Vector space exploration of literary language

Verny, Romain · Kruusmaa, Krister · Barré, Jean

11 mars 2022

Abstract

In this project, we will first intend to reproduce the results of the paper *Vector space exploration of literary language* (Cranenburgh et al. 2019). Then, we will try to implement the methods used with a corpus of french novels to verify the robustness of the results.

We will divide the experiments into three parts, and each will take charge of one of the parts.

1. Introduction

The aim of the experiments is to understand with natural language processing methods how literary novels distinguish themselves from other novels throught textual conventions. Using machine learning techniques, based on vector space representations using topic models (LDA) and word embeddings (DBoW paragraph vectors), we will predict the literariness of novels as perceived by readers.

2. Unsupervised document representations

2.1. Topic modeling with BoW baseline

Krister Kruusmaa will deal with the Vector Space Model of language, assigning coordinates to novels in a high-dimensional space in which semantic similarity of documents and words is realized as spatial distance. From a bag of words baseline, he will implement the Latent Dirichlet Allocation approach of topic modeling, which learns the distributions of topics across words and documents.

2.2. Neural document embeddings with DBoW baseline

In this section Romain Verny will implement paragraph vectors (doc2vec) wich are neural document embeddings based on an extension of word2vec. We will use the Distributed Bag-of-Words (DBoW) paragraph vectors model with negative sampling as implemented in gensim.

3. Supervised predictive models

Jean Barré will apply Support Vector Machines (SVM) models to the task of predicting the degree of literariness for each document, based on the features collected. The accuracy score of the classifier will be report with two evaluation metrics. R2 which expresses the amount of variation in the original ratings that is explained by the model and the Root Mean Square Error (RMSE) which gives the expected error for a prediction in the original scale of the ratings.