Extract of Keywords in text using graph theory and formal concept analysis

Currently, textual data is expanding exponentially. To do this fact, it is necessary to create tools that enable to extract the most relevant information from the text to facilitate the classification and comprehension. Keyword extraction, which is a process that extracts the most relevant words from a text, may be the most essential of the many available tools for extracting information from text. This demonstrates the significance of keyword extraction due to its applications in various disciplines, particularly classification and text comprehension. In recent years, numerous research groups have worked to enhance the efficacy of keyword extraction. See Bharti and Babu (2017) and Shin, Lee, and Cho (2023) for relevant approaches based on statistical methods, linguistic features-based machine learning techniques, graph-based approaches, and hybrid models.

This proposal for this project concentrates on statistical methods combined with graph theory and formal concept analysis to keyword extraction. The graph theory is a mathematical theory that permits the text to be represented as a graph, where the nodes are the words, and the edges are the relationships between the words. This relationship is based on the co-occurrence of words in a text, and it has been used in Schenker et al. (2005), Litvak and Last (2008), Chen et al. (2019) among others, utilizing this type of strategy, we can extract the keywords. The formal concept analysis is a mathematical theory that defines the intuitive notion of "concept" as a relationship between objects (entities that possess particular properties) and attributes (properties that characterize a certain set of objects). This conveniently constructed relationship enables the establishment of an orderly structure between concepts that is easily visualized in a diagram known as a Hasse diagram, which contains all the information about the relationship between concepts constructed by the interaction of their objects and attributes. In Maio et al. (2016), the authors combine formal concept and fuzzy theory to create text summarization, demonstrating some advantages over keyword extraction using formal concept analysis.

Thus, the proposal for this project consists of offering a keyword extraction technique through the combination of graph theory and formal concept analysis tools. Accordingly, the proposed strategy employs a two-stage approach. The first entails obtaining a word graph, which, the nodes are words and the edges between nodes are the result of words satisfying a "closeness" relationship. This methodology has been utilized by Litvak and Last (2008), Ohsawa, Benson, and Yachida (1998), and Abilhoa and De Castro (2014), among others. In this proposal, we intend to combine the generality of n-grams by designating the correspondence between vertices based on the frequency of neighboring words in the context, i.e. their co-occurrence in a window of size k, provided that they satisfy a bound inspired by the sigmoid function (without appealing directly to it). Once this graph, which is generated with weights to indicate the strength of the linkage between the words viewed as nodes, has been constructed, it is desired to use centrality metrics for graphs already described in the literature M. E.J (2020) (beetweness, clossenss, degree, eigenvalue). Given that a node's eigenvalue quantifies the level of importance of this node relative to another important node, we wish to use this criterion to identify the most important nodes (words) in the graph. In the literature already cited, various

Abel Alvarez Bustos (abel.alvarez@javerianacali.edu.co)
Professors of Engineering and Sciences Faculty, Pontificia Universidad Javeriana-Cali

Carlos Ernesto Ramirez (carlosovalle@javerianacali.edu.co)
Professors of Engineering and Sciences Faculty, Pontificia Universidad Javeriana-Cali

strategies for network importance selection are pursued. The criterion for this endeavor will be determined by the node's eigenvalue.

The second stage of the proposal entails the construction of a concept lattice based on the metrics of the nodes with the highest eigenvalue. The following step is to organize the information acquired from the metrics of the nodes so that it can be used as input for the construction of a lattice of concepts. To this end, the metrics of the nodes will be categorized as high, medium, and low so that the nodes with the highest eigenvalue can be used as objects and the metrics obtained for each of these nodes in the aforementioned scale can be used as attributes. This allows us to construct the concept lattice, where a concept will be a collection of words that share a level of node metrics. In this regard, we believe that this proposal, despite adopting a keyword extractive approach, maintains the semantic coherence of the extracted words by differentiating word groups (within a concept) from other word groups (in another concept).

Having described the above methodology, the objective of this project is to build a library or computational package easily implementable in a programming language, which allows the extraction of keywords from a library of scientific research articles on any topic in social sciences and humanities. The proposed methodology will provide the researcher not only with the most relevant words from the collection of documents provided, but also with the semantic connection between these words in a natural and structured way.

References

Abilhoa, Willyan D., and Leandro N. De Castro. 2014. "A keyword extraction method from twitter messages represented as graphs." *Applied Mathematics and Computation* 240: 308–25. https://doi.org/10.1016/j.amc.2014.04.090.

Bharti, Santosh Kumar, and Korra Sathya Babu. 2017. "Automatic Keyword Extraction for Text Summarization: A Survey." *arXiv Preprint arXiv:1704.03242*.

Chen, Yan, Jie Wang, Ping Li, and Peilun Guo. 2019. "Single Document Keyword Extraction via Quantifying Higher-Order Structural Features of Word Co-Occurrence Graph." *Computer Speech & Language* 57: 98–107.

Litvak, Marina, and Mark Last. 2008. "Graph-Based Keyword Extraction for Single-Document Summarization." In *Coling 2008: Proceedings of the Workshop Multi-Source Multilingual Information Extraction and Summarization*, 17–24.

M. E.J, Newman. 2020. *Networks, An Introduction*. Oxford University Press.

Maio, Carmen De, Giuseppe Fenza, Vincenzo Loia, and Mimmo Parente. 2016. "Time Aware Knowledge Extraction for Microblog Summarization on Twitter." *Information Fusion* 28 (March): 60–74. https://doi.org/10.1016/j.inffus.2015.06.004.

Ohsawa, Y., N. E. Benson, and M. Yachida. 1998. "KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor." In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98-*, 12–18. https://doi.org/10.1109/ADL.1998.670375.

Schenker, Adam, Horst Bunke, Mark Last, and Abraham Kandel. 2005. *Graph-Theoretic Techniques for Web Content Mining*. Vol. 62. World Scientific.

Shin, Hunsik, Hye Jin Lee, and Sungzoon Cho. 2023. "General-Use Unsupervised Keyword Extraction Model for Keyword Analysis." *Expert Systems with Applications* 233: 120889.

Abel Alvarez Bustos (abel.alvarez@javerianacali.edu.co)
Professors of Engineering and Sciences Faculty, Pontificia Universidad Javeriana-Cali

Carlos Ernesto Ramirez (carlosovalle@javerianacali.edu.co)
Professors of Engineering and Sciences Faculty, Pontificia Universidad Javeriana-Cali