Learning text representation

for Social Sciences

Salomé Do

February 2020

Introduction

This short course aims at introducing basic tools used to represent textual data in a linguistically powerful way. Representation is at the very heart of most current computational approaches to linguistics. Finding dense, numeric representations enables to use the transformed textual data in many other learning algorithms, and thus help solving many downstream tasks. For instance, finding numeric, robust representations for two texts enables to compute a distance, a similarity between them. Such representations are known to embed a lot of semantic information and thus are particularly relevant when trying to match semantically similar documents, rather than formally similar documents.

Yet, these representations have historically been designed and used to enhance natural language processing models, generally aimed at automatically answering questions, translating texts, extracting informations such as person's, places, organisations mentions, etc. These tasks are predictive rather than descriptive tasks, and might then be of limited interest for a social scientist. Finding ways to use the power of textual representations in a descriptive and exploratory setting is still a scientific challenge, which we will try to account for.

In the course, we chose to focus on the tools that were developped mainly between 2013 and 2019. These tools generally use deep learning, but the mathematical details of deep learning are not on the scope of this course. We will instead describe the core ideas and differences between the main deep learning architectures, and how they can achieve extracting high-level, and sometimes semantically abstract informations. A large part of the course is also dedicated to the recent exploratory and descriptive methods that can reveal to be meaningful for social scientists. Throughout the course, the most important concepts and algorithms will be put into practice, with a real-life corpus.

Contents