Text mining and document analysis

TEACHER: Alessio Farcomeni
EMAIL: alessio.farcomeni@uniroma2.it

COURSE DESCRIPTION

We will present techniques for analysis of textual data. The course discusses at first data engineering, i.e., how to process text and how to obtain a data matrix from a text corpus. Analysis will involve simple descriptive tools for text corpi, summarisation, keyword extraction. We will then discuss the uses of text corpi in supervised and supervised learning tasks. Specific methods will be discussed for topic modeling and sentiment analysis. Case studies will involve both short (e.g., arising from social media) and longer texts (e.g., newspaper articles).

PRE-REQUISITES

The only pre-requisite is some ability to work with the *R* software. Knowledge of statistical learning methods both in supervised and unsupervised contexts would be useful.

PROGRAM

- 1. Introduction to text mining and document analysis.
- 2. N-grams. Tokenisation. Part of speech tagging and named entity recognition.
- 3. Descriptive tools for text mining
- 4 Topic modeling. Latent Dirichlet allocation
  4.1 Word embeddings and *top2vec*
- 5. Supervised and unsupervised sentiment analysis

TEXTBOOK

Silve J. and Robinson D. (2017) *Text Mining with R*, O'Reilly Media, Inc.
(https://www.tidytextmining.com/)