CSCI 4146/6409 - Process of Data Science (Summer 2023)

Assignment 5: Text Mining & Natural Language Processing

In this assignment, you'll explore different aspects of text mining and natural language processing (NLP) and apply your understanding to analyze emotions in text data.

Due Monday July 17, by 11:59 pm - see submission details below.

Datasets: For this assignment, use the IMDB movie reviews dataset (https://ai.stanford.edu/~amaas/data/sentiment/) for sentiment analysis and the SemEval-2018 Task 1: Affect in Tweets dataset (https://competitions.codalab.org/competitions/17751) for emotion classification.

Q1: Preprocessing, Exploratory Data Analysis and Part-of-Speech Tagging (Applies only to IMDB movie review dataset) [1]

- Load the datasets.
- Perform exploratory data analysis: check the distribution of sentiments/emotions, the number of unique words, average review length, etc. Include any interesting observations.
- Preprocess the text data: tokenize the reviews/tweets, remove stop words, and apply stemming/lemmatization.
- Perform part-of-speech (POS) tagging on the preprocessed text. Briefly discuss the importance and potential uses of POS tagging in NLP.

Q2: Sentiment Analysis with Bag of Words (BoW) and TF-IDF [1]

- Convert the preprocessed IMDB movie reviews into a matrix of token counts with the CountVectorizer from sklearn.
- Also, represent the reviews using TF-IDF and briefly explain the concept and its importance in text mining.
- Divide your data into a training set and a test set (If required).
- Train a logistic regression classifier on your BoW and TF-IDF representations. Report the accuracy, precision, recall, and F1-score on your test set. Compare the performances of BoW and TF-IDF.

Q3: Sentiment Analysis with Word Embeddings [1]

- Using the Word2Vec model, convert the preprocessed movie reviews into vectors.
- Train the same type of classifier (logistic regression) on these new features. Again, report the accuracy, precision, recall, and F1-score on your test set.

- Compare the performance of the BoW, TF-IDF, and Word2Vec models. Discuss any improvements or regressions observed.

Q4: Emotion Classification [1]

- On the SemEval-2018 Task 1 dataset, build two classification models to classify tweets into various emotion categories: one using the BoW method and one using Word2Vec.
- Compare the performance of the two models. Discuss which method worked better for this task. Present any theories or observations regarding these results. Consider addressing class imbalance, if any.

Q5: Topic Modeling [1]

- On the SemEval-2018 Task 1 dataset, perform topic modeling using Latent Dirichlet Allocation (LDA).
- Visualize the top words for each topic. Write a brief description of what you think each topic represents based on these words.
- Discuss how well you think the model worked and any interesting findings.

Submission Details

Submissions should be made through Brightspace, adhering to the due date and time specified under the Assignments section. To prepare your assignment solution, make use of the provided assignment template notebook on Brightspace. The detailed requirements for writing and coding can be found within the evaluation rubric document available on the platform. Keep in mind that questions will be graded individually using letter grades, with their respective weights indicated in parentheses.

You may complete the assignment individually or with another individual. In the case of a pair submission, only one student should submit the assignment on Brightspace. All analyses and discussions should be based on your own insights and understanding. While collaboration and discussion among classmates is encouraged, all written, and code work must be your own. Any sources consulted, including websites, textbooks, papers, etc., should be properly cited. Any evidence of plagiarism or academic dishonesty will result in a zero for the assignment.

Your submission should consist of a single Jupyter notebook as well as a generated PDF that contains the compiled results generated by the notebook. This PDF should include both the code and the results as part of the final printout.

Name your files as follows:

- A5-<your_name1>-<your_name2>.ipynb and
- A5-<your_name1>-<your_name2>.pdf.

Failing to submit both files will result in a zero mark for both students.