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.

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.
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.
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.
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 TF-IDF

2.4 Evaluation

3 Proposed questions

In this section, we answer point-to-point to the questions proposed during 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. However, it is important to note that

the y axis has a different scale, due to the smaller cardinality of the summary set.

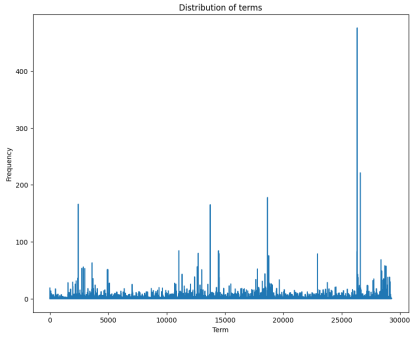

Figure 1: Corpus distribution

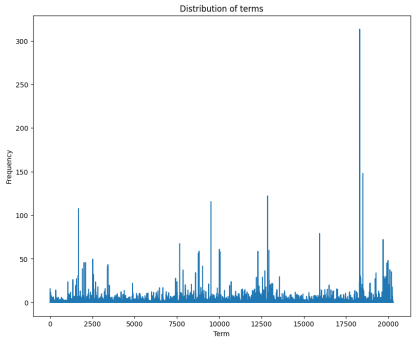

Figure 2: Summaries distribution

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

QUA SI POSSONO VEDERE DEI T TEST

Non mi trovo con il numero di sample. Da controllare

The following graph shows CONCLUSIONI

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

Qua va un attimo contestualizzata la figura

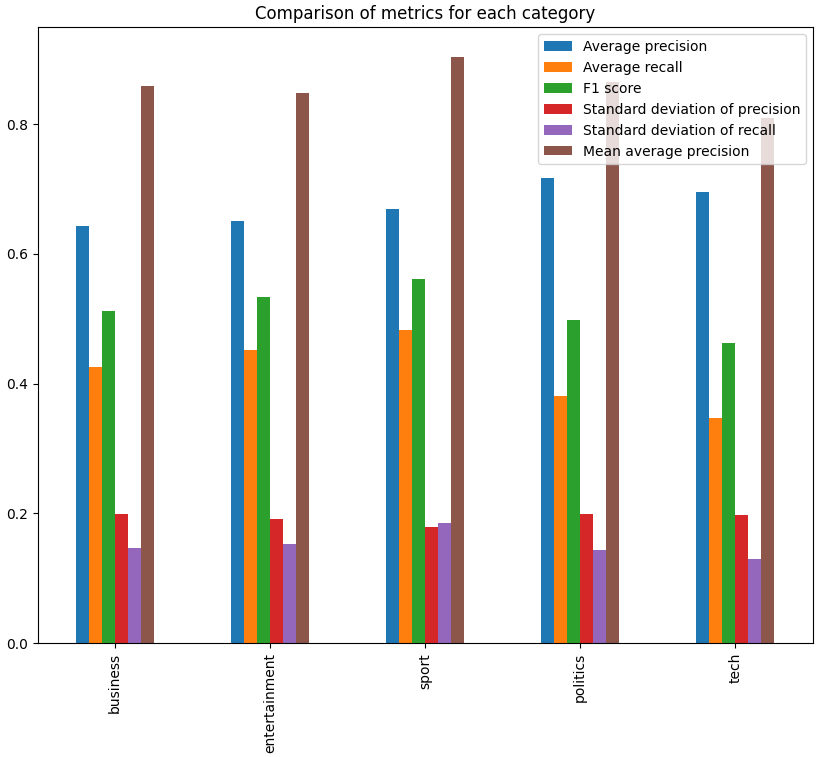

Figure 3: Performances

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 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.

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)?

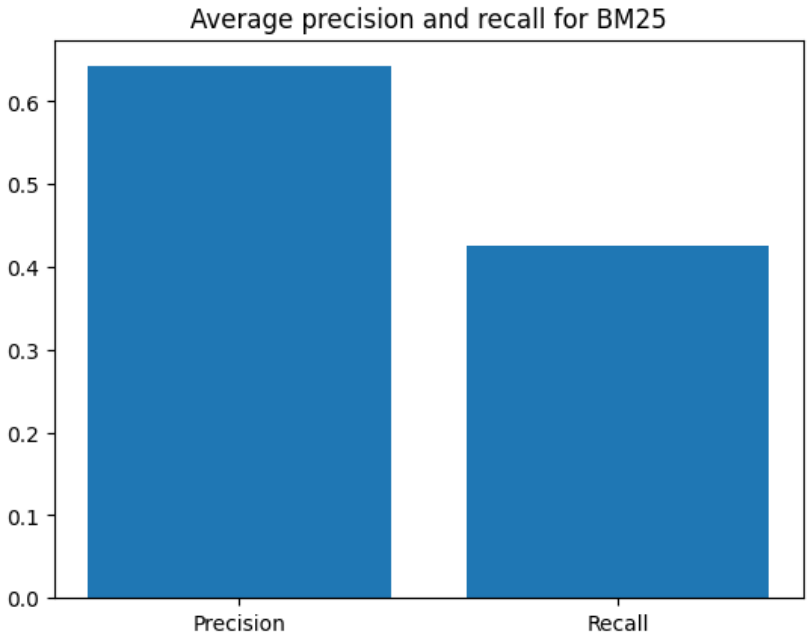

Figure 4: Precision and recall at given threshold

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)?

Qua la figura va rivista, le lambda nell'asse x non corrispondono ai valori proposti e poi l'asse sembra su un insieme continuo. Ho provato a leggere il codice, non capisco che sta facendo

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

? che stai provando a dire?
Every time we try to measure the performance of a system, we always see that the precision is the higest, but the recall is the lowest. This is because the system is trying to avoid to put in the summary a sentence that is not relevant, but this can lead to a lower recall.