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.