

Prospecção e Análise de Dados

2º Sem 2022/2023

Practical Work I

Implement the multi-word Relevant Expressions LocalMaxs extractor, taking into account the following requirements:

- a) To obtain tokens, you may add a space character before and/or after characters such as “;”, “:”, “!”, “?”, “<”, “>”, “&”, “)”, “(”, “]”, “[”, among others that do not change the semantics of the text, in order to improve the reliability of token frequencies.
- b) Choose a sufficiently efficient programming language so you can use the extractor in corpus of at least 1.5 million words.
- c) Let it be possible to use more than one cohesion metric, such as SCP, Dice, ϕ^2 , among others.
- d) Consider n -grams of length up to 7.
- e) Consider a minimum frequency filter as necessary requirement for an n -gram to be considered as Relevant Expression (RE); for example, the frequency of a RE must be at least 2.
- f) Evaluate the results of the extractor through the Precision, Recall and F metric, for at least two corpora. Consider one or more languages.