FIT5196 Task 2 - Assessment 1 (Development History)

Student Name: Alexandra Goh & Sothearith Tith

Student ID: 29796431 & 27208001

TASK 2

Date: 19/03/2024

We conducted a comprehensive analysis of the requirements, including the structure of the Excel, emoji, and stopwords files, to gain a clearer understanding of our resources and the scope of our work. Following this thorough examination, we moved on to discuss and brainstorm potential approaches we can take to complete the task.

Contribution:

Alexandra:

- Proposed using Regular Expression to break up the snippet field and extract the ‘textOriginal’ fields in all top level comments.

Sothearith:

- Proposed using Python's eval() function to analysis the snippet field and return a dictionary.
- Proposed word processing sequence: remove emoji, remove special character, remove stopwords and tokenise words

Date: 26/03/2024

Having gained a comprehensive understanding of the requirements and the resources at our disposal, we transitioned to the coding phase, as well as documenting our process. We adhered to the planned sequence of word processing, ensuring that each step was executed systematically and efficiently to meet the project's objectives.

Contribution:

Sothearith:

- Read Excel file.
- As the data are positioned differently in each worksheet, rows and columns with null value need to be removed.
- Remove duplicate and correctly combining the data from each sheet.
- Extract the neccessary data, ‘textOriginal’ fields in all top level comments, using eval() function and store the data in pandas dataframe.
- Read emoji from the given files.
- Remove URLs, links, emoji, special characters from each record using apply().

Date: 01/04/2024

Following the thorough cleaning of strings in each row of the dataframe, we moved to detecting and tallying English comments. This step involved scrutinizing each entry to identify and quantify comments written in English. Subsequently, we isolated channels that featured more than 15 English comments. For these selected channels, we proceed to constructing a vocabulary list, capturing the essence and diversity of the language used. Additionally, we developed a sparse representation of the data, which efficiently encapsulated the linguistic patterns and frequencies within the comments, providing a structured and analytical foundation for further examination and processing.

Contribution:

Sothearith:

- Detect English comment using langdetect library.

- Generate a csv file that contains unique channel ids along with the counts of top level comments(all language, and english).
- Get all channel that has more than 15 English comment.
- Tokenising words following regular expression, "[a-zA-Z]+".
- Remove context-independent and context-dependent stopwords.
- Stemmed tokens using the Porter stemmer and remove rare tokens as well as tokens with length less than 3.
- Get first 200 meaningful bigrams and calculate the vocabulary containing both unigrams and bigrams.
- Combine the unigrams and bigrams, sort the list alphabetically in an ascending order and output as vocab.txt.
- Generate the sparse numerical representation using FreqDist() and output as countvec.txt

Google Colab Workbook Link

Task 2: https://colab.research.google.com/drive/1N7byY8F4QLcVQcbqkGPm0flv92p6Okjl?usp=sharing