWord scores : {word_scores}\nTopic numbers : {topic_nums}")
    for topic in topic_nums:
        model.generate_topic_wordcloud(topic)
    return topic_words, word_scores, topic_nums
```

This function analyses the top2vec by grouping documents by decades.

GitHub repository of our code: https://github.com/mgiorgi13/MITopics

References

https://blog.osservatori.net/it_it/intelligenza-artificiale-funzionamento-applicazioni
https://monkeylearn.com/blog/nlp-ai/#:~:text=Natural%20Language%20Processing%20(NLP)%20is,spell%20check%2C%20or%20topic%20classification
https://en.wikipedia.org/wiki/Topic_model
https://monkeylearn.com/topic-analysis/#:~:text=to%20Topic%20Analysis-,What%20Is%20Topic%20Analysis%3F,individual%20text's%20topic%20or%20theme.
https://github.com/ddangelov/Top2Vec
https://arxiv.org/abs/2008.09470