MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université de Paris

Projet TER

Field : Machine Learning for Data science

# Biomedical corpus constitution and exploitation


Mohamed GHARBI — *Tutor :* M. Lazhar Labiod

Nassim ELKEFIF — *Tutor :* Mme. Sevrine AFFELDT


1er MARS 2021 — 18 MAI 2021

# Thanks

For this modest work we would like to thank :

— First of all our parents for their unwavering and unlimited support and also given me what I need to get where I am today and are the source of my inspiration and courage.

— Our tutors « **LAZHAR LABIOD** and **SEVRINE AFFELDT** » and also miss *RAFIKA BOUTALBI* for their help and support as well as for the relevance of their remarks and answers.

*And for all of this a thank you is not enough.*

**Abstract**

Nowadays, with the growth of data In different fields like “ educational, economic, political, commercial and medical”, it has become very hard and almost impossible to explore and treat them even with the use of computer science. And the use of this big amount of data after it’s processing and cleaning can be useful in a lot of making decisions and prediction especially in the medical field. So the project that we have is a subject of the constitution and exploitation of biomedical corpus " through the pubmed plateform and by doing this it must respect these clauses :

- Study and application of the different text processing and machine learning methods
- The interpretation of the different results and compare them the visualization of the different results

**Key words** : NLP, text mining, python dash, python, machine learning , deep learning, pubmed, articles

# Summary

# List of figures

# Introduction

Nowadays, with the growth of data In different fields like " educational, economic, political, commercial and medical", it has become very hard and almost impossible to explore and treat them even with the use of computer science in these ones we can see that the problem steel remains due to the volume and the various forms of data such as " articles, videos, images, databases and ... etc.".
In the last years, when talking about articles and exploitation of data the first thing that comes to mind is the medicine, as we can see the medical field is provided with a lot of platforms and web apps the contain a considerable amount of information like " Pubmed, Medscape, WebMD, MedicineNet ... etc.", that touch the different fields of Medicine like, its history, diseases, , symptoms, treatments, the degree of danger ... etc.

**So what is pubmed ?[1]**

PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health both globally and personally.
The PubMed database contains more than 32 million citations and abstracts of biomedical literature. It does not include full text journal articles ; however, links to the full text are often present when available from other sources, such as the publisher's website or PubMed Central (PMC).
Available to the public online since 1996, PubMed was developed and is maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH).

**And going furthermore, what is PMC ?[2]**

PubMed Central (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). In keeping with NLM's legislative mandate to collect and preserve the biomedical literature, PMC serves as a digital counterpart to NLM's extensive print journal collection.
PMC was developed and is managed by NLM's National Center for Biotechnology Information (NCBI).

**And lastly, how are the articles in pubmed ?**

All of this amount of data and resources has made it very difficult for data scientists and researchers to come with relevant results in decision making and prediction, which made birth of a new era in computer science.
The technologies that are used in this kind of problems are mainly involved in the pre-processing of data, data science especially machine learning and deep learning and their rich libraries and methods.
The project that was affiliated to us consists of "the conception and creation of Dashboard for the constitution and analysis of a biomedical corpus" and its main objective is viewing and applying the different methods of pre-processing and the extraction of

sub-themes of medical articles from PubMed.

This thesis is organized as follows :

**Chapter 1 : Constitution of the biomedical corpus**

This is the theoretical part where we present the basic concepts related to the constitution of the biomedical corpus by addressing its different aspects and the different methods and algorithms that are used.

**Chapter 2 : Our proposal for a biomedical corpus creation**

This chapter, we present the modeling of our solution. We start by presenting the algorithms and methods used in the different phases

**Chapter 3 : Interface and experiments**

This chapter is divided in two parts. First, we present the different tools used for the realization. Then, we present the work carried out through the realization of the various stages of the solution, which are illustrated by screen captures of the application.

Finally, we will conclude this thesis with a Conclusion in which we will underline all the important things that we have seen throughout this project.

Chapter 1

# Constitution of the biomedical corpus

## 1.1 Introduction

The implication of computer science more precisely in data science and artificial intelligence in medicine has become essential and a must-apply to have a better vision of the future and a precise decisions. In this chapter we will talk about data science and the different methods that are used to constitute a biomedical corpus by pointing out :

- Data science and Medicine
- What is a corpus
- Text pre-processing

## 1.2 Data science and Medicine [3]

In the last several years, Data science had a big impact in healthcare by touching several fields In it such as :

- X-Ray images
- Genomics
- Drug discovery
- Predictive analytics

And all that by using different methods of machine learning , deep learning and artificial intelligence by the exploitation of a big amount of processed data.

We can see that the data that is collected is raw and it needs processing, so that's where it comes the use of data science, NLP by creating organized and exploitable corpuses.

Before going furthermore, let's have a look on this topic by headlining some of its big titles :

***what is Data science?***

Data Science draws on statistics and computer science to address questions in various domains, such as biology, education, physics, business, linguistics, or, in addition to quantitative skills and domain expertise.

***What is Machine Learning? [4]***

is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a modelfrom example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

In machine learning there are 3 types of problems and tasks, that are presented as follows :

*Supervised learning :*

The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

*Unsupervised learning :*

No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.

*Reinforcement learning :*

A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

***What is Deep Learning?***

Deep learning is a technique of artificial intelligence derived from machine learning, it is based on artificial neural networks with a more complex and deep structure.

***What is NLP?***

In the world we live in there are a lot of ways of communication like : through speaking and listening, making gestures, hand signals and writing and reading. In this matter we will focus of the last one which is text processing and exploitation by using NLP. Natural language processing (NLP) are methods that enable computers to analyze, process and derive meaning from human discourse.[6]

## 1.3 What is a corpus?

After pointing out some of the fields of data science and data processing in a theoretical way, let's point out what is a corpus and its characteristics.

**What is a corpus?**

CORPUS (13c : from Latin corpus body. The plural is usually corpora) (1) A collection of texts, especially if complete and self-contained : the corpus of Anglo-Saxon verse. (2) Plural also corpuses. In linguistics and lexicography, a body of texts, utterances or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analyzed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordance programs. Corpus linguistics studies

data in any such corpus.

(cf. McArthur, Tom "Corpus" , in : McArthur, Tom (ed.) 1992. The Oxford Companion to the English Language. Oxford, 265-266).

### 1.3.1 What are its characteristics ?

Modern corpuses tend to have four main characteristics, which are the following :

- Sampling and representativeness : contains a whole variety of texts, not only an individual text or author.
- Finite size : finite number of words
- Machine-readable form.
- A standard reference : a widely available corpus for other researchers.

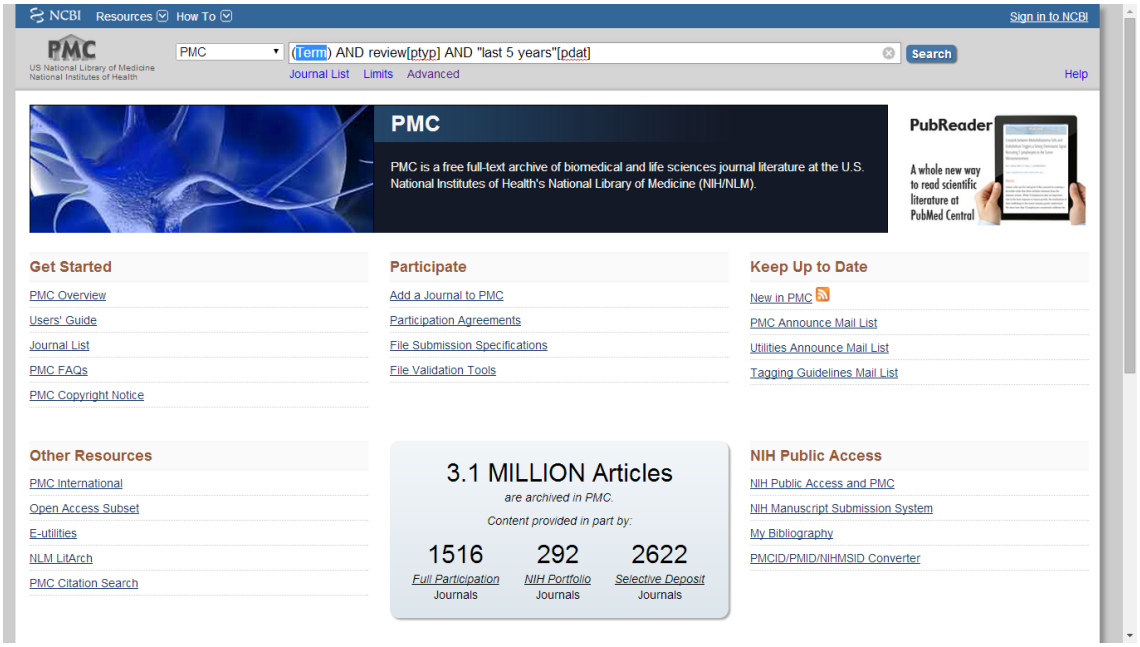



FIGURE 1.1 – Data source.

## 1.4 Document-Term Matrix

[COUR1 NADIF] This Matrix is an unsupervised learning technique, widely used to group documents based on their content so that documents within a cluster are semantically coherent and deal with the same topic.

### 1.4.1 TF-IDF technique

[10] TF-IDF is a mixture of two different words, first TF, which stands for Term Frequency, and then IDF for Inverse Document Frequency :

- TF : used for measuring the frequency of terms in a document by the number of occurrence of a term divided by the number of total terms.

- IDF : used to assign lower weights to frequent terms and greater weight to infrequent terms. researchers.

So TF-IDF is nothing but the multiplication of TF * IDF.

## 1.5 Text pre-processing

[7] By the end of collecting texts and data, and after the establishment of a well-defined purpose of analysis, the next step is the preparation and the pre-processing of data. Note that the quality of the results depends highly on the data pre-processing step and can affect negatively such as positively the results of any text data related analytic task.
Analytics are about working with instances of data, in text analytics, these instances are documents, each document is made up of words, and each word of characters. All the documents combined define a corpus (document collection).
Assuming that words are the focus of the analysis, the pre-processing step is always primordial to assure a result of a good quality, it takes as an input a full text or group of words that constitute a document and returns groups of single words called tokens. This process include Tokenization, standardization, removal of stop words and Stemming and lemmatization. **Step 1 : Tokenization**
This step is used to separate the words and form a group of tokens.
**Step 2 : Standardization**
a set of transformations applied to each token, such as lower case conversion, removal of numbers, punctuation, special characters and extra spaces.
**Step 3 : Stop words removal**
stop words are common words and words that have no real added value to the analysis.
**Step 4 : Stemming and Lemmatization**
Stemming is converting the word into its root by removing its suffix, while lemmatization is grouping together all the forms of a word so they can form a single word.
For the purpose of an example, we use the following input : “ MY flowers are white with yellow spots. ”

| / | OUTPUT |
|---|---|
| **Tokenization** | [MY] [flowers] [are] [white] [with] [yellow] [spots] [.] |
| **Standardization** | [my] [flowers] [are] [white] [with] [yellow] [spots] |
| **Stop words removal** | [flowers] [white] [yellow] [spots] |
| **Stemming and Lemmatization** | [flower] [white] [yellow] [spot] |

Figure 1.2 – Example for text processing.

## 1.6 Used machine Learning methods

### 1.6.1 K-Means

K-means is an unsupervised non-hierarchical clustering algorithm. It allows to group in K distinct clusters the observations of the dataset. Thus, similar data can be found in the same cluster. Moreover, an observation can only be found in one cluster at a time (exclusivity of membership). The same observation cannot belong to two different clusters. [5]

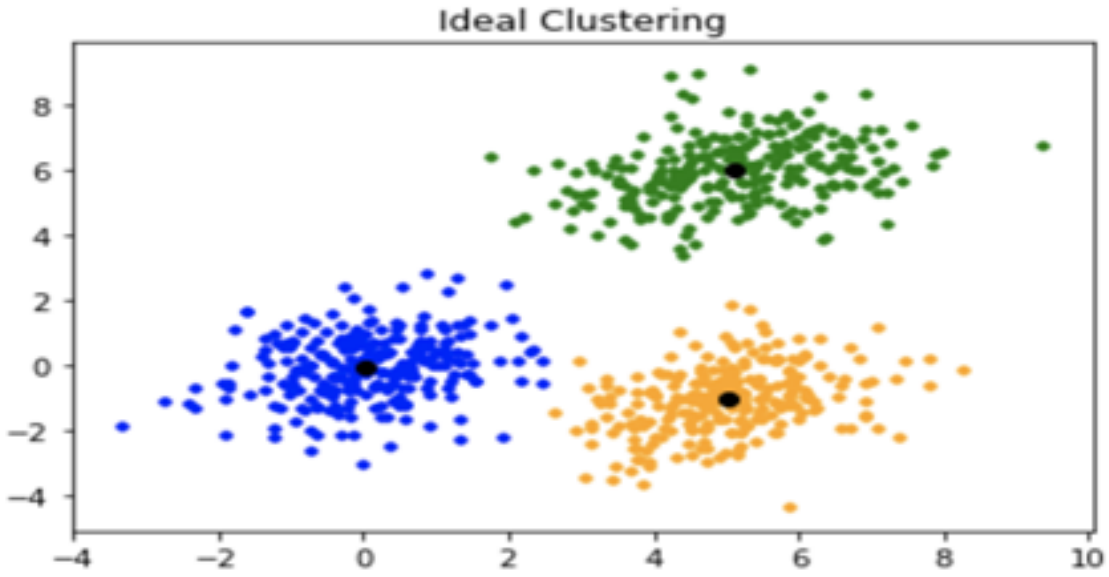



FIGURE 1.3 – Example for kmeans classification.

### 1.6.2 PCA

[COUR1 NADIF] Principal Component Analysis ( PCA ) : is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values on linearly uncorrelated variables called principal components.

### 1.6.3 T-SNE

[11] distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised technique that minimizes the divergence between two distributions : a distribution that measures pairwise similarities on the input objects, and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding.
T-SNE is mostly used thanks to the capacity of facilitating the visualization of high-dimensional data.

### 1.6.4 LDA

The Latent Dirichlet Allocation (LDA) algorithm is an unsupervised learning algorithm most often used to discover a specific number of topics shared by the user through documents in a text corpus.
LDA assumes that documents consist of words that define topics and relates documents to a list of topics by assigning each word in a document to different topics. This relationship is based on conditional probability estimates, as illustrated in the following figure :

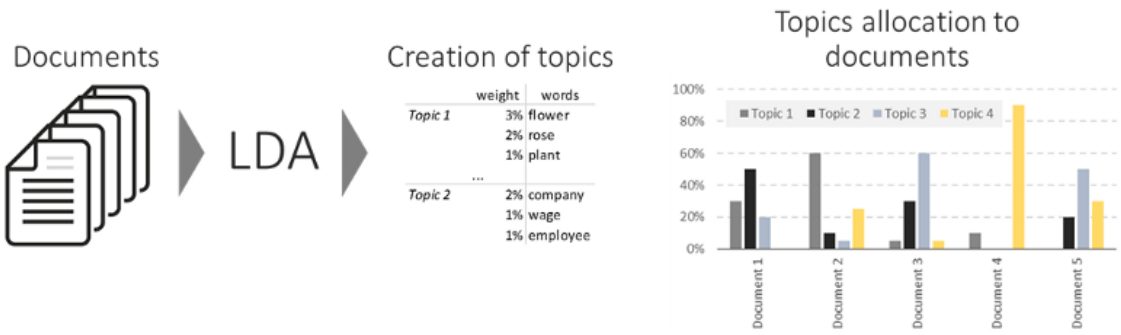



FIGURE 1.4 – LDA processing.

### 1.6.5 Co-Clustering

Also known as Bi-clustering, is a data mining technique that simultaneously cluster rows and columns into partitions, so that a pair consisting of a row cluster and a column cluster is a co-cluster.
[8] In recent years, several implementations of co-clustering algorithms have been developed such as biclust, bicat and bibench, but in the other hand, not many of these implementations can handle co-occurrence matrices such as document-term matrix mentioned earlier. For this reason, the suitable package for these kind of matrices in general, and for our case in particular is the CoClust Package.
Co-clustering algorithms used for co-occurrence matrices can be set into different categories :

- **Spectral methods**

Treat the input data matrix as a bipartite graph between documents and words, and approximate the normalized cut of this graph using a real relaxation.

- **Model-based methods**

Currently two model-based co-clustering methods are implemented, the first one relies on the latent block models (LBM) and the second one relies on the stochastic block model and the LBM.

- **Matrix factorization based methods**

Only used for document clustering based on non-negative matrix factorization.

- **Information Theoretic based methods**

used to co-cluster two-way contingency tables (a two-way is a display of counts for two categorical variables in which the rows represent one variable and the columns represent a second variable [9]).

- **modularity based methods**

allows the co-clustering of binary or contingency matrices by maximizing an adapted version of the modularity measure traditionally used for networks..

#### 1.6.5.1 CoClust Package

Coclust is a python package that includes implementations of co-clustering algorithms. Coclust provides three clustering algorithms, the first two are CoclustMod and CoclustSpecMod , they are part of the block-diagonal co-clustering algorithms family (modularity-based methods) ; these two algorithms have several advantages, some of which are the following :

- They produce descriptions of the resulting document clusters since each document cluster is directly associated with one term cluster.
- They can adapt to various kinds of matrices.

The third algorithm is CoclustInfo, part of the information theoretic based methods, its main advantages are speed of convergence and scalability.

### 1.6.6 Conclusion

In this chapter, we have seen the theoretical part of data science at it's fields by pointing out some specific ones such us : " data constitution, preprocessing and data exploitation".

Chapter 2

# Our proposal for the biomedical corpus creation

## 2.1 Introduction

After, seeing the different sides of data science, now we will tackle another aspect in our study which is the conception of the solution, in which we will talk about the steps and methods that have been used in order to realize this project.
In our conception we followed several steps that are :

- Data extraction
- Text processing
- Generation of sub-themes

## 2.2 Data Extraction

The first step of our project is the extraction of data from PubMed, in this action we aimed to have two types of information namely :

- High-level information

In this type of information we will extract the full text of the different articles.

- Low-level information

In this level of information, the extraction we be mainly pointed in : “title, date of publication, authors, abstract … etc.”.

### 2.2.1 Extraction of High-level information

This kind of information is simply extracted from PubMed and it can be retrieved from all forms of articles.

### 2.2.2 Extraction of Low-level information

The extraction of the high-level information or full articles is only from PMC types which are free articles and they come as a PDF, XML and HTML pages.

The data extraction is based on this prototype algorithm.

**Algorithm 1** Data Extraction

**Result:** All the information : authors, abstracts, ids, date of publication and articles

word, number-of-articles : initialization

step 1 : import necessary libraries

step 2 : calling search method by giving the word and the number

**while** *list of ids not empty* **do**

  **if** *type = PMC " which is a free articles"* **then**

    step 3 : extract all the data : articles as well

  **else**

    Step 4 : extract all the data except the articles

  **end**

**end**

## 2.3 Text pre-processing

After the extraction of data, comes the pre-processing step, in which we will use different algorithms, the following algorithm is the demonstration of this step.

**Algorithm 2** Pre-processing

**Result:** text processed

text : initialization

step 1 : import necessary libraries

step 2 : Extract all the abstract from the corpus

**while** *list of abstracts not empty* **do**

  step 3 : Delete punctuation

  step 4 : tokenize the text

  step 5 : delete the stop words

  step 6 : lemmatize the words

  step 7 : delete the words < 2

  step 8 : convert into strings

**end**

After the application of these different algorithms, now we will constitute our corpus by applying the following steps.

**Algorithm 3** Corpus constitution

**Result:** text processed

lists of data : initialization

step 1 : import the necessary libraries

we will extract only the articles that has an abstract different from "null".

**while** *lists not empty and list of abstracts != "null"* **do**
| step 2 : create a dataframe by combining data.

**end**

## 2.4 Corpus exploitation

now that the corpus is cleaned and ready to be used, we will apply several machine learning algorithms in order to study its use. In our case, we applied three methods :" K-Means, LDA, Co-clustering" that helped us to come up with different and meaningful topics that will be explained by prototype scripts.

### 2.4.1 K-Means application

**Algorithm 4** k-Means function

**Result:** plots using TSNE and PCA

Victorized text : initialization

step 1 : import the necessary libraries

step 2 : apply the K-Means method with the right parameters.

step 3 : apply the K-Means transform method.

step 4 : apply the PCA and T-SNE methods. these methods are used to visualise using two different types.

### 2.4.2 LDA application

**Algorithm 5** LDA function

**Result:** A list of the Different topics and there weight

Victorized text : initialization

step 1 : import the necessary libraries

step 2 : apply the LDA method with the right parameters.

step 3 : apply the LDA transform method.

### 2.4.3 Co-Clustering application

**Algorithm 6** Coclustering function

**Result:** the result depends on the chosen method

Victorized text, numberclusters : initialization

step 1 : import the necessary libraries

**if** *method == "specMod" or method = "mod"* **then**

step 2.1 : apply the chosen co-clustering method with the right parameters

return : the top key words

**end**

## 2.5 Results Interpretation

In this section we will discuss the different results given by the application of the mentioned methods. In our case we have extracted 1000 article from the word **"FEVER"**, our study is based on the abstracts and that after applying the pre-processing.

### 2.5.1 K-Means interpretation

In order to have the right number of clusters, we have used the elbow method, which returned number of clusters " k" : 6.

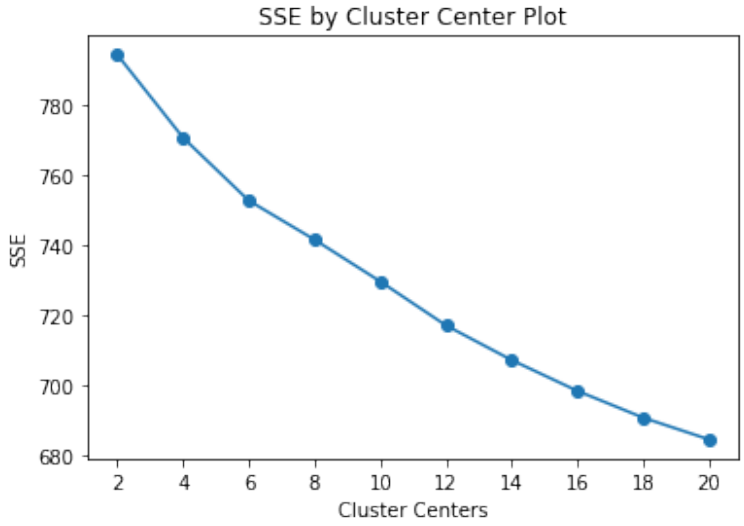



FIGURE 2.1 – Elbow method.

After having the right number of clusters, we applied the k-means method by using *" minibatchkmeans"*.

Based on the results bellow we can see that there are some groups that are formed and others are not, but globally the k-means method is not reliable in this type of data exploration.

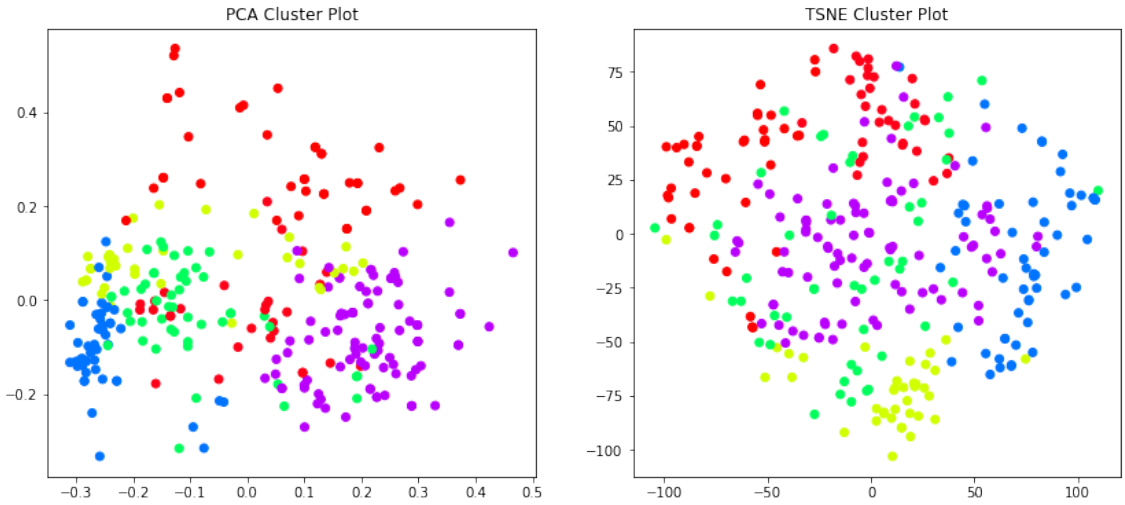



FIGURE 2.2 – Plot with PCA and T-SNE.

By going further on our analysis we listed the 5 top key words for each Cluster as shown on the "figure 2.3" in which there is a repetition of some words.

```
Cluster 0
attribute,stringelement,label,parent,children

Cluster 1
body,rat,response,temperature,induce

Cluster 2
infection,study,clinical,group,patients

Cluster 3
drug,disease,patient,syndrome,case

Cluster 4
patients,attribute,label,stringelement,nlmcategory

Cluster 5
stringelement,label,attribute,test,malaria
```

FIGURE 2.3 – Top key words for each cluster " k-means".

### 2.5.2 LDA interpretation

The purpose of this method is to sort out with different topics, in our case we chose the number of clusters randomly which is 6. The result given in this figure is the 5 top key words of each topic and there weight.

| | Topic 0 words | Topic 0 weights | Topic 1 words | Topic 1 weights | Topic 2 words | Topic 2 weights | Topic 3 words | Topic 3 weights | Topic 4 words | Topic 4 weights | Topic 5 words | Topic 5 weights |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | induce | 17.2 | parent | 23.7 | patients | 85.9 | pattern | 0.2 | case | 18.2 | patients | 0.9 |
| 1 | response | 8.7 | children | 15.3 | label | 78.0 | fevers | 0.2 | malaria | 17.8 | illness | 0.9 |
| 2 | rat | 8.5 | health | 9.9 | attribute | 77.8 | confidence | 0.2 | virus | 15.4 | severe | 0.9 |
| 3 | brain | 8.1 | seek | 9.9 | stringelement | 77.8 | day | 0.2 | test | 15.1 | adult | 0.6 |
| 4 | body | 6.0 | management | 9.8 | nlmcategory | 64.8 | secondary | 0.2 | patients | 12.8 | rapid | 0.6 |

FIGURE 2.4 – Top key words for each Topic " LDA".

The application of the LDA and from the topics that are sorted, we can see that in some of them we can distinguish sub-themes based on their top words such as :

The Second topic : precaution.

The last topic : The propagation of the disease.

### 2.5.3 Co-clustering interpretation

As we have seen the application of the LDA was random based on the number of clusters, but in the co-clustering we have the possibility to to make it automatic by applying the modularity method.

Based on the figure above, we can see that the best modularity is 0.215 and it is attained for 5 co-clusters.

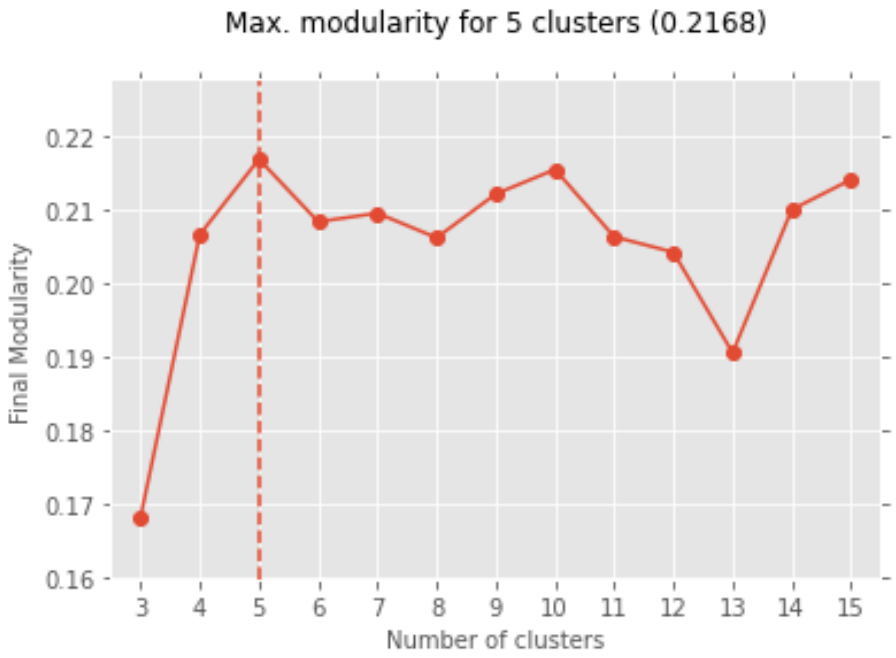



FIGURE 2.5 – Number of clusters for Co-clustering.

As it's shown from the figure "2.5", we notice different results from the application of LDA algorithm and also the construction of 5 clusters of different sizes, each one of the clusters is represented by the most frequent 5 terms (top 5 terms).
Based on these terms and what they represent in term of meanings, we can approximately attribute thematic (sub-topics) for each cluster :

#### 2.5.3.1 Mod method

- Cluster 1 : The first cluster with the terms " Infection, Test, Infections, Virus, Cause ", and from which we can distinguish the sub-topic of " **Causes of infections**".
- Cluster 3 : The third cluster with the terms "Malaria, treatment, diseases, follow, criteria" , and for which we can attribute the sub-topic of " **Diseases that come with fever, and their followed treatment.**".
- Cluster 4 : The fourth cluster with the terms " Case, Patient, Syndrome, Diagnosis, Present", and from which we can distinguish the sub-topic of " **Patients diagnosis and their syndromes**"
- Cluster 5 : The fifth cluster with the terms "Patients, Group, Induce, Temperature, Response" and for which we can attribute the sub-topic of " **General responses of fever patients**"

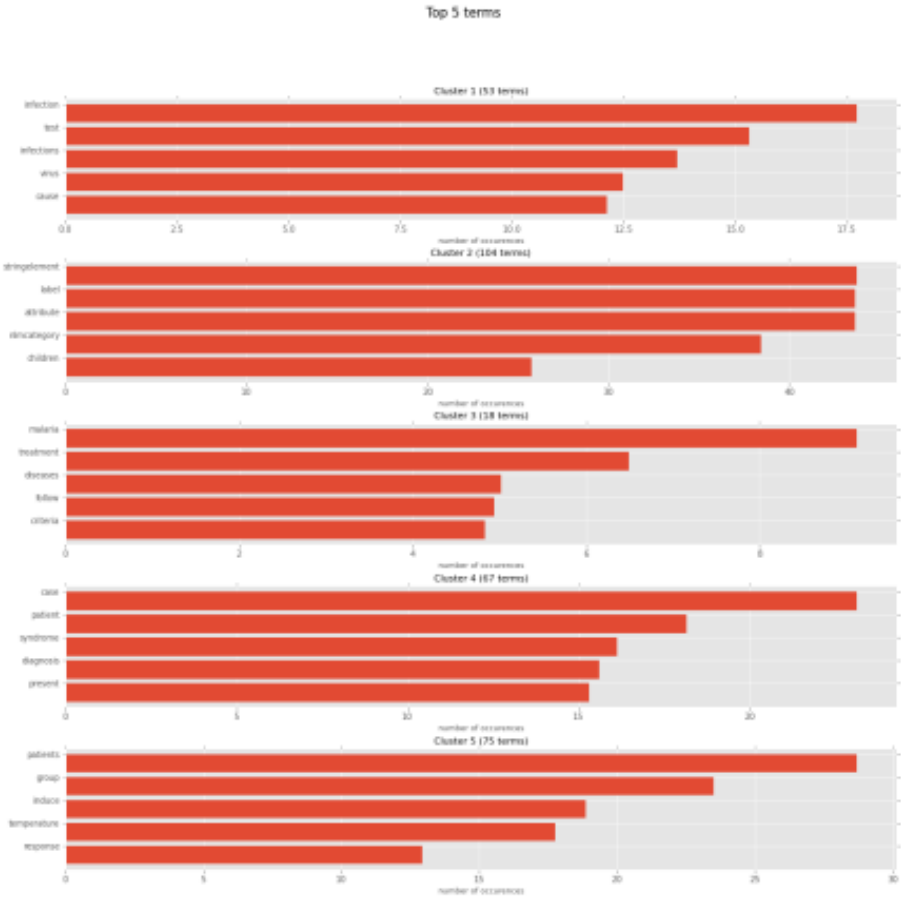



FIGURE 2.6 – Top words of Mod method

#### 2.5.3.2 SPECMOD method

We notice that the results of the SpecMOD method and the top words of their clusters are almost the same with the MOD method results, such as cluster 4 of MOD with cluster 2 of SpecMOD, and cluster 1 of MOD with cluster 3 of SpecMOD.

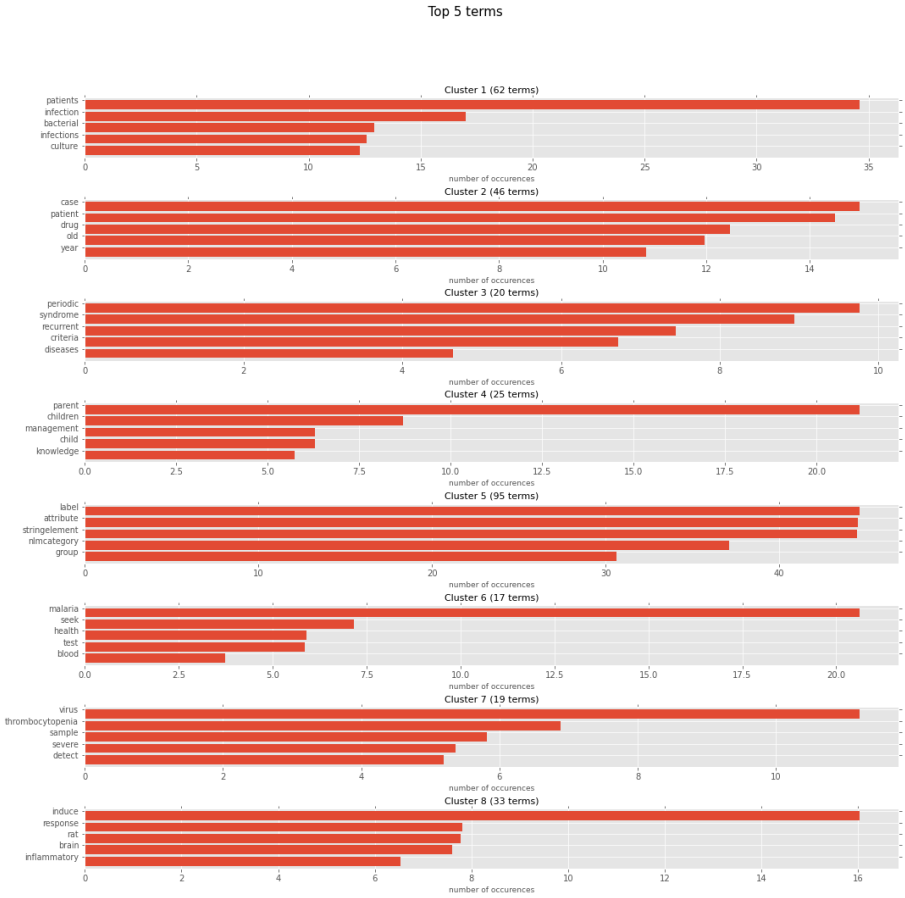



FIGURE 2.7 – Top words of specMod method.

### 2.5.4 Results conclusion

From these results we can confirm that :

- The k-means method is not suitable for this type of data and study, meanwhile, " LDA and Co-clustering " gave different results but with a meaning "we can see that the sub-themes are coherent with the subject" .
- almost all of sub-topics are directly related to the main searched topic of the example which is “fever”, but in particular the first cluster which has the most indicative top terms compared to other clusters.

## 2.6 Conclusion

In this chapter, we have presented the different algorithmic prototypes used in our designand also another part concerning the interpretation of the results obtained on a predefined example.

The next chapter concerns the implementation of our system.

Chapter 3

# Our Interface and experiments

## 3.1 Introduction

After highlighting all the aspects and analyses of our solution, we start its implementation. For this, it is essential to use a number of tools and to set up execution environments.
This chapter presents the tools used, the technical architecture of the developed application, as well as the stages of realization of our solution.

## 3.2 Architecture of the application

The application that we developed is composed of three primal services like :

- Back-end : The back-end contains all the functions and the methods used to fetch data from PubMed.
- Front-end : The front-end is the part where we have the visualisation of the results of the different methods pointing out the interaction between the user and the dashboard.
- Pub-Med : Pub-Med is the source of the data.

In this type of information we will extract the full text of the different articles. The figure bellow shows the connections of the three services.

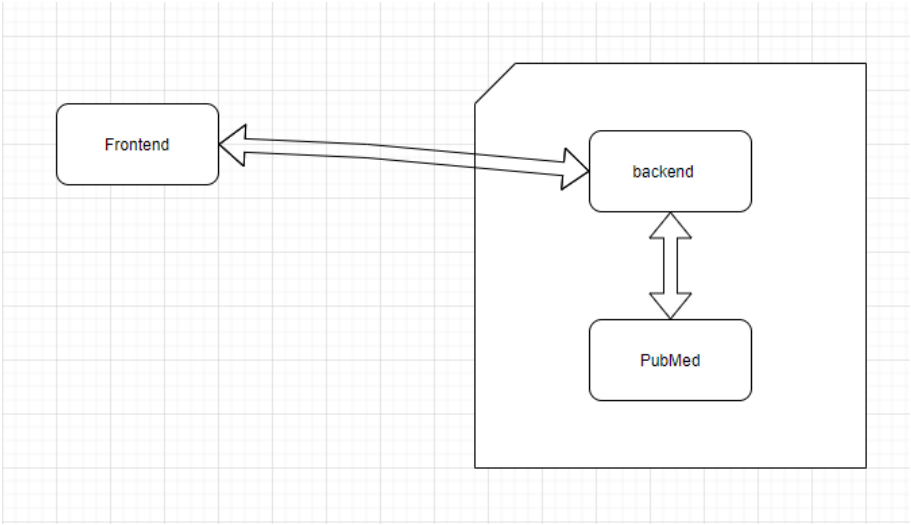



FIGURE 3.1 – Architecture of the Application

## 3.3 Some examples

After seeing the architecture of our application, now we will

### 3.3.1 Corpus creation

As mentioned on the previous chapter, the constitution of our corpus is based on a word and a number of articles given by the user as shown in the figure bellow

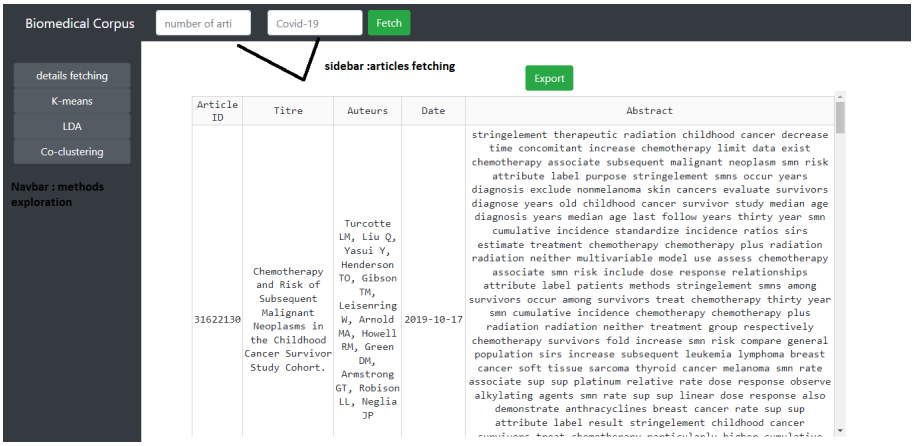



FIGURE 3.2 – different sides of our application

And also we the visualisation of some of the methods as follows :

K-Means :

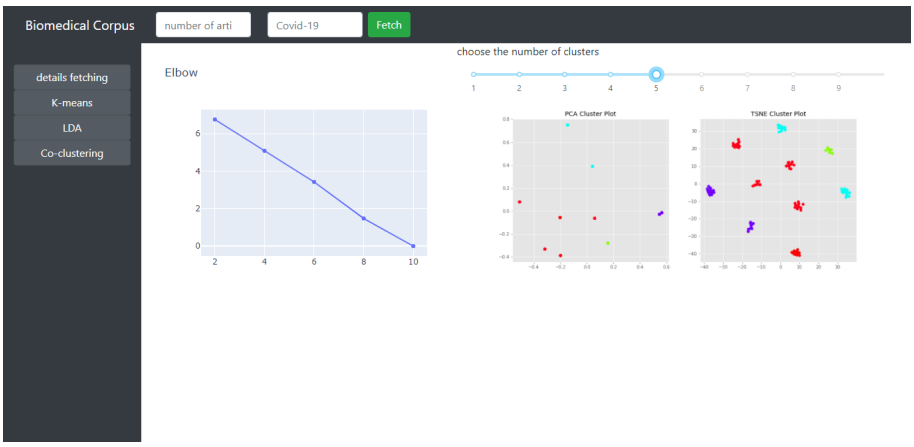
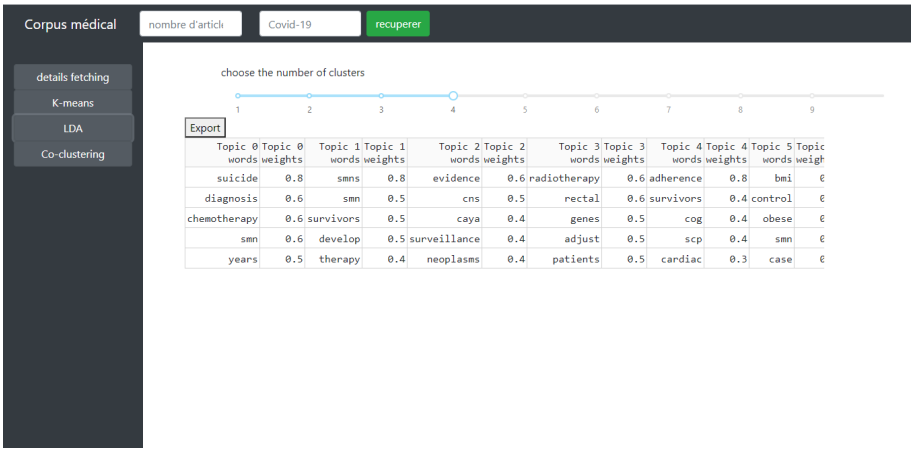



FIGURE 3.3 – Architecture of the Application

LDA method :




| Topic 0 words | Topic 0 weights | Topic 1 words | Topic 1 weights | Topic 2 words | Topic 2 weights | Topic 3 words | Topic 3 weights | Topic 4 words | Topic 4 weights | Topic 5 words | Topic weigh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| suicide | 0.8 | smns | 0.8 | evidence | 0.6 | radiotherapy | 0.6 | adherence | 0.8 | bmi | 0 |
| diagnosis | 0.6 | smn | 0.5 | cns | 0.5 | rectal | 0.6 | survivors | 0.4 | control | 0 |
| chemotherapy | 0.6 | survivors | 0.5 | caya | 0.4 | genes | 0.5 | cog | 0.4 | obese | 0 |
| smn | 0.6 | develop | 0.5 | surveillance | 0.4 | adjust | 0.5 | scp | 0.4 | smn | 0 |
| years | 0.5 | therapy | 0.4 | neoplasms | 0.4 | patients | 0.5 | cardiac | 0.3 | case | 0 |

FIGURE 3.4 – Architecture of the Application

# Conclusion

The work presented in this thesis concerns the realization of a dashboard for the constitution of a biomedical corpus. It required the study of constituting a theoretical framework allowing us to better understand the following steps.
Then, we carried out a study concerning the extraction of data from pubmed as well as a study of the different methods used to make the text exploitable. Then, we came to a decision that was the exploration of some topic modeling methods in order to come out with subtopics.
Once the requirements definition was completed, we moved on to the solution design which includes : the constitution of our corpus and its cleaning in order to make it exploitable. as well as the design of the interface of our dashboard.
The last part of our project was the realization of the solution. We started by choosing and learning the tools and technologies that allowed us to realize our project.
It is important to remember that a computer science project is never finished, so as perspectives to our work we have :

- In our project we focused our exploitation and text mining in the abstracts and we can go further by exploiting the full articles.
- The dynamism of the Dashboard.
- And going further we can also computerize the extraction of the sub-themes.
- The exploitation of some other co-clustering and topic modelling approaches like the spherical k-means.

# References


[1] pubmed official site :
https://pubmed.ncbi.nlm.nih.gov/about/

[2] pubmed official site introduction :
https://www.ncbi.nlm.nih.gov/pmc/about/intro/

[3] https://data-flair.training/blogs/data-science-in-healthcare/

[4] INTRODUCTION TO MACHINE LEARNING
*Authors : Alex Smola and S.V.N. Vishwanathan*

[5] Classification des Images Multispectrales et Hyperspectrales basée sur l'Apprentissage Profond utilisant les Réseaux de Neurones Convolutifs,
*Authors : Nassim BOUHADOUF, Toufik KERDJOU, University of Algiers 1 Benyoucef BENKHEDDA.*

[6] Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context
*authors : Jeffrey C. Oliver - University of Arizona*

[7] Text Preprocessing. Advances in Analytics and Data Science, 45–59. doi :10.1007/978-3-319-95663-3-4
*authors : Anandarajan, M., Hill, C., Nolan, T. (2018).*

[8] CoClust : A Python Package for Co-clustering. 2018. ffhal-01804528f
*Authors : François Role, Stanislas Morbieu, Mohamed Nadif.*

[9] 11.1.2 - Two-Way Contingency Table | STAT 200 (psu.edu)

[10] Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents :
*Shahzad Qaiser, Ramsha Ali*

[11] Accelerating t-SNE using Tree-Based Algorithms : Laurens van der Maaten, Pattern Recognition and Bioinformatics Group Delft University of Technology Mekelweg 4, 2628 CD Delft, The Netherlands.