Quantitative Analysis of Text: A Friendly Introduction

The amount of data being generated is exploding and much of this data comes in the form of text. As a kind of communicative speech, text registers, expresses, and creates social phenomenon and meanings. By summarizing and extracting information from large amounts of text, we can better understand human behaviour and institutions. This short course has three objectives. First, to introduce some of the building blocks for the quantitative analysis of text. Second, to gain some hands-on experience in analysing text data using the Python programming language. Third, to explore how quantitative text analysis can yield social scientific insights. Motivated examples are provided throughout.

No knowledge of Python is necessary although some exposure to programming is very helpful. Basic mathematical maturity is required.

1. **Introduction and Basic Python Syntax. [3 hours]**

   What can text tell us? This module looks at how text data can illuminate questions in social science. Basic concepts of Python coding are reviewed, alongside regular expressions.

2. **Machine Learning. [6 hours]**

   How do machines learn patterns from data? By optimizing on an objective function of course! We will undertake a high-level survey of machine learning techniques, beginning from linear models like regressions and progressing to non-linear models like random forests and neural networks.

3. **Pre-Processing Text. [2 hours]**

   Text comprises many characters and symbols and comes in many shapes and sizes. How can we standardize and clean text documents to make them fit for purpose? In this module, we will discuss why and how to pre-process text documents and the consequences of pre-processing choices.

4. **Bag-of-Words Representations. [3 hours]**

   We have to represent text numerically to perform computational operations on them. How can we turn documents into vectors? In this module, we study one of the most basic numerical representations of text: the bag-of words (BOW) model. The BOW model is, in essence, a word frequency count that disregards order. Despite its simplicity, the BOW model—including weighted variants like TF-IDF—performs well in many applications.

5. **Topic Modelling. [3 hours]**

   Can a machine identify topics in a text corpus without any human supervision? In this module, we will examine how empirical relationships between words, topics, and documents can be exploited to classify and describe the content of large corpora. Models to be considered include Latent Dirichlet Allocation and Non-negative Matrix Factorization. The topics estimated by these models are probability distributions over words and must be interpreted by the researcher.

6. **Word Embeddings. [3 hours]** Apples and oranges are both fruits. But apples are red and oranges are, well, orange. Can we represent words as vectors in a way that captures such similarities and differences? We will study how algorithms such as word2vec generate word embeddings by training neural networks to predict masked words. Word embeddings have improved performance on many natural language processing tasks.

General Readings

- Benoit, Ken. "Text as data: An overview." *The SAGE handbook of research methods in political science and international relations* (2020): 461-497.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. "Text as data." *Journal of Economic Literature* 57, no. 3 (2019): 535-74.