Preliminary syllabus for a two-day workshop
Introduction to Automated Text Analysis with Python

Abstract

During the two days of the workshop, we will cover some of the most popular automated methods to systematically explore and analyze (large) corpora of texts. We start at the very beginning of the NLP pipeline, learning about the pre-processing steps necessary to transfer unstructured, raw text documents into standardized, more useful representations, specifically prepared to retain as much information as possible while still being in a suitable format for different analysis techniques. Diving deeper into one popular representation that made possible many of the latest advances in Natural Language Processing, we will get to know Word Embeddings and how they are useful to surface some of the underlying meanings and relations in a corpus of text documents. Systematizing these aspects, we will leverage the newly opened embedding space to solve one of the most common problems confronted when dealing with a large collection of texts: Finding common themes and topics. We will see how this can be done by clustering vectors in the embedding space.

On the second day, we will turn towards what is probably the most popular NLP task and most general approach for making sense of texts: Text Classification. We will first explore different lightweight methods for assigning texts into different pre-defined categories, before talking about the strengths (and weaknesses) of pre-trained Language Models like BERT and its variants for text classification. We will look into fine-tuning as one strategy to harness the potential of these powerful models, before ending with a practical discussion on why we don't just ask ChatGPT to please do all what we learned over these two days for us.

The workshop will offer a mixture of theory and practice, with a focus on applying the different methods presented to actual text data. Therefore, basic familiarity with the programming language Python and its core concepts (defining functions, writing for loops, basic pandas dataframe operations) is important. Most of the lectures and applications will be based on prepared Jupyter Notebooks, allowing to interactively explore, execute, and discuss single code snippets and whole analysis pipelines. We will use Google Colab in combination with Google Drive to access and execute the materials, which is why there is no need to set up a local programming environment before the workshop.

Program Outline

Day 1

- Types, Formats and Properties of Text Data from the Web
- **Pre-processing** for Text Data from the Web – Dealing with URLs, Emojis and Hashtags
- Exploring Text Corpora through **Word Embeddings** (Word2Vec)
- Finding Similar Documents – **Topic Modeling** through Clustering of Embeddings

Day 2

- Basics of **supervised** NLP methods
- Different Approaches for **Text Classification**
- (Re-)Using and **fine-tuning** pre-trained Language Models (BERT & Co.)
- Outlook: Why don't we just ask **ChatGPT** to do all of this?