Prospeção e Análise de Dados

2º Sem 2021/2022

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, ϕ2, amomg 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.

1) Evaluate the results of the extractor through the Precision, Recall and F metric, for at least two corpora. Consider one or more languages.

2) Eliminate Relevant Expressions produced by LocalMaxs, which contain stop-words such as extreme unigrams (w_1 and w_n). To do this, use the non-thersholds approach you have learnt to detect stop-words. Compare results.