Processamento e Recuperação de Informação - Part 2

Grupo 13 Sofia Aparício 81105 Rodrigo Lousada 81115
Rogério Cabaço 81470

December 7, 2017

Exercise 1- – An approach based on graph ranking

Starting the second part of the project, the group developed a structure based on classes for creating Graphs, Vertices and Edges as objects. Every Vertex has a sentence, a pageRank (score attributed to this specific sentence) and the list of Edges that links it with other vertexes. An Edge is a simple object with two Vertices, which is created if the pair of vertices has a cosine similarity above a specific threshold (constant).

The Graph class receives a list of sentences from a specific document and creates all vertices and edges from that, using a direct application of the cosine similarity formula given in the classes. A Graph also has the method getSummary which receives the number of sentences to return and returns a summary of the document. This procedure calls the method pageRank which ranks all vertices in the graph using the given formula applied up to a maximum of 50 iterations.

As it is possible to conclude from figure 1, were the threshold and the residual probability sufferd variations, the best value for the Threshold is 0,1. Regarding the Residual Probability the best value is between the values 0,2 to 0,3, but it is possible to observe a little drop down, so we chose the value 0,2. To obtain the optimal values for the thrashold we varied this value kepping the residual probability described in the assignment: 0,15. When analysing the residual probability variation we keeped the thrashold value as 0,1 (the optimal value).

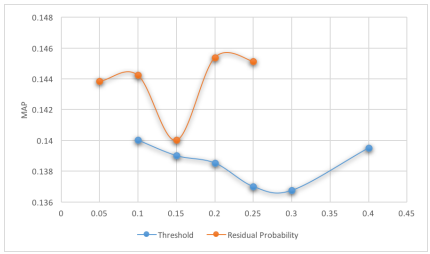

Figure 1: Variation of Threshold and Residual Probability on exercise 1

Exercise 2 - Improving the graph-ranking method

While improving the graph-ranking method of the previous exercise, the following formula was applied. (**inserir formula**)

Regarding the prior probability, the group chose to apply a non-uniform prior weights with basis on the position of the sentence in the document, but instead of having a simple model which assumes that sentences in the first positions are often more descriptive of the entire contents of the document, after some research it was chosen to apply the formula (**insert formula**) described in (**insert paper**) which also gives more important to the last sentences of the document, showing some improvements when applied to the TeMario dataset.

After this, the edge weight was based on the cosine similarity between pairs of sentences, leveraging TF-IDF (since it registered the best results comparing to BM25 in the first part of the project) weights when building the representations.

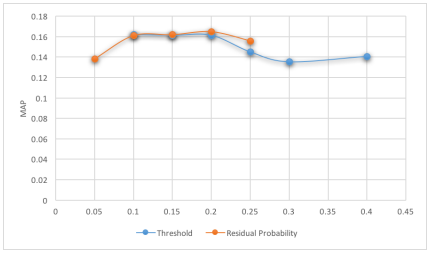

Figure 2: Variation of Threshold and Residual Probability on exercise 2

As it is possible to conclude from figure 2, were the threshold and the residual probability sufferd variations, the best value for the Threshold is 0,1, again. Regarding the Residual Probability the best value is between the values is 0,2. To obtain the optimal values for the thrashold we varied this value kepping the residual probability described in the assignment: 0,15. When analysing the residual probability variation we keeped the threshold value as 0,1 (the optimal value).

Exercise 3 - A supervised learning-to-rank approach

Implementing a supervised learning-to-rank approach, first the data was separated by 4 groups: train_data, train_target, test_data and test_target, using the TeMário_2006 dataset for training, and the TeMario dataset for testing. Leveraging a point-wise learning to rank strategy based on Perceptron algorithm. In the ranking model it was considered some features like the position of the sentence in the document, TF-IDF and BM25 scores for every sentence within the document, and the scores for each sentence when applied the first and second exercise. The target is a Boolean label which says if the document is or not present in the ideal extracted summaries. After this, the top-5 sentences for summary generation for each document is considered getting the confidence score for every sentence. The group varied the loss parameter in SGDC classifier, using different approaches from the perceptron in order to get the best results, however we only coud get a max MAP value of 0.07504221825287999 with the parameter set to log.

Exercicio 4 - Abordagem

In the last exercise, news from 4 different sources are extrated. First, to extract the data, we used a dictionary with the title of the articles as key and the value was a list with the following features: summary, tags, link, published and published_parsed. After that we orderd the articles by their published date (using published_parsed) to use the date as the position of the documents as we used on exercice 2 method of summarization, in order to show the most resent articles. The group used the second exercise ranking approach because it's where we had the best MAP. For sumarizing all these news articles, and in order to show a better summary for a client, since all the rules applied to news writing say all the important information should be presented in the title of the article, we summarize all new by title. After getting a summary of the 5 best articles to show in the webpage.