Business Analytics

Lecture 2: NLP Pipelines

Ulrich Wohak[1]

[1]Department of Economics
Vienna University of Economics and Business

Introduction

- Before we start modelling, we need to make our data understandable for computers
- We will learn a few 'technical' terms from the NLP literature
- Typically, we will think of the whole development process as a 'pipeline'
- The first step of any such pipeline is our topic for today: **preprocessing**

Preprocessing (1)

- Preprocessing is a loose term that incorporates many different methods
- Generally speaking, preprocessing converts raw (text) data into suitable inputs for NLP methods
- Typically, the main purpose is *dimensionality reduction*
- Note that we are deleting data on purpose!
- Performance of your models will **heavily** depend on the steps you include in your preprocessing pipeline

Preprocessing (2)

- Appropriate preprocessing pipeline typically depends on the method (e.g. word co-occurrence vs Transformers)
- For some methods, text *should not be processed*. Why?
- Let's look at what a typical pipeline might look like ...

Preprocessing (3)

1. Remove punctuation, numbers and other non-letter characters
2. Tokenization
3. Remove unwanted parts of speech (more on this later)
4. Remove non-informative text (based on research question and domain-knowledge) such as geographic locations
5. Stemming/Lemmatizing
6. Numeric representation

Code

Let's implement them in code!