Text Mining and Natural Language Processing

Dr. Matthieu Cisel

September 2021

1 Goals

In this class based primarily on Python, students are taught the basics of Natural Language Processing (NLP) with spaCy as well as some Text Mining techniques. Regarding spaCy, we lay emphasis on Part-of-Speech Tagging and the most common concepts of the field (lemmatization, etc.). Regarding text mining, the focus lays on Bag of Words approaches and associated metrics (TF-IDF). The hands-on project is divided into NLP and Text Mining. NLP is focused around the use of POS-taggers to create a large scale exercise database.

2 Instructions

2.1 Datacamp class

Follow the ”Feature Engineering for NLP in Python” and ”Advance NLP with spaCy” classes. Submit the certificates on Teams.

2.2 Expectations

At the end of the version, submit :

1. a PDF version Powerpoint presentation with figures and tables to report your analyses. The presentation must be aesthetically pleasing, complete, and well structured.

2. A well commented and structured PDF version of a Jupyter notebook. Code annotation is mandatory.

2.3 Text Mining and doctoral dissertations

You are provided with a set of 50 URLs of doctoral dissertations. Design a ”for loop” in order to download all 50 documents. Find a Python library online to automatically turn them into a set of .txt files (1 file per page).

Find a solution online to automatically detect the language of the document using stop words. Propose a methodology.

Process the data and use TF-IDF and cosine to assess similarity between the documents. Try to train and use a bigram and a 3-gram model on a relevant part of the dataset. What do you observe in terms of computing time ?

2.4 POS-tagging and exercise generation

You are provided with a corpus of sentences in different languages. Two Python libraries can be used for the following tasks : spaCy and stanza.

Design a strategy to automatically create the necessary data for MCQ exercises (target language is English) with the following goals :

1. identify 1 to 3 nouns in the sentences (among 4 words of the sentence)
2. identify 1 to 3 pronouns in the sentences (among 4 words of the sentence)

Design a strategy to automatically create the necessary data for MCQ exercises (target language is German now) with the following goals:

1. identify the gender of a noun
2. identify the case of a noun

You are provided with a database format meant for exercise integration in Airtable. Give the instructor a set of 20 automatically generated exercises. We will integrate it into a learning platform to assess the potentialities of this approach. To delve into the topic, search for the NLP for BEA (Building Educational Applications) community.

Apply a conjugator-based strategy to the sentence corpus and propose a relevant automatically generated exercise based on this tool.

Apply a Named Entity Recognition strategy to the sentences corpus and propose a relevant automatically generated exercise based on this tool.