Assignment 1 – Natural Language Processing 2024 – 2025 Instructions

- In this assignment, you will work on pre-processing text and solving a classification problem using natural language processing. The classification task is aimed at predicting stock price movements based on newspaper articles.
- Assignments are to be done individually.
- Use Python to solve the exercise.
- Deliverables:
    - The file [STUDENTID]-[LASTNAME].pdf
      which contains your report with answers to the questions.
    - Python code that you have developed to solve the assignment. Keep the code of methods that did not make it into the final model, but do not submit code that does not work. The code shows that you have experimented with different methods. Make sure that this code is readable, and not unnecessarily long. If you developed multiple files, place them in a zip file named [STUDENTID]-[LASTNAME]-code.zip.
    - Submit all deliverables electronically through CANVAS. Make sure that you submit everything in a single zip file that contains the solutions to the assignment (report and source code). Name it [STUDENTID]-[LASTNAME]-A1.zip
- Deadline for submission of solutions is October 10th, 2024, 22:00 hrs.
- Be concise and to the point.
- Use correct terminology.

Questions regarding the assignment can be asked through Canvas, or during the sessions on Thursday mornings. Good luck!

Student assignment 1

For this assignment, you are an analyst working at a company investing in the stock market. Innovative ideas are appreciated, and you want to build a model to predict stock price movements. During your master's at JADS you followed the course NLP. There you learned about text classification and want to use this to your advantage. What if you write a text classification algorithm that classifies news articles based on whether their associated company's stock goes up or down? You pitch this idea to your manager. He is skeptical about the results but sees enough potential to let you try this out. You have three weeks to build a model and write a report on your findings. One of your colleagues mentioned having a dataset named 'us_equities_news_dataset.csv', which you decide to use. You decide to develop a prediction model that predicts the stock prices for Nvidia. You have found a dataset on Yahoo Finance that contains all the historical data. Now, you need to filter documents from the dataset to only retain related articles. Based on each article's publication date, you find the opening- and closing stock price for that day. If the opening price is higher than the closing price, you label your target variable as 0 and label your other targets as 1.

Preprocessing and Analysis

Before you start with your predictions, you want to learn more about the data. Perform the preprocessing steps that are likely to result in the best predictions and motivate why. For each step describe what you did and why. Also, provide an overview with descriptive statistics with at least the number of articles, average words per article, number of unique words and the lexical richness. Also, show the most common 50 words in the corpus after stop word removal and show the 20 words most indicative to each class.

Text representation and modeling

From your NLP course at JADS you remember that there are different methods of text representation. You remember that TF-IDF is a widely used method of representing documents. You decide to test this representation for your documents. You calculate the TF-IDF representation of all the documents in your dataset. Then, you look at two documents that have very similar TF-IDF representations. You inspect these documents to see if the articles are also similar. You also look at two documents that have very different TF-IDF representations. You include this analysis in your report.

You remember that Word2Vec representations might also be a good choice to represent the documents. You decide to train your Word2Vec on the dataset instead of downloading a pretrained one. However, you are aware of the pros and cons of both approaches. What are these? Now, there are several modeling decisions you need to make. Which Word2Vec algorithm will you use (CBOW or Skip-Gram)? How do you select your hyperparameters? Do you train your model on the whole corpus or only on the articles related to Nvidia? Lastly, how do you convert your word embeddings (matrix) into a document embedding (vector)?

Predictions and evaluation

Once you have created your document representations, you will need to make predictions and evaluate your model. For the classification, use Naïve Bayes and one other model. Which procedure do you follow for training and evaluation? How will you evaluate your model? If you had to choose one evaluation metric only, which one would it be?

Report

Prepare a report for your manager in which you explain the modeling process, the outcomes obtained and a reflection on the applicability of the model. Justify your choices by referring to the scientific literature, where appropriate. Your manager has a full agenda, so the report should be concise. Make sure to use at most five A4 pages (excluding figures, tables and references). Consider at least the following points for your report.

1. Dataset creation - How do you create/define the relevant corpus for your experiments (document filtering)? How do you create the labels for the target variable?
2. Preprocessing - Which preprocessing steps did you use, and why? Also mention the one that you chose not to use, and why, if relevant.
3. Analysis – Provide an overview of descriptive statistics, the 100 most common words, the 20 most indicative words per class. Are all words in the corpus evenly used? Are the labels in the dataset balanced?
4. Numerical representation – What were your findings when using TF-IDF for document representation? Also, what are the pros and cons of training your own Word2Vec representation vs. downloading a pretrained one? Which algorithm did you use and why (CBOW vs Skip-Gram)? Do you train the model on the whole corpus or only on the selected articles? How do you convert your word embeddings into a document embedding?
5. Training procedure – How do you train your model, i.e. obtain the optimal parameters? What is the experimental design that you use for training and evaluating the model?
6. Evaluation – How did your models perform? Which model performed better? Why is this the case? Motivate why you chose your evaluation metric.
7. Discussion & Limitations – What needs to be done to implement this in practice? What are the limitations or shortcomings of the approach that you have used? What could have been a better design (if any)? Think carefully about your target variable.
8. Any other relevant details.

You realize that three weeks might pass by soon if you don't start quickly.

Good luck.