POL 42050, Quantitative Text Analysis

- dr. Martijn Schoonvelde (module coordinator)
  - School of Politics & International Relations
  - martijn.schoonvelde@ucd.ie
  - Office hours: by appointment, Newman Building, room F303.
  - Lectures: Friday 9-11.
    - Weeks 1-4: G6, Daedalus Building.
    - Weeks 5-12: G-30PCLab, Agriculture and Food Science Centre.

Course introduction

Automated text analysis has become very popular in political science over the past years. With the massive availability of text data on the web, political scientists increasingly recognize automated text analysis (or "text as data") as a promising approach for analyzing social and political behavior. This module introduces students to a variety of its methods and tools to learn about, among other things, topical content, latent ideology and sentiment in text. The meetings – which combine lectures and coding sessions – will be hands-on, dealing with practical issues in each step of the research process (ranging from collecting and pre-processing text data to validating and visualizing output of an analysis).

NB: In this module we will use R. Students who have not used R at all will need to familiarize themselves, for example by working their way through a free online R resource, like `https://www.datacamp.com/courses/free-introduction-to-r` or the resources that are listed on `https://www.rstudio.com/online-learning/#R`. Another good resource is Quantitative Politics with R, developed by Erik Gahner Larsen and Zoltán Fazékas: `http://qpolr.com/index.html`. Students working on their own laptops will need to have R and RStudio installed.

Learning Outcomes

This module introduces students to various approaches of automated text analysis in social science research, emphasizing hands-on analysis of real (political) texts. Students will learn how to extract useful quantities of interest from text, evaluate the outcomes and write up the results of an analysis that uses automated text analysis. Furthermore, students will be able to critically evaluate (social science) research that uses automated text analysis methods.

Assessment

Students are assessed on the basis of 3 components:

1. **Two coding assignments (15% each, 30% total towards the final grade)**

   - The coding assignments are designed to experience the workflow of a text research project. The first assignment will concern getting from text to data that can be analyzed. The second assignment will involve applying some of the methods we discuss in class to this data. Both assignments rely on the EUSpeech dataset.

2. **Presentation (20% towards the final grade)**

   - All students will give a presentation of a small research project they conducted using (one of) the methods discussed in class. This presentation should at least contain a research question, a discussion of the text sources, as well as the steps to address this question using the methods discussed in class. **NB**: Depending on time and enrollment, students will also act as discussant of the research project of another student with the goal of providing constructive comments to improve their work.

3. **Paper (50% towards the final grade)**

   - All students will hand in a paper note of about 2500-3000 words (excluding references and appendices) in which they succinctly but clearly write up the results of a small research project of their choosing using quantitative text analysis methods. Students are free to collect their own data or use existing data (like the EUSpeech dataset or a replication file from a published research paper). Creativity is encouraged.

     This paper should contain the following elements:

     (a) Introduction & research question (±500 words): introduction to the topic.
     (b) Data & methods (±500 words): description of the data sources as well as the methods employed.
     (c) Analysis: (±1000 words): a discussion (with figures and tables) of the results of the analysis.
     (d) Conclusion (±500 words): a brief evaluation of the results and steps to push the research forward.

Participation

This module will involve a lot of reading, some of which is technical. I expect that you come to class prepared, having read all required papers, and ready to discuss your questions, criticisms and thoughts.

Late submission policy

All written work must be submitted on or before the due dates. Students will lose one point of a grade per working day late or part thereof (taking B+, B and B- to be "points" of a grade), and receive an NG (no grade) for essays over 1 week late. Exemptions will only be made in extenuating circumstances and need to be requested in writing. Note that "bad planning" and "work commitments" do not count as extenuating circumstances.

Plagiarism

Although this should be obvious, plagiarism – copying someone else's text without acknowledgement or beyond "fair use" quantities – is not allowed, including self-plagiarism. UCD policies concerning plagiarism can be found online. A more extensive description of what is plagiarism and what is not can be found at the UCD Library website.

Course outline

** This outline serves a general plan for the course; deviations (announced) may be necessary.*

Week 1

- Introduction to text as data. Introduction to EUSpeech, a dataset which will use for running examples: `https://dataverse.harvard.edu/dataverse/euspeech`. Introduction to R and Rstudio.
  - **Required reading**:
    - Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. and Pinker, S., (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331(6014)*, 176–182.
    - Schumacher, G., Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). EUSpeech: a new dataset of EU elite speeches. In: *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, 75–80.
    - Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annual Review of Political Science 2*0: 529– 544.

Week 2

- A survey of automated text analysis in political science. Supervised and unsupervised methods. Validation, validation, validation. Text Analysis in R.
  - **Required reading**:
    - Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21(3)*, 267–297.
    - Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11(4)*, 245–265.
    - Benoit, K., Watanabe, K., Wang, H, Nulty, P., Obeng, A., Müller, & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3(30)*, 774.

Week 3

- Pre-processing data. Going from text to data (including a few notes of caution). Discussion of the research paper.
  - **Required reading**:
    - Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis, 26(2)*, 168–189.

* Schoonvelde, M., Schumacher, G. and Bakker, B.N., (2019). Friends with text as data benefits: assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology, 7(1)*, 124–143.

Week 4

- Getting data. Web scraping. API's
  - **Required reading**:
    * Webscraping using R, tutorial by Chris Bail: `https://cbail.github.io/SICSS_Screenscraping_in_R.html`
    * Collecting tweets using R: `https://bit.ly/2JJhm11`
    * Steinert-Threlkeld, Z. (2017). Spontaneous collective action: Peripheral mobilization during the Arab Spring. *American Political Science Review, 111(2)*: 379–403.
  - **Coding Assignment 2 Due**

Week 5

- Describing and comparing texts: readability, distinctive features, text similarity
  - **Required reading**:
    * Chapters 3 and 4 of Silge, J., & Robinson, D. (2018). Text Mining with R: A Tidy Approach. O'Reilly Media, Inc. Available at `https://www.tidytextmining.com`
    * Cross, J. & Hermansson, H., (2017). Legislative amendments and informal politics in the European Union: A text reuse approach. *European Union Politics, 18(4)*: 581–602.
    * Bischof, D. & Senninger, R., (2018). Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research, 57(2)*: 473–495.
  - **Coding Assignment 1 Due**

Week 6

- Using and developing dictionaries to measure emotions, morality, populism, personality, and other speaker characteristics.
  - **Required reading**:
    * Pennebaker J. & King, L. (1999) Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology, 77(6)*, 1296–1312.
    * Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29(2)*, 205–231.
    * Kraft, P. (2018). Measuring morality in political attitude expression. *Journal of Politics, 80(3)*: 1028–1033.
    * Hawkins, K. & Castanho Silva, B. (2018). Text Analysis: Big Data Approaches. In: *The Ideational Approach to Populism: Theory, Method & Analysis*, edited by Kirk A. Hawkins, Ryan Carlin, Levente Littvay, and Cristóbal Rovira Kaltwasser. London: Routledge.
    * Ramey, A. J., Klingler, J. D., & Hollibaugh, G. E. (2019). Measuring elite personality using speech. *Political Science Research and Methods, 7(1)*,163–184.

Week 7

- Supervised and unsupervised methods to locate text on an underlying (political) dimension. How do they work? And how should we interpret them?
  - **Required reading**:
    * Slapin J. & Proksch S. (2008). A scaling model for estimating time-serial positions from texts. *American Journal of Political Science 52*, 705–722.
    * Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M. E., & Kurrild-Klitgaard, P. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. Research & Politics, 2(2).
    * Daniel Schwarz, Denise Traber, & Kenneth Benoit (2017). Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods 5(2):* 379–396.

Week 8

- Topic models, unsupervised models for summarizing what a text is about. How do they work? And how should we interpret them?
  - **Required reading**:
    * Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4)*, 77–84.
    * Roberts, M et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58(4)*, 1064–1082.
    * Boussalis, C. & Coan, T. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*: 89–100.

Week 9

- New developments in data: (i) images as data, (ii) automated speech recognition, (iii) machine translation.
  - **Required reading**:
    * Proksch, S.O., Wratil, C. and Wäckerle, J., (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, 1–21
    * De Vries, E., Schoonvelde, M. & Schumacher, G., (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis, 26(4)*, 417–430.
    * Casas, A. & Williams, N.W., (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly, 72(2)*: 360–375.
  - **13:00: Coding Assignment 2 Due**

Week 10

- New developments in modeling: (i) word embedding models, (ii) ltta.
  - Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š. & Sedlmair, M., (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12(2-3)*, 140–157.
  - Kleinberg, B., Mozes, M., & van der Vegt, I. (2018). Identifying the sentiment styles of YouTube's vloggers, EMNLP 2018.

Week 11

- Conclusion and loose ends

Week 12

- Student presentations.