Information Processing and Retrieval

Part 1 Project Report

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

1 Problem statement

In the same dataset context as Part I of this project, we are now addressing the tasks of clustering and classification with both an unsupervised approach and a supervised one, using summaries that are being provided for each document as reference.
Tasks conducted in this second part are:

- **Part A: Clustering;** for each document, the goal is grouping sentences based on their features and similarities. With this done, it is easy to select the most relevant sentences based on some criteria and algorithm that we defined.
- **Part B: Classification;** given a document, the goal is to split it into sentences and, using a binary classifier, define wether each sentence belongs to a summary or not.

This report can not contain all the data and graphs that we produced, so for more complete informations it is strongly suggested to check the comments on the provided notebook.
Some tasks were again very intensive in term of computation, so we decided to not use BERT embedding representations, since it would increase by a lot the time needed to run the code. Thus, our attention was on space representation using TF-IDF.

2 Adopted solutions

Part A: Clustering

This part is conducted in an unsupervised approach, with the idea of grouping sentences with clustering algorithms. In particular, we used the **sklearn** library, with the AgglomerativeClustering class as suggested.
The main paths to explore in this part are:

- **Number of clusters and used metrics:** the main challenge here was the correct choice of the **number of clusters**, because it's a parameter than could have a big impact on the results. Using **silhouette score** we have solved this problem in a iterative way.
- **Sentences selection:** the second challenge was to select the sentences that would be used to build the summary. Using **centroids** of each clusters, we were able to build summaries that had more topics and were more representative of the original text.

Number of clusters and used metrics

A problem related to part A is to define a good number of clusters to represent the feature space of the document.
In this part, we tried to construct a custom metric taking into account Silhouette score, Calinski-Harabasz score and a function of the number of clusters. The idea was to consider the number of clusters, that can vary in the range $[2, numberOfSentences]$, to avoid high sparse clusters representation with single-sentence clusters, but in the end we noticed that the silhouette score by itself was performing overall better.
We are leaving the code for the custom metric in the notebook, but it is commented since it was not used in the final version of the code.

Part B: Classification

3 Proposed questions