Text Analysis for Applied Social Science.

Aquellas personas interesadas en asistir al curso deberán contactar con Magdalena Nebreda (**secretaria@march.uc3m.es**) antes del 20 de mayo, y entregar un CV y una breve explicación acerca de su interés en el mismo. Tendrá lugar un proceso de selección en el caso de que se reciban un gran número de solicitudes: la relación de personas seleccionadas se notificará antes del 25 de mayo.

Descripción del curso

Statistical analysis of text data has become increasingly common in the social sciences (Grimmer and Stewart, 2013). Applications can be found in political science, economics, sociology, and psychology, for example. In this week long workshop we introduce scholars to the necessary tools for doing text analysis in a rigorous, replicable, way. We cover both pragmatic aspects but also cover the statistical details of workhorse text analysis models.

Programa del curso

Para el programa completo ver (**aqui**)

Day 1: Introduction to Text Analysis

This introduction will introduce an overview of text analysis as a methodology. It will begin to introduce text and the basics of text processing necessary to use these tools. This unit will cover:

- Text analysis: an introduction (Monroe and Schrodt, 2008)
- Overview of approaches text analysis methods (Grimmer and Stewart, 2013)
- Collection of text-scraping software (Jackman, 2006; Pilgrim, 2000) (the latter at **http://www.diveintopython.net/**)
- Introduction to basic stemming and lemmatization **http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html**

Day 2: Word Counts and Basic Text Manipulations

This unit will also discuss using word counts for text data. It will introduce software to count words and software to identify discriminating words. This unit will cover:

- Yoshikoder: software for word counts **http://www.yoshikoder.org/courses/apsa2006/apsa-yk.pdf**
- Word counts in action (Young and Soroka, 2012)
- The pitfalls of word counts (Loughran and McDonald, 2011)
- Multinomial Inverse Regression (Taddy, 2013)

Day 3: Supervised Text Methods

This unit will focus on supervised methods for text analysis. Supervised methods leverage some form of human training or guidance which is then used directly in the analysis of textual data. We will cover the statistical foundations of the models and describe their use. This unit will cover:

- ReadMe (Hopkins and King, 2010)
- Classifying political parties from speech (Yu, Kaufmann and Diermeier, 2008)

- Classiffiers and ensembles (Hillard, Purpura and Wilkerson, 2008)
- RTextTools (Jurka et al., 2011)

Day 4: Unsupervised Text Methods

This unit will focus on unsupervised methods for text analysis. Unsupervised methods leverage use statistical tools to discover common patterns in textual data, which then require human interpretation and validation. We will cover the statistical foundations of the models and describe their use. This unit will cover:

- Introduction to inference for latent variable models (Bishop et al., 2006, Chapter 1)
- Latent Dirichelet Allocation (Blei, Ng and Jordan, 2003)
- Structural Topic Models (Roberts et al., 2014, 2013)
- Clustering (Grimmer and King, 2011)

Day 5: New Applications

This Final day will cover applications of the text analysis methods described above to interesting social science questions.

- Reverse engineering censorship in China (King, Pan and Roberts, 2013, 2014)
- Text analysis for comparative politics (Lucas et al., 2015)
- Measuring political communication (Grimmer, 2010)
- Measuring anti-Americanism (Jamal et al., 2014)

Statistical Packages

Throughout the course we will leverage several statistical packages that we or others have contributed to the open source community. These packages will be helpful for students wishing to complete optional workshops. These include:

- Python for web-scraping
- Python package BeautifulSoup
- Yoshikoder for word counts http://www.yoshikoder.org/
- R package textir for multinomial inverse regression
- R package ReadMe
- R package RTextTools for classiffiers
- R package implements the Structural Topic Model (Roberts, Stewart and Tingley, Submitted) (available at **www.structuraltopicmodel.com)**