Project COVID-19 OPEN RESEARCH DATASET CHALLENGE (CORD-19)

Morgane Govone
Gaetan Jacquet

Introduction

Several tasks :

- Obtain the best answer
- Improve these results

Dataset :

- Documents
- Queries

Introduction

Summary :

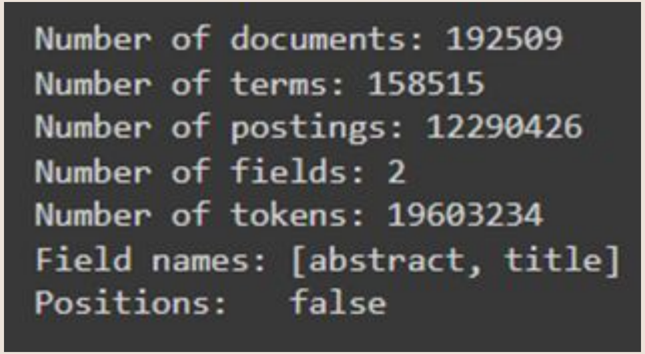
1. Description of our collection
2. Description of our search engine
3. Evaluation of results
4. Model improvement

↓

Have the most relevant documents

I. Description of our collection

Our collection



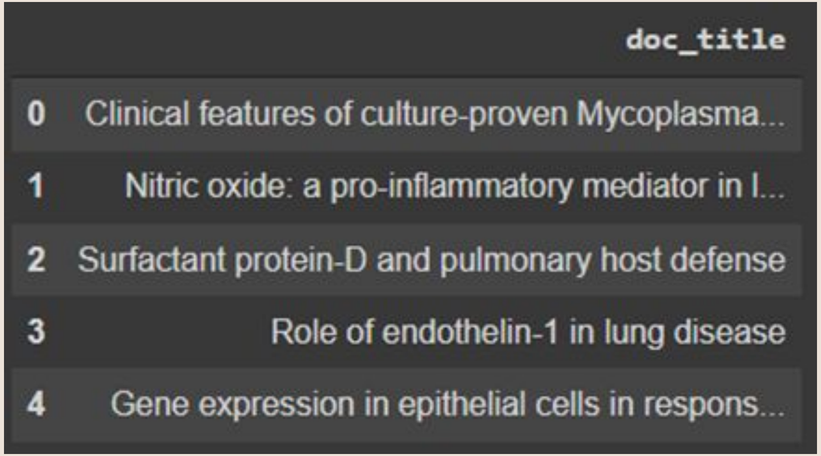
```
Number of documents: 192509
Number of terms: 158515
Number of postings: 12290426
Number of fields: 2
Number of tokens: 19603234
Field names: [abstract, title]
Positions:   false
```

All the documents are scientific papers in english

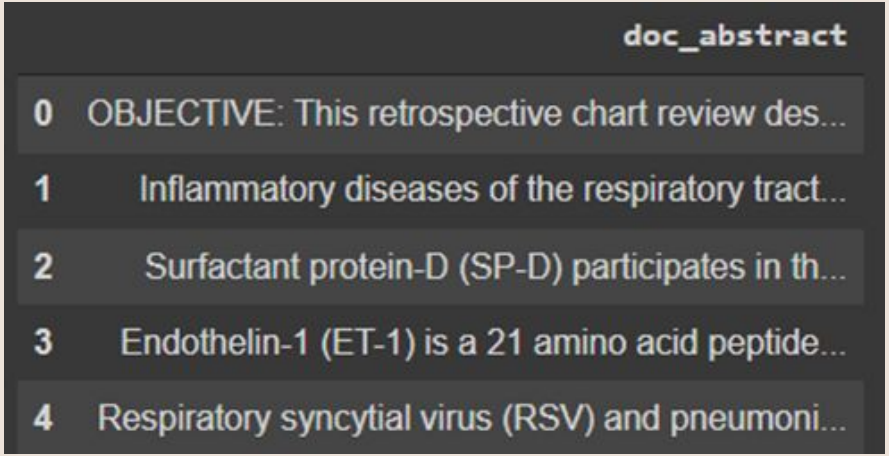
The language and the type of paper are important

I. Description of our collection

DataFrame of our collection of data



	doc_title
0	Clinical features of culture-proven Mycoplasma...
1	Nitric oxide: a pro-inflammatory mediator in l...
2	Surfactant protein-D and pulmonary host defense
3	Role of endothelin-1 in lung disease
4	Gene expression in epithelial cells in respons...



	doc_abstract
0	OBJECTIVE: This retrospective chart review des...
1	Inflammatory diseases of the respiratory tract...
2	Surfactant protein-D (SP-D) participates in th...
3	Endothelin-1 (ET-1) is a 21 amino acid peptide...
4	Respiratory syncytial virus (RSV) and pneumoni...

I. Description of our collection

DataFrame of our collection of data

Most of these documents are related to covid 19.

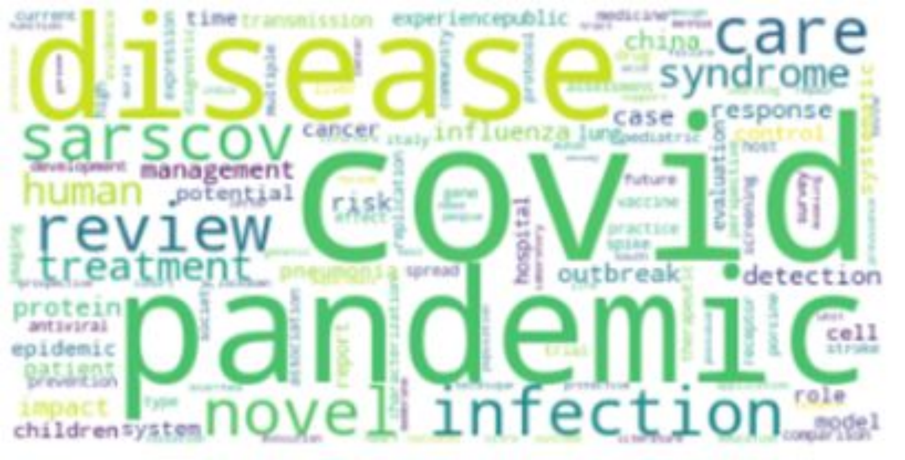
Problem for splitting the collection into two part :

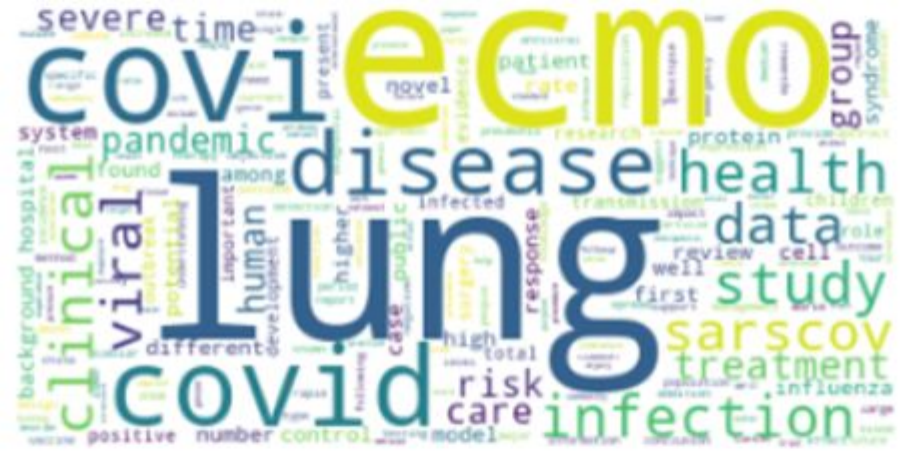
- Miss relevant documents
- What form of this notion are we looking for

→ We have chosen to keep all the document of the collection

I. Description of our collection

__Words clouds of our collection of data__

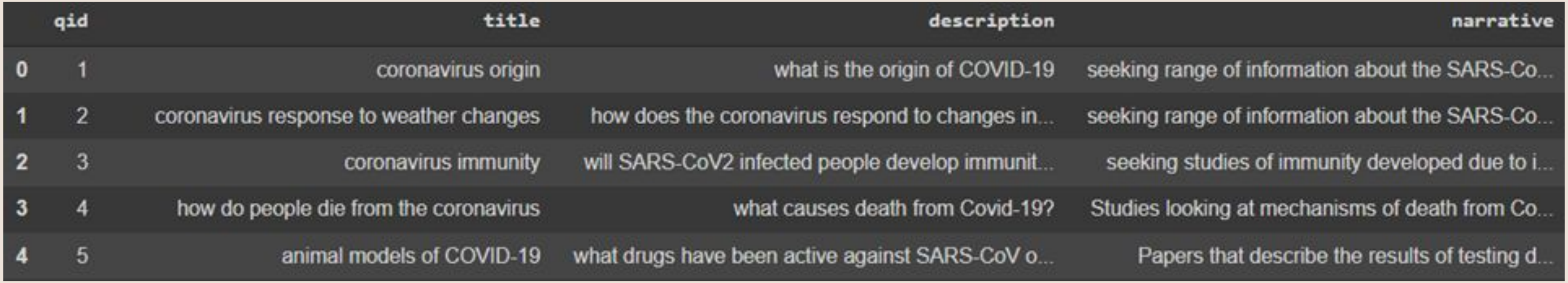




I. Description of our collection

DataFrame of our collection of query

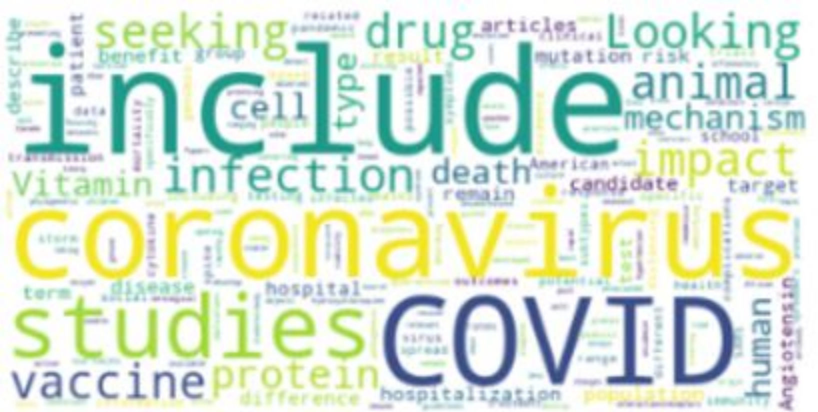
Problem of terminology is still present



	qid	title	description	narrative
0	1	coronavirus origin	what is the origin of COVID-19	seeking range of information about the SARS-Co...
1	2	coronavirus response to weather changes	how does the coronavirus respond to changes in...	seeking range of information about the SARS-Co...
2	3	coronavirus immunity	will SARS-CoV2 infected people develop immunit...	seeking studies of immunity developed due to i...
3	4	how do people die from the coronavirus	what causes death from Covid-19?	Studies looking at mechanisms of death from Co...
4	5	animal models of COVID-19	what drugs have been active against SARS-CoV o...	Papers that describe the results of testing d...

I. Description of our collection

Words clouds of our collection of data



II. Description of our search engine

Search engine :

Simple machine divided into two part

Queries

Documents

II. Description of our search engine

Query

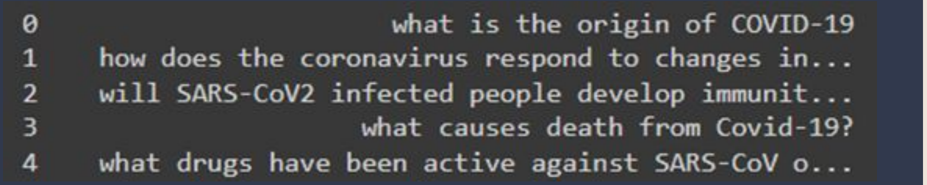
process_query(type_query) :

- pre_process(dataset)
- token_per_sent(dataset) : *tokenize the dataset per query*

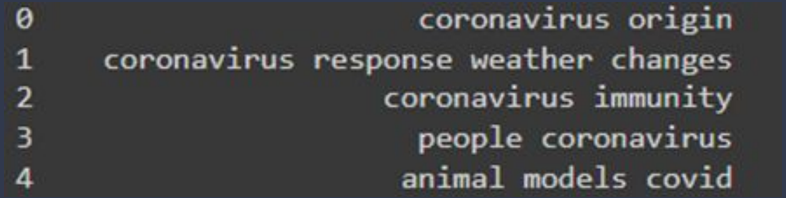
Document

- pre_process(dataset) : *contains the majority of the pre processing steps*
- word(dataset) : *makes it possible to apply tokenization on our dataset and thus remove the words not relevant for our request*

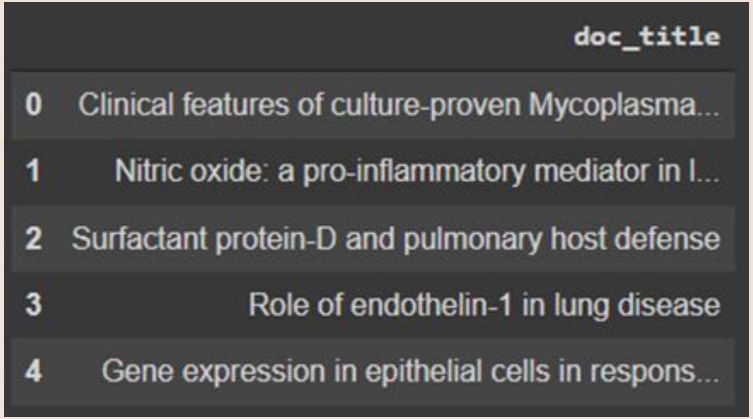
II. Description of our search engine



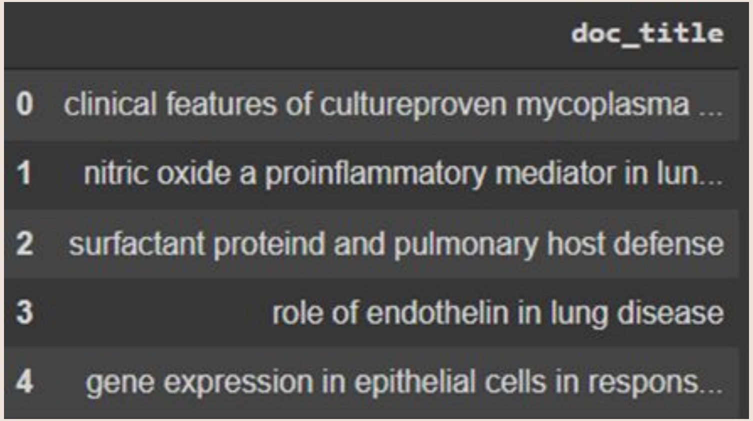
```
0                              what is the origin of COVID-19
1      how does the coronavirus respond to changes in...
2      will SARS-CoV2 infected people develop immunit...
3                            what causes death from Covid-19?
4      what drugs have been active against SARS-CoV o...
```



```
0                          coronavirus origin
1      coronavirus response weather changes
2                       coronavirus immunity
3                          people coronavirus
4                         animal models covid
```



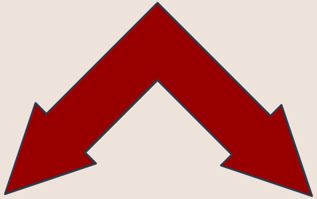
	doc_title
0	Clinical features of culture-proven Mycoplasma...
1	Nitric oxide: a pro-inflammatory mediator in l...
2	Surfactant protein-D and pulmonary host defense
3	Role of endothelin-1 in lung disease
4	Gene expression in epithelial cells in respons...



	doc_title
0	clinical features of cultureproven mycoplasma ...
1	nitric oxide a proinflammatory mediator in lun...
2	surfactant proteind and pulmonary host defense
3	role of endothelin in lung disease
4	gene expression in epithelial cells in respons...

III. Evaluation of results

Objective : evaluate if the documents we get match well and are relevant to our queries



Measures :

- P@5 (precision at 5)
- P@10 (precision at 10)
- NDCG (Normalized Discounted Cumulative Gain)
- Reciprocal Rank
- MAP (Mean Average Precision).

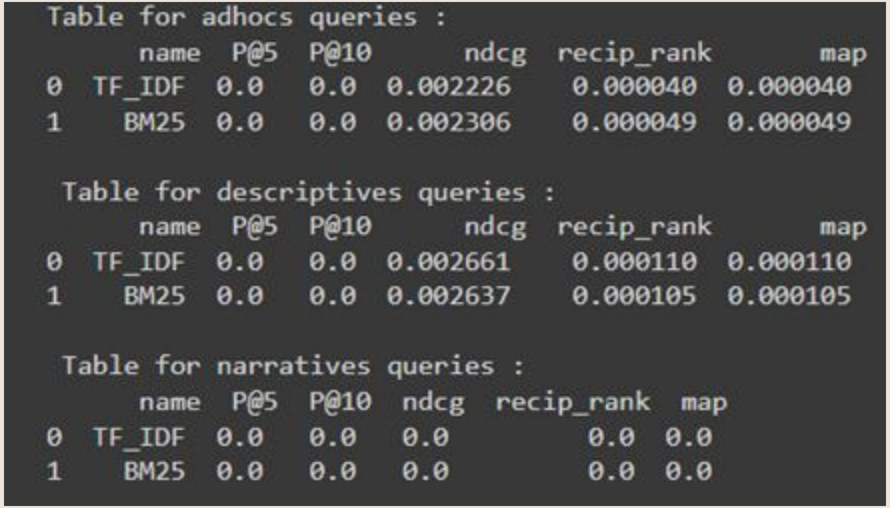
Weighting model :

- TF__IDF
- BM25

III. Evaluation of results

First results

The evaluation run and return the correct response format but not the results we expect



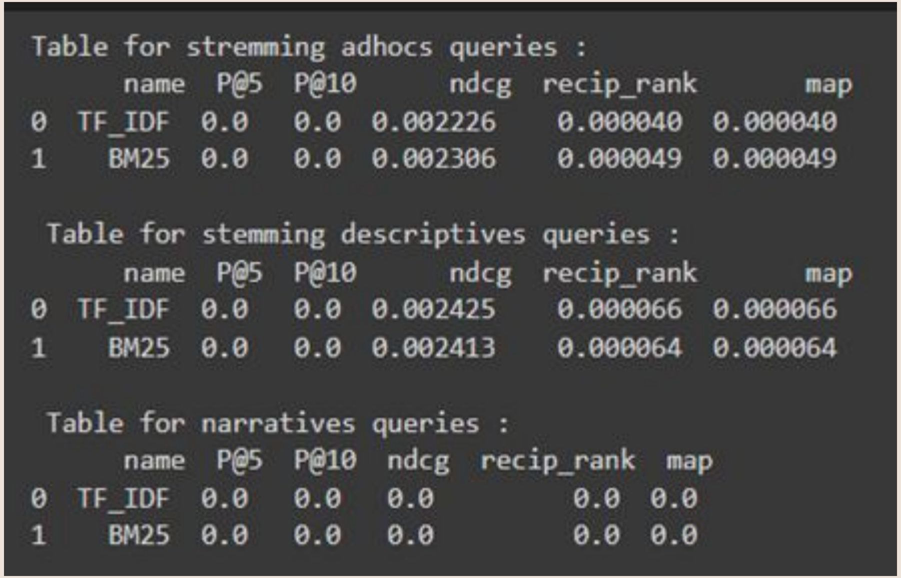
```
Table for adhocs queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002226    0.000040  0.000040
1    BM25  0.0   0.0  0.002306    0.000049  0.000049

 Table for descriptives queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002661    0.000110  0.000110
1    BM25  0.0   0.0  0.002637    0.000105  0.000105

 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```

III. Evaluation of results

Second results

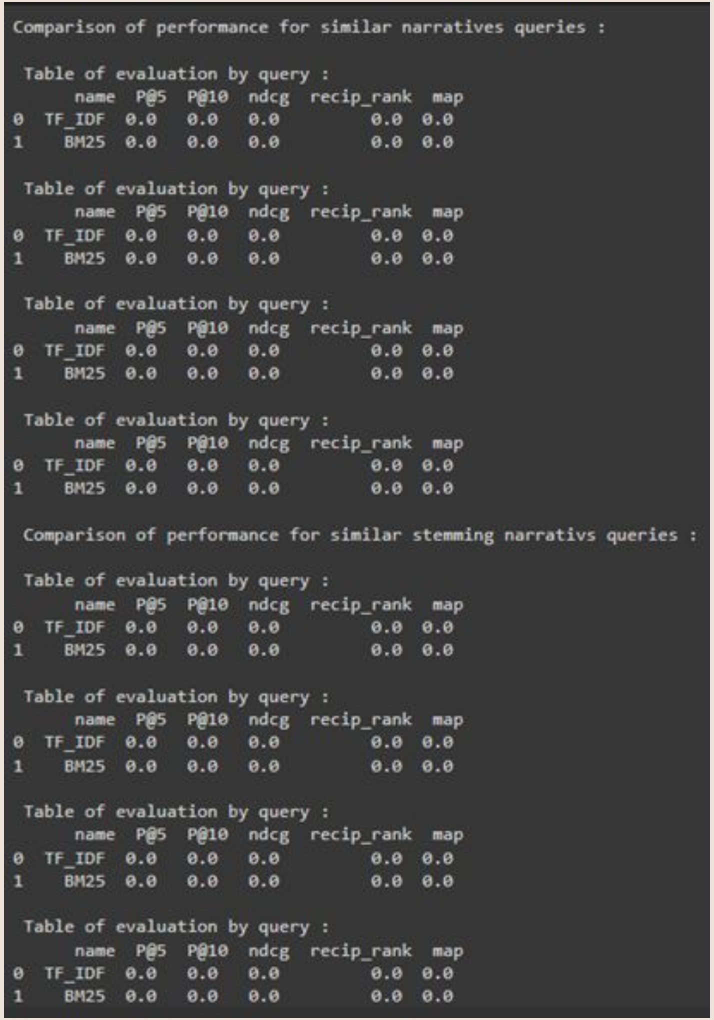
Try to modify our pre-process

Stemming step : have only the root of the words



```
Table for stremming adhocs queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002226    0.000040  0.000040
1    BM25  0.0   0.0  0.002306    0.000049  0.000049

 Table for stemming descriptives queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002425    0.000066  0.000066
1    BM25  0.0   0.0  0.002413    0.000064  0.000064

 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```

III. Evaluation of results



```
Comparison of performance for similar narratives queries :

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Comparison of performance for similar stemming narrativs queries :

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table of evaluation by query :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```

Third results

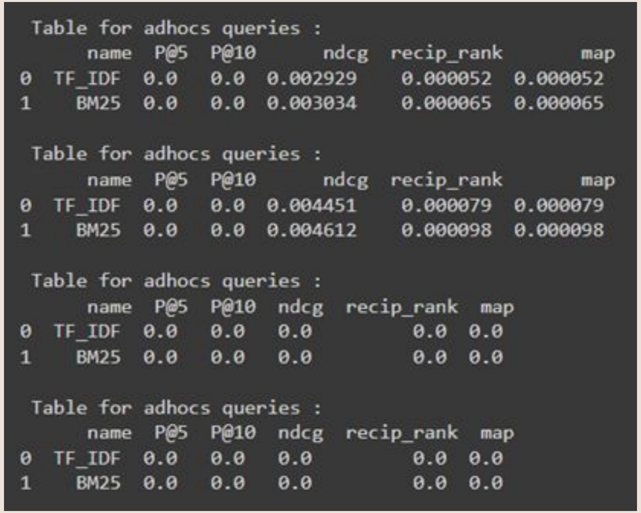
Queries with similar performance could change the evaluation ?

Obtain the value 0 everywhere, maybe because of problem

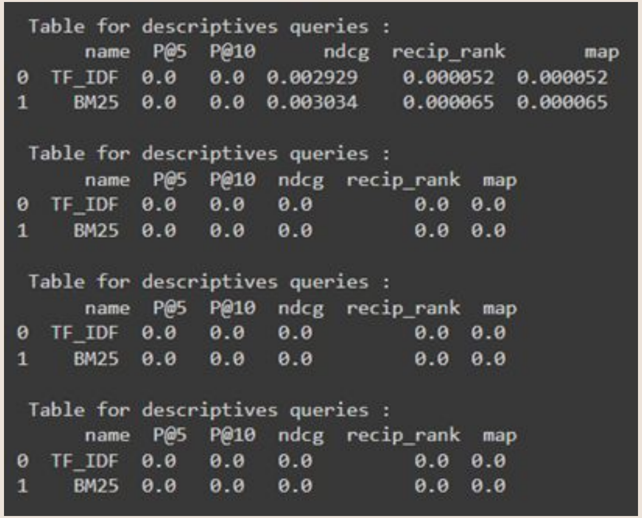
IV. Model improvement

Reducing the number of queries for each of our three types

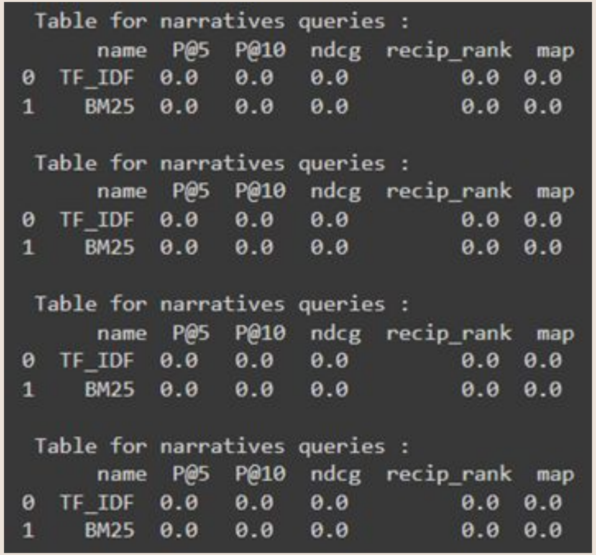
Take only 75% of the queries then 50% and 25% to finish 10%



```
 Table for adhocs queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002929    0.000052  0.000052
1    BM25  0.0   0.0  0.003034    0.000065  0.000065

 Table for adhocs queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.004451    0.000079  0.000079
1    BM25  0.0   0.0  0.004612    0.000098  0.000098

 Table for adhocs queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for adhocs queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```



```
 Table for descriptives queries :
     name  P@5  P@10      ndcg  recip_rank       map
0  TF_IDF  0.0   0.0  0.002929    0.000052  0.000052
1    BM25  0.0   0.0  0.003034    0.000065  0.000065

 Table for descriptives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for descriptives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for descriptives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```



```
 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0

 Table for narratives queries :
     name  P@5  P@10  ndcg  recip_rank  map
0  TF_IDF  0.0   0.0   0.0         0.0  0.0
1    BM25  0.0   0.0   0.0         0.0  0.0
```

V. Conclusion

We can conclude in two ways :

- There really is a problem → we can't really compare or improve our search machine with a great deal of certainty
- There is no problem → we can say that our search engine is really not efficient and only returns random documents

THE END