First Information Retrieval Model

The goal of this session is to run, understand and improve a simple information retrieval system.

In case of text, from a given query, the system has to find the best documents related to the query.

We start by the simplest word representation known as TF-IDF for Term Frequency and Inverse Document Frequency. It puts a score on every word of each document. A term will have a bigger score if is frequent in the local document and also rare on the entire corpus.

Exercice 1: The code below implements a very basic information retrieval system:

https://colab.research.google.com/drive/1MrDvnQqSEMvnUlN0RkH1faxaX1GGJ8Hn?usp=sharing

Try to understand by ourselves what it does.

Can we use this model to predict the score of a review? How?

Exercice 2 : This code is very inefficient. Why? How to improve it?

The vocabulary is quite raw. You can use pretreatments seen last session is order to improve the precision of the system.

Exercice 3 : Design a completely different system using learned embeddings like (w2v, glove or fastext) and the ideas that you had developed on last session.