Project Report **PRI - PART 2**

Teammates: Patrcia Santos - 77091, Pedro Faleiro - 65007, Ricardo Sequeira - 79750
Group: 32
Date: 07/12/2017

Introduction

The following sections describe the work done on the second course project for Information Processing and Retrieval. While still on the subject of extractive summarization, the focus is now both on adapting and improving a proved algorithm, PageRank, and to explore a supervised learning procedure (Perceptron) to rank sentences in a text with the final intent of producing a summary. Finally, methods for information extraction from the web are also given some attention in the last exercise. Follows an explanation on the resolution of each of the proposed exercises.

Implementation

Most tools developed in the first part of the project were re-used (and improved) as some of the problems to solve required similar processing. To aid in Mean Average Precision calculations a new class was written and use throughout. To run each script, make sure your working directory is the project folder and use a python 3 interpreter.

Exercise 1

For the first exercise, the PageRank algorithm was adapted to perform extractive summarization. The undirected graph traversed in this procedure is built from nodes which represent sentences and edges that connect similar sentences (given a minimum threshold of similarity). It then iteratively computes a rank for all sentences, bounded by a maximum of 50 iterations or an absolute difference less then 0.001 between iterations, for every sentence. Existing sinks in the graph (nodes without outlinks) are handled the same way as the original algorithm, i.e., all sinks are considered to connect to all other nodes and an equal "travel" probability is considered. This guarantees that their initial rank is distributed equally among all other nodes. The summary consists, here and in all following exercises, of the 5 highest ranked sentences.

Exercise 2

The second exercise builds on the the previously adapted algorithm to allow a non-uniform prior probability distribution of the sentences and give additional emphasis on edge weights. From the suggested ways of computing this parameters, the following were used: priors based on the position of the sentence in the source text; priors based on sentence similarity towards the entire text; edge weights representing similarity between connected sentences. This new approach was then evaluated against the first method (using the TeMrio dataset) and the performance improved when using sentence similarity as priors.

Exercise 3

Here, a point-wise learning to rank procedure is explored, using a supervised learning method: a multi-class Perceptron. The functioning of the Perceptron relies on a set of sentence features

as input, which are weighted and combined to output a discriminated class. The input features considered were, given a sentence, its the position in the document, its similarity towards the whole text and the number of noun words it contains. The classes that make a valid output range from 1 to the maximum number of sentences in a summary, joined to a -1 representation for sentences that don't match the requisites to be part of the summary. The training is done using the TeMrio 2006 dataset and the evaluation on the original TeMrio dataset. The resulting poor performance highlights the need for additional (and probably of higher quality) features, such as acknowledging noun-phrases or the suggested graph centrality scores.

Exercise 4

This exercise concerns with the development of a program for illustrating extractive multi-document summarization, as a practical application. Information from the World news section of several different sources (New York Times, CNN, The Washington Post and Los Angels Times) is retrieved and, after parsed and combined, summarized using the procedure from the first exercise. The resulting summary is presented in a html file, together with additional info for each sentence that composes it.

Conclusion

We faced several challenges on implementing both new approaches and were, once again, reminded of how hard the subject of automatic summary extraction can be. However, the attempt to overcome them (even if not successful at all times) allowed a significant gain of knowledge on both the theory underlining this matter and the tools that can be used to implement a solution. The contrast between the adaptation of a largely used and well tested algorithm and the experience with a machine learning procedure to address the same problem, serves to show that a problem can be solved in several ways and the best solution is often only a compromise on the best tool available at the moment.