Discussion

There is plenty of ongoing research in the field of Automatic Summarization. But, there is no clear approach that outperforms the others. In fact, even the evaluation metrics for automatic summarization are not perfect or necessarily conclusive. Two people summarizing an article may not use the same words but may yet be conveying the same meaning and two people using a lot of similar words may convey completely different meanings. There is no clear way to decide which of the summaries is more accurate. There may also be differences owing to perception and bias which are very difficult to be accounted for or evaluated.

Supervised methods can usually be solved with conventional machine learning techniques but supervised data with summaries are hard to find as it involves a lot of manual work in summarizing articles. Finding these summaries in a relatively less widely spoken language like Portuguese is harder yet. And hence we have tried to implement unsupervised methods for automatic summarization. These methods are usually extractive summarization techniques wherein the summaries are an aggregation of parts of the article. Hence the key idea is to pick the sentences that best convey the meaning of the article.

The naïve tf-idf based approach works surprising well. But with the introduction of a good heuristic like semantic similarity with the title and using cosine similarity as the similarity metric, we see that the results are considerably better. A graph based approach like TextRank also does considerably better than the naïve approach. However, we found that the sentence extraction through clustering performed the best. This is because using a clustering approach divides the article into clusters which have very little inter cluster similarity and high intra cluster similarity. This allows us to pick the best sentence from each cluster and avoid sentences which are similar to each other being included in the summary.

Directions for future.

While the data readily available to us in to be summarized Portuguese is limited, with the advent of newer and better automatic summarization may present better data for future development in the field. Some of the approaches like the graph based TextRank approach could perform significantly better with the availability of much larger data (a key indicator to this is seeing how good search engines, especially Google, are performing with huge amounts of data). An approach to use better similarity metrics in conjugation with graph based approaches could also yield better results.

We could also consider using hybrid approaches to get better summaries. An approach which can assign scores to sentences as a weighted average of a combination of approaches like sentence extraction through clustering and semantic similarity and additionally using position (sentence at the start and end of each article may in most cases have higher correlation to article summaries) and length of sentences could also be useful in assigning rank scores to sentences.