

Big Data 1 (Social Media Research)

Duration: 12 hours (3 X 4 hours)

Description:

This workshop 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.

Day 1 (4 hours)

- Introduction to Computational Methods
- Basic programming on Python 3.4

Day 2 (4 hours)

- Data retrieval of tweets with the Streaming and Rest APIs of Twitter
- Natural language processing with NLTK

Day 3 (4 hours)

- Supervised machine learning with Scikit-Learn
- Supervised sentiment analysis and visualisation of tweets in realtime

Course prerequisites: Some bases in statistics desirable.

Software used: We will introduce the language Python



Instructor:

Dr Carlos Arcila Calderón, Associate 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.