TASK 5 - Development History

Student Name: Anushka Jemima

Student ID: 33617457

Student Name: Shruthi Shashidhara Shastry

Student ID: 33684669

Date: 30-08-2024

Task 1

Date: 31/07/2023

Today, we went through the assignment specification,discussed which files we are to use and decided the method we would follow for development and started exploring task 1 data.For each task we had decided that each of us would try out multiple methods and then put our final conclusions on a common notebook.

contribution:

Group member 1:

- Explain the files and data structure required for task 1.

  Group member 2:

- Identified the structure of the txt files for task 1.

During this meeting we identified a gap in our scehdules and we realised we couldnt spend time working on the same file at the same time hence we decided to split the tasks there by allowing for more ideas on how to solve the issue and also split the testing required since there were a huge number of files.

Date 11/08/2024

Today, we started reading the data by python and extracted the key fields we needed. We loaded the extracted information into a pandas.And identified the regular expression patterns.

Contribution:

Group member 1:

- Reading and loading data from all txt files into a dataframe and excel file. Test first 7 txt files.

  Group member 2:

- Regular Expressions required to extract each column.Test remaining txt files.

During this task we faced a hurdle in how to store each record , initally we tried extracting the columns that is the gmap_id , review text , rating etc. But this was causing a data mis match , we then decided to first store an entire record in a data frame and then from each record extract out column values.

Date 14/08/2024

Today, we started cleaning the data frame , removing emoji's and any duplicates generate the CSV and perfroming data manipulation required before loading to json file.

Contribution

Group member 1:

- Cleaning of the generated data frame , removing emoji's and identified any inconsistencies.

  Group member 2:

- Code for generating CSV and remaining data manipulation required for json file.

During identfying inconsistencies we faced an issue with the regular expression user_id and user_name had the same values , thats when we decided to split the testing tasks to identify where we selecting the wrong user_id and user_name.

Google Colab Workbook Link

https://colab.research.google.com/drive/1YBQU7JqI7I66w_X5pYns0f8blcARZEZW?usp=sharing

Task 2

Date: 17/08/2023

Today, we began exploring sepcification fro task 2.We discussed and completed the extraction of data.We decided on steps to follow, we then decided we would each try different steps and cross verifying our output files.Our focus was on unigrams first.We dicsuused and shared code reference for each of the tasks we had to do.

contribution:

Group member 1:

- Extracting data from the json file. Stopword removal and tokenization

  Group member 2:

- Ensure data is clean and selection of gmap_id with more than 70 reviews.Rare token removal and stemming.

During this meeting we focused on extracting unigrams , we faced a road block in deciding if we should perfrom all these operations on the entire dictionary after combining it or on each

individual review in each individual gmap_id list. We each tried different methods and then compared our vocab files and had similar outputs , but decided the more optimized method would be to combine all reviews into one list as this would eliminate the duplicates and since we need a vocab dictionary of all gmap_id's and not a vocab dictionary for each gmap_id.

Date: 21/08/2023

Today, we focused on bigram extraction , verification and combining with unigram.

contribution:

Group member 1:

- Extracted bigram from the text and combined with unigram.

  Group member 2:

- Checked if bigram is valid and can be found within a review.

During this meeting we were unsure if bigram should be generated at the beginning or if it should be generated at the end and combined with the unigrams.We again tried different methods and noticed that the bigrams generated at the end after all unigrams had been put into a single list and stop word removed and stemming was done , we recieved odd bigrams in our output and weren't able to to find many of them in an individual review, hence we then created bigram at the beginning after unigram's were created on the dictionary it's self before combining all the tokens , this allowed us to check if a bigram is present in the review and then we combine the tokens to perfrom the remaining steps.

Date: 25/08/2023

Today,we completed the data storing and formating in our output files and testing it against the input files , we then consolidated all our code into a single notebook.

contribution:

Group member 1:

- Data storing of the vocab.txt and testing and checking for bigrams.

  Group member 2:

- Data Storing of the countvec.txt and testing agaisnt each gmap_id.

  During testing of our output files we the countvec file did not pass the test case because we had displayed the token and token frequency instead of token_index: token frequency hence we revised our code to incorporate the right methods.

Google colab link for task 2:
https://colab.research.google.com/drive/1ElGXOBA5XdkNVyiVXHL6rWhLAglSeMqA?usp=sharing

Task 3

Date: 25/08/2023

Today, we went through the task 3 , we decided to use both data sets and perfromed the inital loading and cleaning of the data.

contribution:

Group member 1:

- Loading and extracting main google review data from task 1 and cleaning.

  Group member 2:

- Loading and cleaning meta data.

During this meeting we faced an issue with the meta data , our meta data had a lot of null columns and we had to come up with a strategy on how to handle these columns. We decided to set a threshold to select only a set of columns which had at least more than 50 % of data present and drop the remaining columns, there by we could focus on the exploration without any bias.We also explored a few potential insights we could use.

Date: 28/08/2023

Today,we began intial exploration of the data to identify trends and discuss what our main insights should focus on.

contribution:

Group member 1:

- Initial exploration of the google review data set and identify insight.

  Group member 2:

- Inital exploration of the meta data set and identify insights.

During this task we were each able to identify multiple insights covering various topics , hence we had to decide our goal in order to select the most important insights , we wanted our EDA to not only provide statitical descriptions but also provide summary that could be beneficial and easy to understand for a business owner as well hence we decided that we wanted to include semantic analysis to our insights along with location of a business and other statistical visualization. We used a combined data set for 2 visualizations to avoid inconsistencies and allow for proper data exploration.

Date: 30/08/2023

Today,perfromed the documentation and final review of all insights and recording of a video to demonstarte our findings.

contribution:

Group member 1:

- Documentation and review of the google review data insights and extraction.

  Group member 2:

- Documentation and review of the meta review data insights and extraction.

During our final meeting we realised that some of our insights had very basic infomation about the data set and would not contribute to a business or our goal of the EDA as a deep insight, but these explorations and visualizations were still useful hence we decided to classify them under initial exploration so that it could be used to provide context to our final insights and make it easier for an external viewer to understand our final insights.

Google colab link for task 3:
https://colab.research.google.com/drive/1Xdkd3RrZHs9t0ubnxBF84BahpI3VY4yq?usp=sharing

```
Start coding or generate with AI.
```