Departamento de Informática

Masters Dissertation Proposal

Topic Modelling in Medical Data

Supervision: Hugo Pedro Proença (UBI-DI)

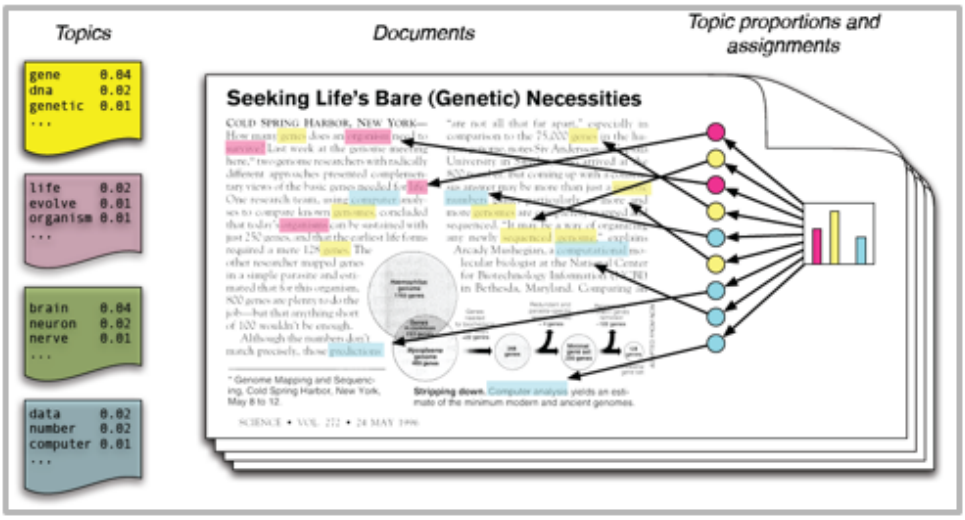

Fig. 1: Illustration of topic modelling main task: given a text source, determine the most important semantic concepts in the document. (image from https://www.analyticsvidhya.com/)

In the natural language processing context, topic modelling refers to a statistical analysis that is carried out for estimating the most important concepts (i.e., "topics") in a document or collection of documents. This type of techniques is frequently used as a text-mining solution for discovering the semantics associated to a particular document.

The main intuition in this type of statistical analysis is that a set of particular words should appear in a document more or less frequently: i.e., "dog" and "bone" will appear more often in documents about dogs, while "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear approximately equally in both. A document typically concerns about multiple topics in different proportions.

Topic models are mainly used to understand large collections of unstructured texts. Originally developed as text-mining solutions, state-of-the-art models have been used to detect instructive structures in data such as genetic information, images, and networks.

This proposal lays at the intersection of important scientific research topics, namely natural language processing and medical diagnosis. The goal is to evaluate the state-of-the-art in topic modelling solutions [1] for clinical pathways and clinical guidelines data.

Prerequisites

- Basic skills in English writing/reading;
- Good programming skills, preferably in *Python*, *Keras* and *TensorFlow* languages.
- Previous knowledge about Machine Learning and Artificial Intelligence.

References:

[1] https://paperswithcode.com/task/topic-models/latest