Text Mining 2020

Project guidelines

The course is graded in two parts: a written examination (30%) and a project (70%) conducted in groups of three to four students that will be organized after the final registration date.

Originality and style:

The project consists of an analysis of an original corpus of text. This means that re-analyses of existing datasets, whether provided during the course or found online elsewhere, are not enough. Instead, you are expected to identify your own source of text data and perform any necessary scraping or parsing tasks to construct the dataset. Each part of your analysis should be supported by well-designed and relevant charts and graphs. These should be adequately labelled with captions and legends.

Completeness:

Your analysis should to match the complexity and uniqueness of your dataset; your analysis must be comprehensive and not leave major aspects of your dataset unaddressed. The project must include (all points):

- Original data gathering (like scraping, etc.)
- Cleaning and exploratory data analysis (like stemming, lemmatization, etc.)
- Unsupervised analysis (like topic analysis, etc.)
- Supervised learning (like text classification)

Deliverable:

Return a full PDF or HTML report of your analysis with reproducible code and figures, not exceeding 30 pages, including title page and appendices.

Deadlines:

Monday, 14th of December at midnight (23:59). Any late deposit will encounter a penalty of 0.2 points (out of 6) per started hour.

Presentations:

On the 17th of December a series of oral presentations will be held from 8:30–12:00, to present your work. Each group will be given 30 minutes (20+10 minutes). Slides should be sent before the presentation (pdf or ppt).