METHOD

Material

This project focuses on unsupervised, extractive text summarization. We have done a comparative study of different unsupervised methods of sentence extraction techniques on a dataset of Portuguese news articles.

The basic term weighing metric that we have used is Term Frequency – Inverse Document Frequency (tf-idf). Tf-idf acts as a good measure that positively weighs terms that are dominant locally and globally discriminatory. The term frequency promotes words that appear frequently in the local document. The inverse document frequency promotes words that appear in fewer of the documents in the corpus, thereby selecting words that are representative of the current document.

The first method we used calculated the tf-idf for every word in the document, and picks the sentences who have the highest values for the averaged tf-idf scores for the words in that sentence. This is a naïve approach but works surprisingly well. The second method that we used was TextRank[2][3], which is a graph-based sentence extraction method. The third method we developed was a title-based semantic sentence extraction that uses the title as a heuristic to find the most relevant sentences, and then chooses sentences with the best cosine similarity with respect to the sentence representation vectors.

The fourth and final method we used would using k-means clustering to cluster the sentences in the document based on Euclidean distances[4] of the sentences, represented as a vector of tf-idf values. K-means clustering requires the number of clusters to be pre-specified, which is advantageous for us as we can specify it to be the length of the target summary that we require. K-means clustering initializes random centroids in the vector space, and updates the positions of the centroids as it encounters new vectors. We chose to use clustering as it allows us to chose sentences that are different from each other. So, if we pick one sentence from each cluster, we are more likely to get better coverage in the summary.