Information Processing and Retrieval

Part 1 Project Report

Group 08:
Daniele Avolio, ist1111559
Michele Vitale, ist1111558
Luís Dias, ist198557

1 Problem Statement

In this project we handled the task of **summarizing** and **extracing keywords** from a set of documents. In particular, we were handling documents regarding news from the BBC. Our dataset is composed both from the plain text of the news and the corresponding summarization, retrieved using state of the art techniques. Note that in this part of the project we didn't use the summarization of the news for guiding our system, but we only used it for evaluating the performance of it.

The tasks that were conducted can be explored one by one. Let's start by listing them:

- **Indexing**: The creation of a structure that allows to quickly retrieve the documents.
- **Text Summarization**: Given a document, the task is to compute the best set of sentences that are more relevant to the document itself.
- **Keyword Extraction**: Given a document, retrieve the most important words that are present in the document.
- **Evaluation**: Given a set of produced summaries S_p and a set of real summaries S_r, compute the metrics that evaluate the performance of the system.

Moreover, for the task of **Text Summarization** we explored some techniques to improve the performance of the system. Namely, we tried to use **Reciprocal Rank Fusion (RRF)**, a technique that allows to combine the results of different systems in order to improve the performance of the system itself, and **Maximal Marginal Relevance (MMR)**, a techniques that theoretically reduces the redundancy of the produced summaries.

Cose da aggiungere:

- Specificare che se mancano alcuni grafici e dati, sono sul notebook
- Specificareche il numero di documenti sui quali sono stati tirati fuori i grafici. In particolare, RRF fatto su tot documenti, MMR su tot documenti, comparison con BERT su tot documenti. Questo appunto per dire che non è sull'intera collection.

2 Adopted solutions

Each task of the project presented a set of choices or challenges that we had to overcome. This section of the report summarizes briefly those aspects, describing our solution in a non-exhaustive way. To better understand our proposed system, please refer to the code in the ipynb file.

2.1 Indexing

The indexing phase sticks to the Whoosh library implementation, because it was fast and straightforward. The only notable operation is the title removal in documents.
We did not implement any scorer function by scratch, since the one already present in the library are performing working in our use case.

2.2 Text summarization

The general idea was to produce a summarization system based on the *extractive approach*. In particular,the task was fulfilled with 3 different IR systems: BM25 scorer, TF-IDF scorer and a BERT-based system, with further tests on Reciprocal Rank Fusion and Maximal Marginal Relevance.

MB25 and TF-IDF

BM25 and TF-IDF are two models based on the respective metrics that are querying on the inverted index created in the first step. The score of each sentence is related to the length of it and we prefer the top k scoring sentences. We also have the possibility to get the extracted sentences in the exact order of appearance in the initial document.

BERT

Our BERT model is based on a sentence selective approach in which we prefer the sentences that are closer to the document in the latent space representation produced by BERT. This might represent a problem in contexts with long documents, since the BERT token windows is limited to 512. However, we accept that possible information loss for evaluation purposes.

Reciprocal Rank Fusion and Maximal Marginal Relevance

We also explored the RRF and MMR techniques to improve the quality of the summaries. The first one is a way to combine the results of different systems, in our case BM25 and BERT, while the second one is a way to avoid redundancy in the output summaries.
Both of them are limited by important computational times, since the first requires the usage of BERT and the second is an iterative process.

2.3 Keyword extraction

We used the *nltk* library to tag words based on the context that they provide in the sentence. We then selected a few relevant tags, like nouns and pronouns.

Using TF-IDF, we scored each of the selected word to get the k most relevant ones.

2.4 Evaluation

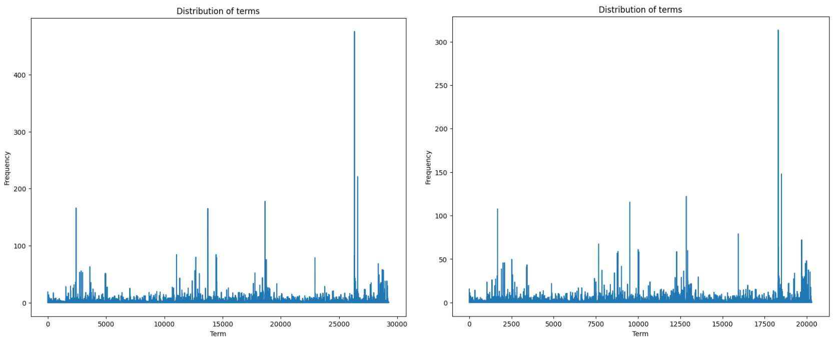
The used metrics in this phase are precision, recall, F1 score and Mean Average Precision (MAP). As required in the project description, we conduced a set of evaluation tests to assert the quality of the system, but the complete collection could not be tested for some specific evaluation tasks due to lack of computational power. More on this can be found in the upcoming section, as well as in the proposed notebook.

3 Proposed questions

In this section, we answer point-to-point to the questions proposed during the project description.

3.1 Describe the corpus D and summaries S. Are terms uniformly distributed regarding TF-IDF?

Fixed the x axis on terms, it can be seen by the figures that the plotted distribution of words remains close to unchanged. The cardinality of terms changes due to the fact that many words are not relevant to the summarization task, so the do not appear on the vocabulary of the summaries S. Also, it is important to note that the y axis has a different scale, due to the smaller cardinality of the summary tokens in respect of the corpus D.



(a) Corpus distribution

(b) Summaries distribution

3.2 How does the summarization system perform for the full collection? And within each category? Any intuition for the observed differences?

The following graphs show the overall results, both on whole dataset or by class. We might say that there are differences among classes, but that should be cautious since to prove it we would require a non-parametric statistic test to assert differences on MAP. We can suppose that eventual differences might be result of the different words distribution among categories,with different keywords vocabulary size.

Category	Documents	Precision	Recall	F1	MAP
business	510	0.642	0.426	0.512	0.859
entertainment	386	0.651	0.452	0.533	0.848
sport	511	0.669	0.482	0.561	0.904
politics	417	0.717	0.381	0.498	0.864
tech	401	0.695	0.347	0.463	0.810
whole_collection	2225	0.675	0.421	0.518	0.992

Table 1: Performance metrics for different categories.

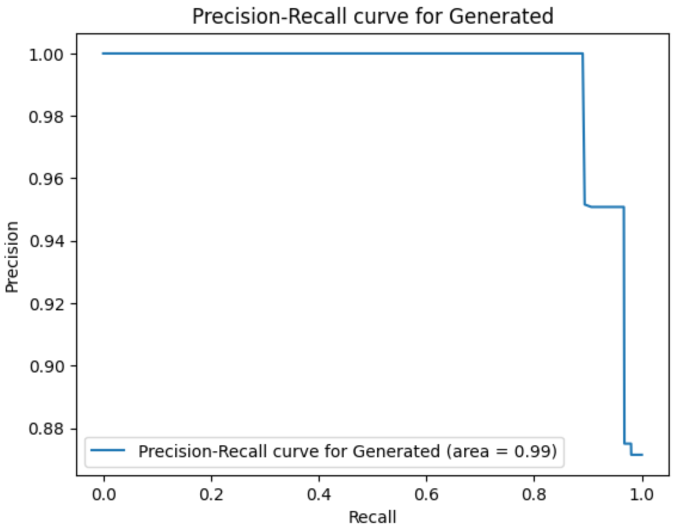


Figure 2: Precision and recall for the whole collection.

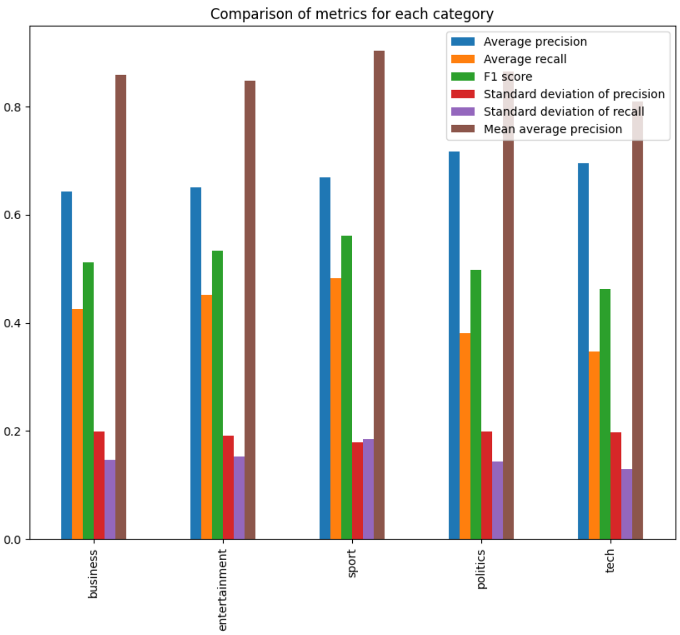


Figure 3: Performances by categories

3.3 How IR models affect summaries? How vector space models compare with language models?

The comparison between BM25 model (that performs similarly to TF-IDF) and BERT model with our algorithm tends to be in favour of this second implementation.

Metrica	BM25	BERT
Average Precision	0.25	0.30
Average Recall	0.14	0.17
F1 Score	0.18	0.22
Standard Deviation of Precision	0.06	0.17
Standard Deviation of Recall	0.04	0.15
Mean Average Precision	0.12	0.86

Table 2: Performance comparison for BM25 and BERT.

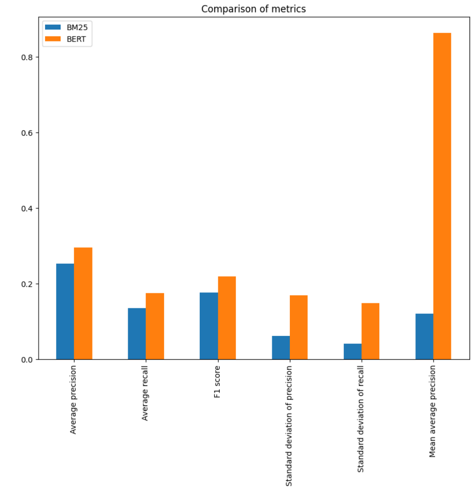


Figure 4: Performance comparison for BM25 and BERT.

3.4 Is Reciprocal Rank Fusion (RRF) useful to aid decisions?

We expected that RRF would help to improve the summaries, since it's a way to combine the results of different systems and, in machine learning contexts, ensembles are generally outperforming single models.
Our test was not done on the entire collection due to the scarcity of computational power. Thus, we conduced a test on a subset of the dataset,that didn't really show any particular improvement on the scoring metrics. Due to this, we can say that RRF did not improve our system, but we cannot say that it is not an useful technique in general, since the size of our testing dataset was not significative and the BERT implementation is based on our suppositions. Moreover, a BERT-based solution is not suggested since the model usage times, combined with the additional overhead caused by calculating distances in the latent space, cause a significant slower response.

Method	Avg. Prec	Avg. Rec	F1 Score	Std Prec	Std Rec	MAP
BM25	0.7554	0.4250	0.5440	0.1962	0.1574	0.7869
RRF	0.5105	0.2832	0.3643	0.1223	0.1140	0.5014

Table 3: Performance metrics for BM25 and RRF.

3.5 Considering MMR, how λ impacts the accuracy (against ideal extracts) of summaries? Should λ be a fixed threshold or depend on the provided topic document (d-specific)?

In this case we have a strange outcome. In fact, the λ parameter seems to not affect at all the scores. Our idea is that this is caused by the fact that we could conduct the test only on a small set of documents (20), so the summarization on them might stay unchanged for different λs.

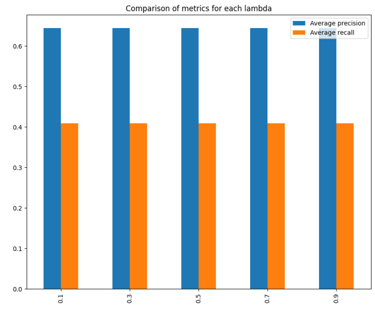


Figure 5: Performances at different λ

3.6 At the suggested p length threshold, is the system better at promoting recall or precision?

We noticed that precision is overall usually higher than recall, due to the fact the system tends to avoid to put non relevant sentences in the summary, leading to a lower recall.

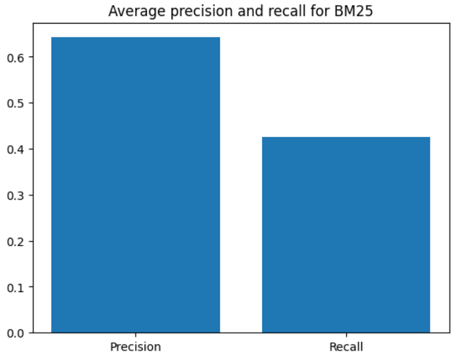


Figure 6: Precision and recall at given threshold