Natural Language Processing

Midterm Project - Spring 24-25

Project Description

The goal of the midterm project is to implement different natural language processing techniques in a dataset using Python. You should carefully read this document to complete the project successfully.

- The project must be completed in a four-person group.
- Select a dataset from the Dataset List given at the end of this specification file. Your selected dataset must not match the dataset of any other group in your class.
- A spreadsheet will be posted in MS Teams to provide group and dataset information.

Project Deliverables

- Submit the implemented python program ("nlp_mid_project_group_XX.ipynb" file) and the report ("nlp_mid_project_group_XX.pdf" file) in MS Teams. Replace "XX" in all filenames with your group number such as for group-1, the report file name will be "nlp_mid_project_group_01.pdf". See the instruction section below for the report details.
- During the VIVA session, you will bring this implemented program, and we may ask you to execute it. The program file must not contain any comments.

General Instructions

- Make sure your group information is complete in the spreadsheet.
- Make sure you have selected a unique dataset by reviewing the already selected datasets in the spreadsheet.
- At the beginning of the report (after the cover page), write a short description of your selected dataset.
- For each implemented task, write a paragraph in the report. In the paragraph, describe how the related task is implemented in your project, along with the relevant code segment. While writing a description, only write the content (do not write unnecessary content) that is sufficient to understand the solution and its associated code segment. No need to write anything in the report for the pre-tasks.
- **The submission deadline for all deliverables is April 26, 2025, 11:59 PM.**
- **Comments are not allowed in the implemented program.**
- **I will announce the project VIVA schedule in MS Teams.**
- **Please do not copy content from any sources. It will be strictly handled.**

Project Requirements

- **Pre-Task:** Provide your group and dataset information in the spreadsheet by **12-04-2025**.
- **Tasks:** Our task is to apply the Naïve Bayes algorithm to the selected dataset to classify the sentiment appropriately and evaluate the learned model with different metrics that we learned in the class. To complete this task, apply any of the following techniques that are applicable:
    - Tokenization
    - Case folding
    - Synonym substitution
    - Stemming
    - Lemmatization
    - Punctuation removal
    - Stop words removal
    - Vector Semantics
- Each task must be completed in a separate program cell in your python notebook file. For example, a cell can be used to mount to Google Drive, another cell can be used to read the dataset files, and so on. No need to do anything for the pre-task.
- Once you complete your program, execute the entire code, which will generate the output. You must submit your program with the generated output.
- Scikit-Learn supports various implementations of the Naïve Bayes method. You must implement the MultinomialNB version of Naïve Bayes for this project.
- Do not use any library that needs to be installed other than nltk, spaCy, matplotlib, numpy, pandas, and scikit-learn (or any other library that is discussed in the classes).

Dataset List

1. https://www.kaggle.com/datasets/anishdabhane/apple-tweets-sentiment-dataset
2. https://www.kaggle.com/datasets/drishtiagarwal20/brands-and-product-emotions
3. https://www.kaggle.com/datasets/jannesklaas/disasters-on-social-media
4. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
5. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
6. https://www.kaggle.com/datasets/sahilnbajaj/movie-reviews-raw
7. https://www.kaggle.com/datasets/niraliivaghani/flipkart-product-customer-reviews-dataset
8. https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset
9. https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis
10. https://www.kaggle.com/datasets/yash612/stockmarket-sentiment-dataset