INTRODUCTION

The present day, when everyone is busy with their fast-paced life, reading articles, news or blogs is something people don't give much importance to anymore. But reading is an essential part in learning process/ improving knowledge. To address this problem, we see a lot of news websites giving summaries at the beginning of an article to attract people to read their article. To receive more users, they must present good summaries, writing them manually for every article is a lot of effort. So, many websites use automatic summarizers to write these summaries.

Automatic Summarization is the process of reducing a text document using a computer algorithm to generate a summary which retains the meaning and points conveyed by the original document. The best summarizers consider many variables like length, the style of writing and syntax for generating summaries. The most important task here is to find the subset of data in the text which contains information related to the entire set.

The main two approaches for automatic summarization are extraction and abstraction. The extractive methods summarize the text by selecting a set of existing words, phrases and/or sentences from the original document. The harder method, i.e. abstractive summarizers build semantic representations for the words/sentences and use natural language generation techniques to create summaries which are like summaries written by a human. These summaries usually contain vocabulary which isn't present in the original text. However, these methods are complex due to which more of the research is focused on improving extractive methods.

In this project, we've attempted to build extractive summarizers for a news articles in Portuguese using unsupervised methods and compare their efficiencies using ROUGE and BLEU evaluation metrics. We obtained the data from "PRIBERAM COMPRESSIVE SUMMARIZATION CORPUS"[1] which has 801 documents split into 80 topics each of which has 10 documents (one has 11). The documents are news stories from major Portuguese newspapers, radio and TV stations. The data set also provided us with two human generated summaries up to 100 words for each topic which were used for evaluation purpose.