Introduction to Computational Social Science:

Application of Machine Learning Approaches to Text and Survey Analytics

Duration: 18 hours (3 X 6 hours)

Description:

This workshop provides an introduction to Computational Social Science by bringing together the analysis of unstructured (text) and structured (surveys) data in social research. The practical exercises in class will be implemented in Python and some supplementary examples will be provided in R (based on *Computational Analysis of Communication*, by Van Atteveldt, Thrilling & Arcila). Firstly, it focuses on the application of the supervised sentiment analysis technique to real-time Twitter messages that can help social researchers to detect social trends or citizens' attitudes. Participants will learn the vary basis of programming in Python and will be able to retrieve tweets with Streaming and Rest APIs of Twitter. In addition, they will learn the basis of natural language processing (tokenizing, stemming, lemmatizing, stop words, frequency of words, lexicon based sentiment analysis, etc.) using the package NLTK and will learn the first notions of supervised machine learning applied to unstructured data with different algorithms implemented in Scikit-Learn (naive bayes, logistic regression, support vector machines, etc.). The participants will train a model with positive and negative message and will use this model to predict and visualise in real-time the sentiments of tweets filtered by a keyword. Secondly, it addresses the basics of predictive analytics using structured data to mine survey data. It includes reading and managing different data formats, data wrangling and exploratory data analysis, as well as the application of machine learning algorithms over survey data in Python.

Day 1 (6 hours)
-Introduction to Computational Social Science
-Basic programming on Python 3.4
-Data retrieval of tweets with the Streaming and Rest APIs of Twitter

Day 2 (6 hours)
-Natural language processing with NLTK
-Supervised machine learning with Scikit-Learn
-Supervised sentiment analysis and visualisation of tweets in real-time

Day 3 (6 hours)
-Introduction to predictive analytics
-Data formats and ingestion of data
-Exploratory data analysis
-Machine learning to survey data

Instructor:
Dr Carlos Arcila Calderón, Assistant Professor at the Department of Sociology and Communication at the University of Salamanca, Spain. PhD in Communications and Master in Data Science. Specialist in computational methods in social sciences. (http://diarium.usal.es/carcila/english/)

References

Arcila, C.; Barbosa, E. & Cabezuelo, F. (2016). Técnicas Big Data: Análisis de textos a gran escala para la investigación científica y periodística. El Profesional de la Información, 25 (4), 623-631

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Van Atteveldt, W., Thrilling, D. & Arcila, C. (2020). *Computational Analysis of Communication*. Wiley. Manuscript in preparation.