Text as Data: Computational Text Analysis

Instructor: Ashrakat Elshehawy
Department of Politics and International Relations
University of Oxford

Trinity Term, 2021

Contact: ashrakat.elshehawy@politics.ox.ac.uk
Class Place: Virtual via Zoom (Meeting times below)
Office Hours: Fridays 16.00-18.00

GitHub Link for Class
Class Meeting Link
Book office hour slots & attend here

Course Description

In recent decades governments have started to maintain an online presence of their archives and documentation of their proceedings and decisions. Newspapers around the world continue to produce daily textual data. Different groups and individuals are also employing online platforms at a rapid rate, like Twitter, Facebook, and Reddit that constantly store data about users' activities. All of this has led to an availability of extensive text data online that social scientists can make use of to answer pressing research questions that were previously difficult to approach. The aim of this course is to offer students an introduction into the methods of Natural Language Processing (NLP) and Computational Text Analysis to be able to process and analyse large collections of textual data.

We will be using Python as a programming language during this course. The course will start by introducing text processing mechanisms. We are going to learn about mechanisms like tokenization, lemmatization, stemming, part-of-speech tagging, and named-entity recognition. The course will also provide insight in methods of managing and manipulating text data in Python. We will then cover aspects of numerical representation of text, for example like word-embedding, and also discuss metrics of text similarity. After that, we will focus on methods of unsupervised machine learning like clustering and topic modelling, as well as, supervised machine learning methods, with a focus on classification techniques and sentiment analysis.

Prerequisites

Basic knowledge of Python as a programming language; either through taking in Hillary Term the optional course *Introduction to Data Science and Measurement with Python* or students can also

show that they have undergone prior training in Python.

You should also have knowledge of basic statistical concepts, through a course like *Intro to Stats* offered by the Department.

Course Materials

- All course material will be posted on GitHub. This is the link: https://github.com/aelshehawy/text-as-data-computational-text-analysis-oxford
- For our coding sessions we will use Google Colab Notebooks, coding sheets and their solutions are hosted on the GitHub Page of the class.
- You should have a Google Drive account and familiarize yourself with the Google Colab Work environment before coming to class. You might find this link to be useful.
- You can easily open any sheet from within GitHub by clicking on *"Open in Colab"* button, remember to always save a copy of your edits in your Google Drive Account.
- The Teams Website will serve our needs for communication. It has different channels, each serves discussions and questions for different purposes (General, Code Debugging, Office Hours, Windows Support, Relevant Literature). You should have received an invitation for the Teams Website set up for this class. Please get in touch if you still need to be added.
- The Syllabus will be hosted on Canvas.

Class Times and Office Hours

The class takes place online via Zoom. You are encouraged to attend the Python Refresher Sessions in Week 0. We will be refreshing Python basics needed for this course.

- Python Refresher Session 1: Thursday the 22nd of April 15.00-17.00 Link
- Python Refresher Session 2: Friday the 23rd of April 15.00-17.00 Link

We meet for 4 weeks during Trinity term. We will use this Zoom Link for all of our 4 sessions. We will meet on the following days:

- Week 1: Monday the 26th of April 14.00-16.00
- Week 2: Tuesday the 4th of May 14.00-16.00
- Week 3: Monday the 10th of May 14.00-16.00
- Week 4: Monday the 17th of May 14.00 16.00

Office hours take place from week 1 to 4 on Fridays 4-6 pm. You need to book a slot beforehand using this link. Please use this Zoom Link to attend the office hour sessions in the time you have booked.

Class Assessment

You will receive a coding assignment on the last session of the class, due by the Friday of the 5th week via email to the instructor (ashrakat.elshehawy@politics.ox.ac.uk).

Course Structure

0.1 Introduction To Computational Text Analysis

- Corpus Creation: Opening and reading in large text corpora in Python
- Manipulating Text, Understanding Text-Cleaning
- Basic Text Processing Operations: Tokenization, Stemming, Lemmatization

Readings

- Wilkerson, J. and Casas, A., 2017. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science, 20,* pp.529-544.
- Monroe, Burt and Phil Schrodt, 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis 16,* 4, 351-355

0.2 Natural Language Processing: Building Pre-processing Pipeline + Dictionaries and Keyword Lists

- Pre-Processing Pipeline
- Pos-Tagging
- Name-Entity-Recognition (NER)
- Building and Using Dictionaries and Keyword Lists
- The Power of Counting Words

Reading

- Denny, M. J. and Spirling, A., 2018. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis 26*(2), pp. 168–189. https://doi.org/10.1017/pan.2017.44.
- Nielsen, Richard. 2019. "What Counting Words Can Teach Us About Middle East Politics" APSA MENA Newsletter (LINK)
- Young, L. and Soroka, S., 2012. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication, 29*(2), pp.205-231.

0.3 Vector Space Representation and Unsupervised Techniques

- Word-Embeddings
- Cosine Similarity
- Topic Modelling
- Clustering

Readings

- Boussalis, C. and Coan, T.G., 2016. Text-mining the Signals of Climate Change Doubt. *Global Environmental Change, 36*, pp.89-100.
- Elshehawy, Ashrakat, Konstantin Gavras, Nikolay Marinov, Federico Nanni, and Harald Schoen. How Hybrid Regimes Use Propaganda for Election Interventions: The Refugee Crisis in Germany. *Available at SSRN* here (2019).
- Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A. and Tingley, D., 2015. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis, 23(2),* pp.254-277.

0.4 Supervised Machine Learning Techniques

- Classification
- Evaluation Techniques
- Sentiment Analysis

Readings

- D'Orazio, V., Landis, S.T., Palmer, G. and Schrodt, P., 2014. Separating the Wheat from the Chaff: Applications of Automated Document Classification using Support Vector Machines. *Political Analysis, 22*(2), pp.224-242.
- Yu, B., Kaufmann, S. and Diermeier, D., 2008. Classifying Party Affiliation from Political Speech. *Journal of Information Technology Politics, 5*(1), pp.33-48.
- Grimmer, J., Westwood, S.J. and Messing, S., 2014. *The Impression of Influence: Legislator Communication, Representation, and Democratic accountability.* Princeton University Press. [**Only Chapter 3** - SOLO link]