Meeting Updates

Date: October 27, 2023

Completed:

Data gathering: https://www.kaggle.com/datasets/raynardj/imdb-vision-and-nlp/data

Assignments:

a. Start with data pre-processing
   - Data Pre-processing (a, b, c) - Bhumika, Aura (check spacy for NER)
   - Data Pre-processing (e, g, h) - Olivia, Rochan
   - Data Pre-processing: Spell corrector - Varun, Aura (check pyspellchecker)
   - Data Pre-processing: POS Tagger - Abhishek, Miraj

b. Research on different model techniques (to be started by Roger)

c. SQL Queries c/o Roger and Abhishek

Additional Notes:

- Project objectives to be finalized
- Everyone can help on data model technique research once done with the assigned tasks
- Do not forget to add in the references all the links of the sources, and commit it in Github
- Each member has his own assigned task branch on GitHub. Utilize it to experiment with your code and make commits as needed. However, be sure to check the Git Process guidelines for instructions on merging code into the main branch (https://github.com/abccastro/Movie-Sentiment-Analysis/blob/main/reference%20documents/Git%20Process.pdf)
- The data pre-processing is just a guide provided by our professor. **We still need to access if we need to incorporate all the tasks in the list**

Data Pre-processing

a. Basic regular expressions (handle non-grammatical stuff), e.g. URL, email add

b. Contractors, abbreviations and slangs

c. NER (name entity recognizer) - proper nouns; rely on punctuations

d. Spell corrector

e. More Regex (punctuations, lower casing, whitespaces)

f. POS Tagger (optional)

g. Remove stopwords

h. Stemming / Lemmatization (most important for data pre-processing)