Project 2: Sentiment analysis & Topic modelling

Text Mining (G00C8A)

Martial Luyts

---

1 Sentiment analysis (3pts)

Description of the dataset

IMDb (an acronym for Internet Movie Database) is an online database of information related to movies, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

Here, sentiments on movie reviews are collected and stored as .csv file, consisting of 2 attributes:

- review: review of a particular movie;
- sentiment: sentiment related to the review, i.e., positive or negative.

Instructions of this analysis

Perform the following tasks:

1. Train a feedforward neural language model from scratch (you can choose the complexity of your model) that predicts the sentiment score (positive or negative) of a given review. To perform this task, the dataset needs to be split a priori in a 80-20 train-validation dataset, respectively, at random. The inputs of your feedforward neural network can be either word embedding (with vector dimension 100), or specific features, depending on your personal preferences. Remark: To perform this task, it can be that you need to perform some pre-processing steps.
2. Evaluate the performance of your trained neural network model by calculating the precision, recall and F1-score measurements on the validation dataset.

2 Topic modelling (3pts)

Description of the dataset

News headlines are important for newspapers and magazines to attract as many people as possible to their articles and broadcasts.

Here, a corpus of over one million news articles published by the Australian news source ABC (Australian Broadcasting Corporation) is obtained over a period of nineteen years. It is stored as .csv file, consisting of 2 attributes:

- publish_date: Date of publishing for the article in yyyyMMdd format
- headline_text: Text of the headline in Ascii, English, lowercase

Instructions of this analysis

Perform the following tasks:

1. Apply the Latent Semantic Analysis and Latent Dirichlet Allocation technique to study the topic focus of ABC's news headlines. Characterize them by exploring the most frequent words in each topic.
2. How do these topics evolve through time in the ABC news headlines?

You are expected to make this project in Jupyter Notebook, consisting of clear comments and codes such that the steps and conclusions made by you can easily be followed and verified. You need to upload this file (in .ipynb extension) **before January 20th, 2024 (23h59 BE time)**, on **Toledo**.

This project will count for **6 points of the total grade**.

Good luck!