Data Science Toolbox Assessed Coursework 2: Text as Data

Deadline: Wednesday Noon, Week 17

Group Project description

You will **choose an application domain** that your group will work with for Assessment 2. Your challenge is to apply **topic-modelling** (or other models of text) to that data.

You should:

a) choose an appropriate **scientific/analysis question**;
b) use an appropriate strategy to **learn any parameters** of the model(s);
c) use an appropriate strategy to **learn about the performance** of the model(s);
d) apply the model(s) to achieve your question.

Appropriate Methods

Models and scientific questions include but are not limited to:

- Bag-of-words models
- Latent Dirichlet Allocation
- Recommender Systems
- Systems that translate data into vectors, such as Neural Networks and Autoencoders
- Data visualisation/ Exploratory Data Analysis
- Comparison of approaches
- Importance of pre-processing

Advice on assessment

You will be assessed on:

a) the implementation of the model, that is, you can be awarded credit for:
    - additional implementation if an off-the-shelf implementation falls short,
    - exploring multiple implementations,
    - examining the mathematical details of choices.
b) the application of the model to a complex problem, that is you can be awarded credit for:
    - identifying an appropriate dataset;
    - using your understanding of the structure of datasets to make arguments comparing the dataset you chose to one that you might encounter in a "real" setting;
    - plotting or otherwise describing various inputs, outputs, or parameters.

c) the correctness of the methods used to achieve their stated goals.
d) the robustness of the results in supporting the conclusions.

In order to be attributed credit for your efforts to choose appropriate data, ensure that you document the data exploration process. You should aim to demonstrate diligence that there is no more appropriate data source in your chosen category.

You do not need to excel in all areas in order to get a high mark. Instead, you need to perform robustly in all areas and additionally demonstrate insight somewhere to score highly.

Topics

Here are suggested example datasets. You are free to find and analyse other datasets on discussion with the course tutor. It is understood that public domain data is sparse in some spaces, and you will be rewarded for creative use of available data.

- Cyber Security, for example:
  - Meta-data surrounding attacks
  - Meta-data surrounding malicious actors
  - Source code modelling
  - Byte code/instruction set modelling

Individual reflection description

- Discuss the rationale behind the inference goal that you selected;
- Discuss the model(s) that you explored;
- Relate your data source to those you might encounter in a real-world problem;
- Discuss a mathematical issue raised in the project, different to those of your group.