Team Carmen Mola

Challenge 1

Authors: Fruela Palacio Pérez, Ordoño Palacio Pérez y Pelayo Palacio Pérez

March 2022

INTRODUCTION

Introduction

Nowadays the number of text documents, proposals or reviews available on the Internet is huge and out of the scale of human supervising.
In this context we believe that automatic keyword extraction is a crucial tool for both public and private institutions.
Several approaches and models have been proposed in these last years:

- A first group of methods based on statistic methods.
- More advanced methods that incorporate machine learning techniques.
- Lately we can find methods which use the linguistic approach.
- The last methods to be incorporated are the graph-based ones.

Introduction

Our approach is to combine some of the existing methods instead of applying one of them based on the following:

- Creating a more robust corpus of relevant keywords.
- Using multiple methods to provide a better diversification for the search environment of the solutions, avoiding bias or falling into local maxima.

Our inspiration comes from the paper "*D. Vega-Oliveros, P. Spoljaric, E. Milios and L. Berton A multi-centrality index for graph-based keyword extraction*".
In order to process the large amount of data and because of the need of scalability we consider the best to develop a Python script to achieve the goals of the challenge.

METHOD

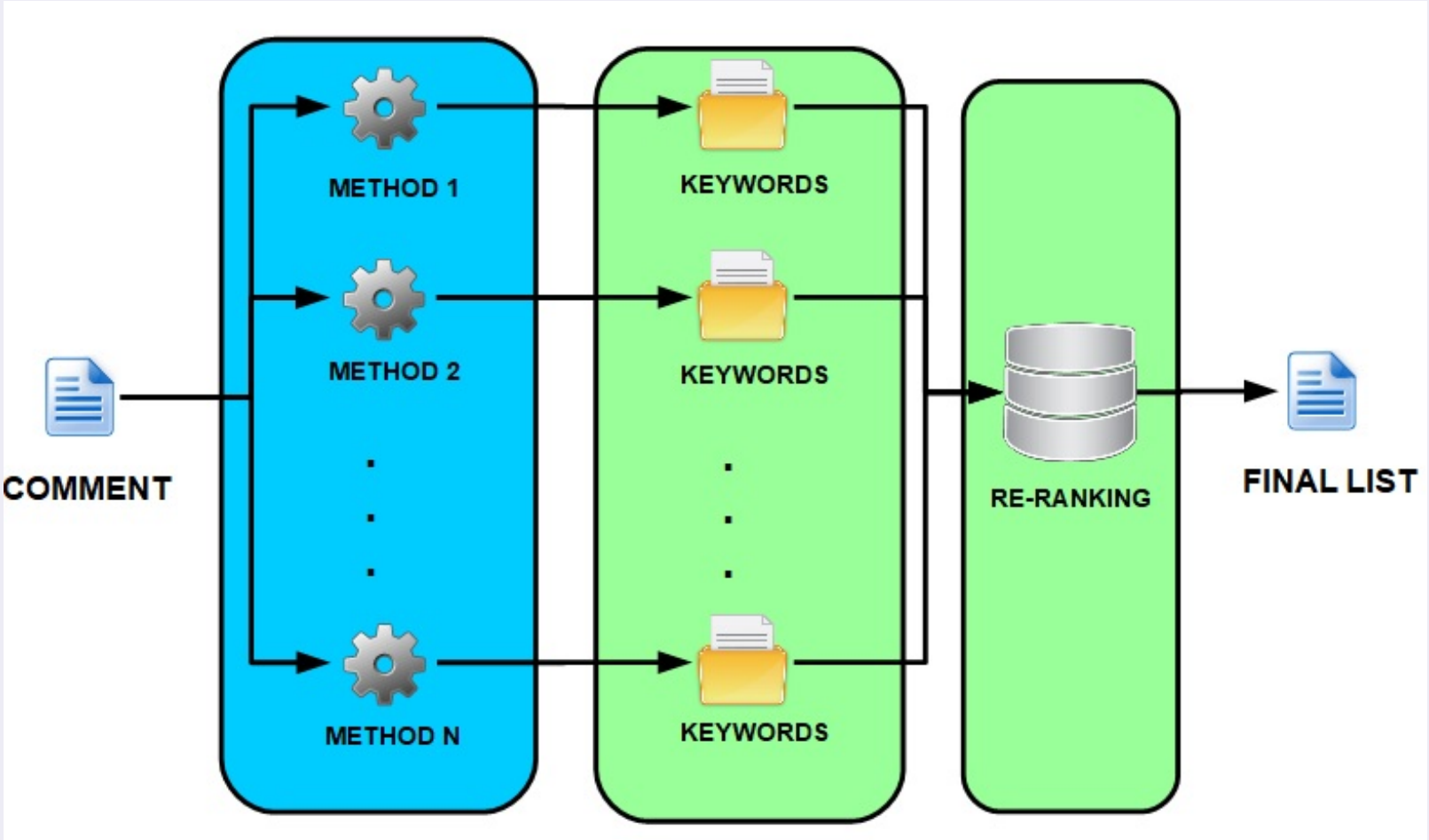
Method

Our combination is a mixture of both statistical and graph-based methods for keyword extraction. We reproduce the list of the methods below and the detailed bibliography can be found in the last section:

- First Phrases
- MultipartiesRank
- PositionRank
- SingleRank
- TextRank
- TF-IDF
- TopicRank
- YAKE!

Method

The algorithm is described as follows:



Method

We describe how our algorithm works: let be "n" the number of methods we use.

- For every comment we extract the 5 most relevant keywords using the methods described above. This step will return $5 \cdot n$ keywords.
- Since every algorithm uses its own ranking scale we need to create a re-ranking method in order to decide which 5 keywords are the most relevant overall.

Method

- We select the keywords that are more frequent, that is, if a keyword is chosen by 5 of the methods it should be more relevant than others that appear only once or twice.
- In case of a tie we order based on the times that those keywords appear as the first option for each of the methods, then as the second option and so on. For example, if we have a keyword that appears 4 times, 3 of them as the first option of three methods and we have another keyword, appearing also 4 times but only as the last option of four methods, the first one is going to be considered more relevant.
- In case of a tie in all the parameters between two or more keywords, a random choice among those keywords will be done.
- With the previous re-ranking process we get the final list of keywords per proposal as it was stated for this challenge.

BIBLIOGRAPHY

Bibliography

- A. Bougouin, F. Boudin and B. Daille, *TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction*
- F. Boudin, *Unsupervised keyphrase extraction with multipartite graphs*
- R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes and A. Jatowt, *YAKE! Keyword extraction from single documents using multiple local features*
- C. Florescu and C. Caragea, *PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents*
- R. Mihalcea and P. Tarau, *TextRank: Bringing Order into Text*

Bibliography

- J. Piskorski, N. Stefanovitch, G. Jacquet and A. Podavini, *Exploring Linguistically-Lightweight Keyword Extraction Techniques for Indexing News Articles in a Multilingual Set-up*
- R. Stephen, *Understanding inverse document frequency: on theoretical arguments for IDF*
- D. Vega-Oliveros, P. Spoljaric, E. Milios and L. Berton, *A multi-centrality index for graph-based keyword extraction*
- W. Xiaojun, X. Jianguo, *Single document keyphrase extraction using neighborhood knowledge*

****

THANK YOU FOR YOUR ATTENTION