Text Mining and Analytics, Spring 2024, Assignment 3

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Due: March 20, 2024

This assignment is a team assignment, where students will work in teams of 2 or 3 students. The goal is to continue the task of song classification based on genre using lyrics. All three tasks should be completed; in a team with 3 students, each member will do one task, and in a team of 2 students, the third task should be done by both students. In your submission, you should explicitly mention who has done which task. There are three tasks as follows.

Task 1: To develop neural network models, there is usually a need for a larger dataset than the one we used in the previous assignment. You have practiced using web scraping in assignment 1. To develop models for tasks 2 and 3, you require a larger dataset of lyrics and their corresponding genres. Considering the same genre as the previous assignment, continue scraping the web for lyrics. You are allowed to use any website; however, your code will be run from scratch, therefore all data collection processes and generating the data files should be done automatically. You should create .csv files as your data file. It is highly possible that the more data you gather, the better you can train models for the next steps. Also, you should exclude any of the songs that are provided as the test set (provided on BrightSpace).

Task 2: After collecting data, you will develop a neural network that uses manually defined features to classify songs. You should go through some sample songs and carefully think about what features are more useful. Here are a few examples:

- Number of words, unique words
- Rhyme density
- Use of pronouns

After extracting features, you should design a feed-forward neural network, and train it with the data from task 1. The only criterion for designing the neural network is that it should be a feed-forward network.

Task 3: To classify the songs, another approach is instead of manually extracting features, use word embeddings. For the words in a lyric, average their embeddings to get a vector representation of the whole lyric. You can use a pre-trained embedding Word2Vec model, or

train your Word2Vec from scratch (or a fastText model); but for this, you need enough data. After you have the vector representation of each lyric, design a feed-forward neural network to classify the lyrics based on their genre. This vector will take in the vector representation as the input.

Deliverables:

Task 1:

- .py file to generate the dataset from scratch
- .CSV files of your dataset. You should have train and validation sets separated
- A ReadME file specifying how to run your code and what output to expected (format of .CSV files)

Tasks 2 and 3 (bullet points for each task):

- Training vs. Validation loss-epoch diagram (.pdf)
- Table of Results showing F1-Score of your model, and F1-Score per genre (.pdf)
- Analysis of the results (.pdf)
- .py file to run your model along with a ReadME file specifying how to run your code

On Brightspace, one student will submit the files (.py, .csv, .pdf, .md/.txt). Please make sure to include your teammates' name and the task they have done on the top of the .pdf file.

For grading, each task will have 6 points, graded individually for each student. 3 points are considered for putting the steps together and providing the final report.