Group 5

Ana Carolina Baptista, ist195529
ana.p.baptista@tecnico.ulisboa.pt

António Martinho Marçal, ist195735
antonio.marcal@tecnico.ulisboa.pt

Ådne Tøftum Svendsrud, ist1108703
adne.svendsrud@tecnico.ulisboa.pt

ABSTRACT

This report delves into the development and evaluation of an Information Retrieval (IR) system designed to tackle keyword extraction and text summarization challenges. Leveraging the BBC News Summary dataset, our study embarked on a two-phased approach: initially employing unsupervised techniques for feature extraction and summarization, followed by, in the forthcoming period, an exploration of supervised methods to refine the summarization process through relevance feedback. The project showcases innovative methodologies integrating traditional IR models to enhance the accuracy and efficiency of information retrieval. Our findings contribute to the understanding of effective strategies in IR systems, demonstrating the potential of combining unsupervised and, in the time ahead, supervised learning for improved content summarization and keyword extraction.

Keywords

Information Retrieval. Keyword Extraction. Text Summarization. Unsupervised Learning.

1. INTRODUCTION

The increasing necessity to process and understand vast volumes of textual data across various fields has stimulated the development of advanced Information Retrieval (IR) systems. These systems are crucial for efficiently extracting, summarizing, and presenting relevant information from a sea of data. This project explores IR techniques, focusing on keyword extraction and text summarization using the BBC News Summary dataset.

In the realm of keyword extraction, our system ranks normalized nouns and noun phrases within documents to highlight their relevance. For text summarization, we employ both abstractive and extractive approaches, wherein the extractive method selects key informative sentences from the original document to create a summary, ordered by either relevance or original location. This process aims to capture the main ideas of a document effectively while minimizing redundancy.

Our project has been structured into two main phases. The initial phase focuses on unsupervised keyword extraction and summarization over the document collection, utilizing reference extracts solely for evaluation purposes.

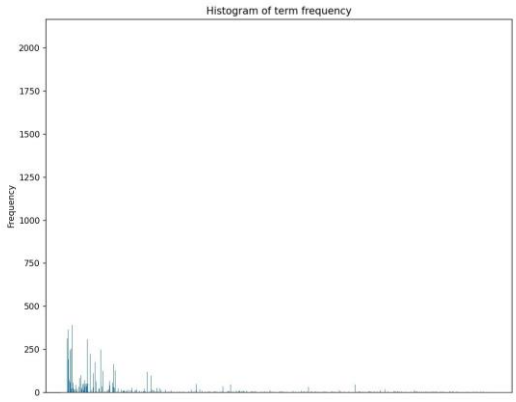
To ensure robustness and efficacy, our system leverages various Python libraries, including NLTK for natural language processing and Scikit-learn for machine learning tasks. The core functionalities encompass indexing for inverted index creation, summarization to generate concise document summaries, keyword extraction to identify top informative keywords, and evaluation to assess summary quality against reference extracts. Additionally, the solution incorporates advanced IR models like TF-IDF and BM25, alongside potential exploration of BERT embeddings for summarization enhancements.

Through innovative use of algorithms such as Maximal Marginal Relevance (MMR) and Reciprocal Rank Fusion (RRF), we aim to enhance the diversity and consensus of our IR system's outputs, thereby improving the overall performance and relevance of extracted keywords and generated summaries.

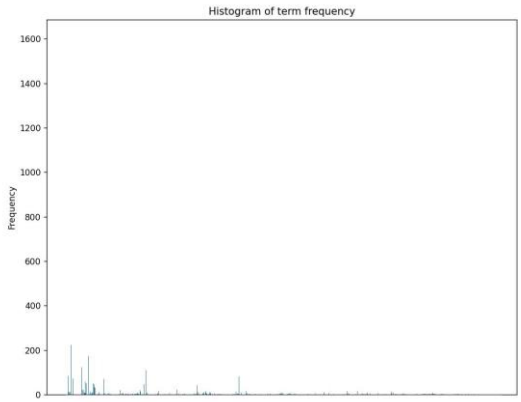
2. Corpus C and Summaries S

Considering the Dfull collection, our corpus is a static collection with 2225 text files representing journalistic articles divided in five categories: business, entertainment, politics, sport and tech. and a vocabulary of 41543 terms. Following the analysis of the Dfull collection, a significant aspect of our study involves understanding the distribution of terms across the documents. To this end, we have constructed a histogram that illustrates the frequency of terms within the corpus. This graphical representation plots a range of terms on the x-axis against the number of occurrences per document on the y-axis.

The histogram provides a clear visual insight into the term distribution, highlighting how certain terms are more prevalent across the corpus, while others are less frequent. This variability in term frequency underscores the diverse nature of the vocabulary used within the journalistic articles.



In reference to the summaries denoted as 'S,' it is noted that they are akin to the corpus labeled 'C,' constituting a fixed compilation of 2,225 text files. These files serve as summaries of news articles and are also categorized into five distinct sect sions: business, entertainment, politics, sport, and technology. Furthermore, the summaries encompass a vocabulary comprising 35,239 terms. To analyze S, we have constructed a histogram that illustrates the frequency of terms within the summaries.



Beyond the initial analysis of term frequency within our corpus, we delve into the significance of terms through of TF-IDF. TF-IDF, a statistical measure used to evaluate the importance of a word within a document in relation to a corpus, offers nuanced insights into term distribution that frequency counts cannot provide. Our investigation into the TF-IDF distribution reveals that terms are not uniformly distributed. In essence, more frequent terms within specific documents gain greater weight, underscoring their significance within the corpus. However, this distribution is not merely a reflection of term frequency; it also accounts for the uniqueness of terms across the document set.

3. SUMMARIZATION SYSTEM PERFORMANCE

After a detailed evaluation of the summarization system applied to our document collection, we observed a satisfactory overall performance, achieving a success rate of 40.2%. This result was obtained by directly comparing the summaries generated by the system with a set of reference summaries.

When analyzing the system performance between different categories of documents, we noticed significant variations that are worth highlighting. Particularly, the categories of politics (politics), entertainment (entertainment) and sport (sport) demonstrated better results (up to 57, 14%) compared to the overall system and to others. Further investigation revealed that these segments tend to present shorter texts and, consequently, a more direct structure, facilitating the task of summarization by allowing a more efficient capture of the main points.

On the other hand, the business and technology categories presented additional challenges for the summarization system. The specialized nature of these documents, often saturated with specific jargon and complex textual structures, contributed to a reduction in the accuracy of the summaries generated. Furthermore, the wide thematic variety found in these categories imposes extra difficulty in generating consistent summaries, as the heterogeneity of topics can make it difficult to identify common central elements that should be highlighted.

These observations highlight the importance of adapting summarization techniques to the specific characteristics of each category of documents. While categories with more concise and direct texts benefit from standard approaches, documents with high complexity and thematic diversity may require more sophisticated strategies, possibly including adjusting summarization parameters or implementing advanced techniques that better consider context and specificity of jargon.

In summary, our summarization system has demonstrated that it is capable of generating quality summaries for a wide range of documents, although performance varies significantly between different categories. These differences highlight the need to continue refining and adapting summarization techniques to meet the particular requirements of each type of document, with a view to improving the consistency and accuracy of the summaries produced.

4. IR Models Used

For the summarization task, we designed three different models based on distinct embedding methods and ranking criteria.

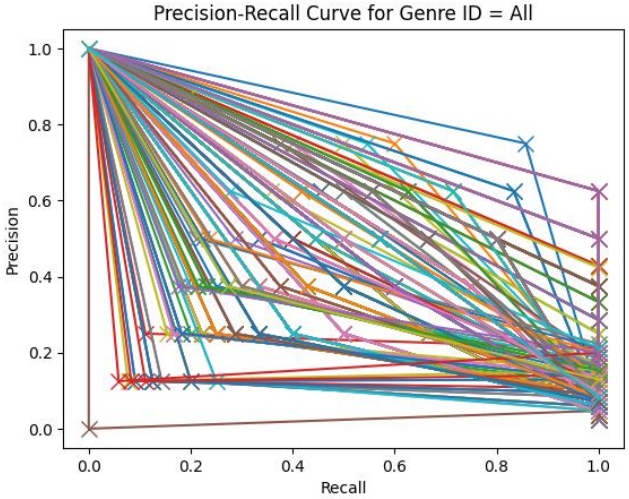
4.1 Vector Space with TF-IDF

Our first attempt was a simple model that vectorized each sentence based on the tf-idf scores of its terms. Using the inverted index created with the term frequencies of the entire corpus, we computed the scores and ranked each sentence accordingly. At the end, we made sure to fit the final summary in the provided bounds.

Due to the limitations this method provides, our model's results were not the best, with the Mean Average Precision (MAP) being around 39.21%. In addition, due the computational demands of calculating all these scores, the algorithm ran for an estimated 14.9 seconds / iteration, totaling just over 9 hours for the whole collection.

When the F-measure was computed for a Beta value of 1.75 (with summary limits of 8 sentences and 1000 characters), the average value obtained was 47,30%.

The chart below offers further information about our evaluation, demonstrating the obtained precision-recall curves for a select group of 500 documents.



Despite what the low values in precision might suggest, our model still had somewhat satisfactory results, with an Area Under the ROC curve of 68%.

4.2 Vector Space with BM25

Since we were not satisfied with our results, we decided to improve our scoring system while keeping our space represented as vectors. For this purpose, we chose the common improvement on TF-IDF: BM25.

Our MAP reached a total of 41,23%, while our F-Measure (computed as before) totaled 46,22%. This suggests that although we've seen a slight improvement in precision from the previous model, it is not significant.

Despite the extra steps, our model had a very similar efficiency to the one with simple TF-IDF, taking a comparable amount of average time per iteration.

4.3 BERT embeddings

In an attempt to drastically improve our results, we resorted to the large language model BERT to compute the embeddings for each sentence, instead of using simple vector spaces. For this process, we considered each sentence as a term and computed each embedding in the context of only the document where the sentence is inserted. This choice was due to the apparent low dependency between each document, specially across categories.

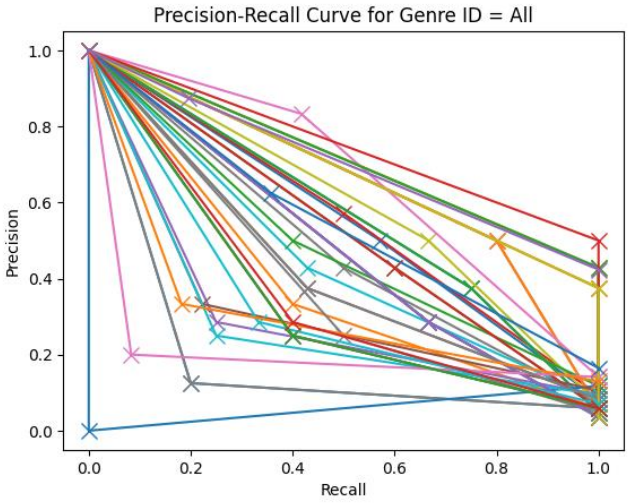
After we obtained the embeddings, we computed the average cosine similarity between each sentence and the other sentences in the document and used that as our score, giving us a notion of how relevant each sentence is in the document.

Our MAP reached a total of 40,11%, while our F-Measure (computed as before) totaled 46,51%.

Despite what our initial hypothesis might have led us to believe, the difference in performance is well within statistical error, suggesting this model is about equally as effective as the other ones.

This model was, however, considerably more efficient, boasting an average time per iteration of 1,1 seconds. This is likely due to the fact that BERT is already pre-trained, and the cosine similarity is not too computationally expensive.

Once again, we present the precision-recall curves in the chart below.



We can observe here that just like the other statistics, this suggests little to no improvement on the previous models.

5. Reciprocal Rank Fusion (RRF)

Since we had conceived several different models, we hypothesized that we could use different advantages from each model to compute a summary that was closer to the optimal one.

To achieve this, it was necessary to reach a consensus between each summary option available. Since the scoring systems were very different, we opted for comparing the ranks obtained for each sentence. For that purpose, the RRF formula seemed appropriate.

Fixating the value of μ as 5, we obtained a combined score for each sentence in our summaries, rearranging the ranks as appropriate.

The resulting MAP was 43,55% and the F-Measure was 49,26%. This means that, as expected, RRF signified an improvement in our previous models. However, the gains were still relatively small, which meant we wanted to try yet another strategy to improve our summarization model even more.

In terms of efficiency, though, this was clearly the lengthiest approach, at over 20 seconds per iteration, which is not surprising given that it works as a combination of several previous models.

6. Maximum Marginal Relevance (MMR)

In order to further close in on an optimal solution, we decided to implement a system to handle redundancy problems with our summaries. For this purpose, we chose the MMR algorithm, which will iteratively select the most relevant sentence not yet selected, that is the least similar to the ones already selected.

This will, in theory, give lower rankings to redundant sentences, though still prioritizing the most relevant ones. However, to accomplish a good balance between redundancy and relevance, we tried different values for the λ parameter.

For higher values of λ (closer to 1), MMR prioritizes relevance over diversity. This means that the summary will emphasize documents or sentences most directly related to the topic of interest, which can be useful when accuracy is more important than coverage of diverse information.

A lower value of λ (closer to 0), means the algorithm gives more weight to diversity, trying to include new information and reducing redundancy. This can be beneficial when the goal is a broad view of the topic, covering multiple aspects or points of view.

Despite these differences, for collections of documents or queries where topics are well-defined and focused, a fixed λ may be sufficient, as the balance between relevance and diversity tends to be consistent. However, for broader or heterogeneous topics, adjusting λ according to the specific nature of the document or query can optimize the quality of summaries.

In our case, we tested the results of applying MMR to the model which used BERT embeddings. This way, we not only ranked sentences based on the cosine similarity with the overall document, but also negatively penalized them for a higher similarity with more relevant sentences.

Fixing a λ value of 0.5, we obtained a MAP of 35,49%, an F-Measure of 40,94% and an AUC (ROC) of 64,29%. This suggests that although an interesting idea in theory, MMR might not be the best when applied to this particular problem.

The following chart illustrates the precision-recall curves for a total of 500 documents:

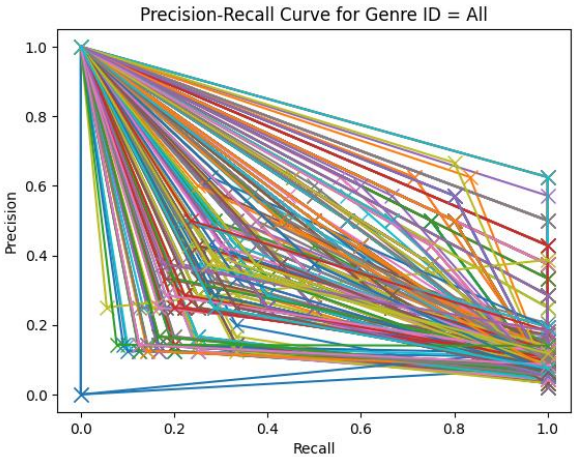


Figure 1. Precision-Recall Curves for 500 documents for a model using MMR

7. Recall Vs. Precision

Once all our approaches had been tested and evaluated, we turned our focus towards a comparison between the precision and the recall of our models.

As can be seen in the various precision-recall curves shown throughout this report, the values for recall are mostly higher than those for precision.

Given that this system in question performs an Information Retrieval task, this demonstrates that our models have not performed as they should, since it would be more important to have a higher percentage of relevant sentences in our summaries, and not so much missing fewer relevant sentences.

On the other hand, if this summarization model were to be used in a context where missing information can be a critical problem, then our emphasis on recall could be a benefit.

As for the practical reasons on why this might be happening, it is likely due to the parameters we have fixed for the maximum length of the summaries. In all our evaluations, we maintained the fixed value of p=8 maximum sentences per summary. If this value was lower, it is likely that the precision of our model would be higher, since it would be more likely that the top ranked sentences would be relevant for the document.