SDS 2020 - M2: Group Assignment

Introduction

In the M2 sessions, online courses and assignments you learned how to work with relational data and text in different contexts and on various levels of aggregation. These skills allow you to access and process a vast range of data. The analysis of textual as well as relational data indeed has a lot of synergies. Often measures of relationships can be created based on text data. Likewise, often the analysis of text data can be enhanced by analyzing the relationship between keywords, documents, and the like.

Now it's time to get creative. We would like you to carry out an own analysis on a self-chosen topic (and on self-chosen data). This analysis should be interesting and informative and contain elements and methods from network analysis as well as natural language processing. The balance between the two fields is up to you, as long as the application of both is non-trivial.

Task description

Data & Problem identification

In this exercise, you are asked to choose and obtain a dataset you consider interesting and appropriate for the tasks required. You are welcome to use existing datasets for language and networks but at this stage you could also consider getting your own data (e.g. Twitter API, Instagram, news repositories etc.)

The data should be large enough and of proper granularity to be interesting for NLP and network analysis techniques. If you are in doubt, please reach out.

What we expect you to do:

- Identify an interesting problem that can be tackled using data science techniques applied to natural language and networks.
- Select and obtain relevant data to do so.
- Clean and manipulate the data to make it useful.
- Carry out an exploratory data analysis to provide intuition into the content of the data, and interesting relationships to be found in it.
- Use unsupervised ML techniques to discover relationships within the data such as interesting topics or latent network structures.
- Use supervised ML techniques to create models that predict an outcome of interest.
- Document your workflow in a reconstructable manner.
- Report your findings in an accessible manner.

Analysis pipeline

The analysis to be carried out by you has to contain elements of **data manipulation**, **exploration**, **unsupervised** and **supervised ML** as applied to **relational** and **language data**.

In the best case, you combine network data with language elements. Twitter is a good (and easy) example, as you can, for instance, combine mention-networks with sentiments expressed in the tweets. The article below is a creative example of that (with a rather small NLP part).

Liu, Z., & Weber, I. (2014). Is Twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In International Conference on Social Informatics (pp. 336-347). Springer, Cham.

- Definition of a problem statement and a short outline of the implementation
- Description of data acquisition / how it was collected (by you or the publisher of the data)
- Data preparation (general)
  - Data cleaning (if needed)
  - Recoding (label encoding, dummy creation etc.)
  - Merging and wrangling (if needed)
- Missing data imputation (if applicable and deemed relevant)
- Network Data - preparation
  - Extraction and formatting
  - Creation of functional graphs with relevant attributes
- NLP - preparation
  - Extraction & Cleaning
  - Tokenization
  - Filtering & Lemmatization / Stemming (if needed)
- Network analysis
  - Calculation of relevant indicators on different levels / EDA
  - Projection (in the case of bipartite graphs)
  - Identification of community structures
- NLP
  - EDA / simple frequency-based analysis
  - Simple vectorization (BoW, Tf-idf)
  - Topic modelling / Clustering (LDA / LSA)
  - Embedding-model based vectorization (Word2Vec, Fasttext, GloVe)
- Supervised / Unsupervised ML
  - Try to link your results from network analysis or NLP with a more traditional ML problem.

Many of the steps are optional. So choose which methods you deem helpful and relevant to explore your chosen problem.

Note: Quality > Quantity. Consider which analysis, summarization, and visualization adds value. Excessive and unselective outputs (e.g. running 20 different models without providing a reason for, providing all possibilities of different plots without discussing and evaluating the insights gained from it) will not be considered helpful but rather distracting.

Some inspirational examples (non-binding, and non-exhaustive):

1. You obtain a dataset with tweets on a current debate (e.g. #MeeToo) and try to map the discourse.
   - You perform "naive" NLP, counting handles, hashtags, basic plotting etc. to get some overview.
   - You perform "out-of-the-box" sentiment analysis and plot tweets on a map, colouring by sentiment.
   - You perform topic modelling and identify the sub-discussions.
   - Isolating handles/retweets, you identify some interaction patterns, use network indicator to identify thought leaders or conflicting communities as well as people that try to negotiate between positions.

2. You obtain a bibliographic dataset on a field of study (or from an entity such as a university) of interest, e.g., from scopus.
   - You perform a network analysis on different levels of aggregations, identifying key publications, scientists etc.
   - You run a topic model to identify relevant discourses.
   - You might then answer questions such as: Did the discourses change over time? In case so, who or what drives these changes?

Documentation and Deliverables

You are asked to hand in two different report formats, namely:

1. Functional computational notebook
2. Stakeholder report

Computational Notebook

The notebook targets a **machine-learning literate audience**. Here you can go deeper into the technical details and method considerations. Provide thorough documentation of the whole process, the used methods. Describe the **intuition** behind the selected and used methods, **justify** choices made, and **interpret** results (e.g. Why scaling? Why splitting the data? Why certain tabulations and visualizations? What can be seen from ... ?, How did you select a particular algorithm? Why did you scale features in one way or another?). Please provide the notebook as a PDF or HTML (Knittered from rmd or converted ipynb, when HTML zipped) with a public link to a functional Colab (Python) version (test it beforehand in incognito/private mode of your browser)

Stakeholder Report

The stakeholder report (simple PDF, no code) summarises the analysis for a **non-technical audience**. Here you don't need to discuss alternative approaches to standardization and alike. Instead, you should try to explain the analysis and results, emphasizing its meaning and interpretation. Imagine it as a report of the **project outcome**, as you would explain it to a general audience.to Aim at a length of **not more than 5 pages**, including tables & visualizations.