CS1951 Final Project Proposal

Project Type : Option 1

Team Members : Hongjun Choi (hc121), Jaisung Hui (jh138), Taegone Lee

Part 1 : Search Engine

- **Hypothesis** : Create a search engine using inverted index that includes the functionality ranking and querying.
- **Data** : yelp_academic_dataset_review.json
    - The data needs to be cleaned. First all the review texts will be set to lowercase. Then the texts will be tokenized by space(‘ ‘) deliminator. Among the words that all review texts includes, generic words such as “is”, “are”, “of” , and expressions such as “ ~”, “(“, “)”, and “-” will be deleted.
- **Methodology** :
    - Index creation : Our first job in implementation is to go through the entire file line by line, tokenizing string into clean, consistent form. During the process, the position of the word within that string will also be icluded in the dictionary. It will increase the size of the output considerably, but will greatly enhance the performance of PQ querying.
    - Querying :
        - QWQ: Go through the dictionary and fetch business ids containing that word.
        - FTQ: First parse the input string into a list. Then, lists of business ids containing each word in input list will be casted into into set. The intersect of those sets will be returned.
        - PQ: Parse the command by removing quotation marks and tokenizing each word. The process is almost similar to that of FTQ, but the code will make use of additional data about the position of the word within the reiew text to drop irrelevant data from set.
        - LSQ: Use of “IN” word in input will trigger this functionality to take additional ‘business’ data, dropping results that are not from the specified region.
    - Ranking: Calculate the td-idf as defined and use review rating in conjunction with review voting(funny, cool etc.) to determing the significance of certain reviews to evaluate final score for a business.
- **Task List**
    - Time Plan : Finish by 15 APR 15
    - Hourly Estimate : 20 hours
    - Deliverable Stages: All functionalities has to be met to receive complete grade (A)

Part 2 : Yelp Dataset

- **Question to Solve** : Can you take all of the reviews of a business and predict when it will be the most busy, or when the business is open?
- **Hypothesis 1:** The time period when there is highest number/ frequency of reviews corresponds to the busiest time of the day for the business.
- **Hypothesis 2:** The opening and closing time interval for the business is within 80% interval of the posting time of reviews.
    - If 80% of the posted reviews are posted between timeline x and y, the opening and closing times for the business would be in that time interval with 95% confidence.
- **Data** : Yelp dataset (yelp_academic_dataset_review.json)
- **Methodology**
    - Parse through yelp_academic_datset_review.json file and retrieve a list of review posting hours
    - Parse through review texts and look for keywords such as “open” , “close” and retrieve possible open and closing time of a business.
    - Using 1 by 24 size matrix, calculate the distribution of review posting in a day (24 hour range).
    - Using the distribution of review posting, conduct p_test & t-test regarding hypothetical open/closing time and actual business hour.
    - Visualize the data with distribution chart to show which time interval includes 80 % of review posting in Yelp. Also product a heat map showing frequencies of review postings on hourly interval during the week. (Darker color will represent higher review posint frequency)
- **Task List**
    - Time Plan : Finish by 20 APR 15
    - Hourly Estimate : 20 hours
    - Deliverable Stages
        - B : Validate / Invalidate the chosen hypothesis using at least 1 statistical inference
        - A : Validate/ Invalidate hypothesis by statistical inference including t-test and permutation test. Also with given results, conclude a probable open/closing hours for a business and measure the error percentage by comparing the results with actual yelp business data.
        - Extra Credit : Complete the tasks above and show it with creative visualization methods includind 3D hovering and heat maps that show the frequency of reviews in hourly period during the week. (The colors in the heap map will signify the frequency of Yelp reviews posted during the time interval.

Part 3 : Integrating Yelp & Zagat Dataset

- **Problem :** Quantitatively measure which among Zagat or Yelp is a better source for predicting food qualtiy and customer service.
- **Hypothesis 1:** The average star rating of Yelp is equivalent to average rating of Zagat
- **Hypothesis 2:** Yelp has more than 5% outlying reviews that deviate heavily from the average review and rating and are therefore less credible than Zagat.
- **Data** : Yelp dataset and Zagat Dataset
- **Methodology**
    - If Zagat dataset is not given, parse through "https://www.zagat.com/new-york-city" using "Beautiful Soup" and retrieve dataset of business and reviews for restaurants.
    - Compare the Yelp review ratings and Zagat review ratings for same restaurant, and create a list of difference in rating between the 2 sources.
    - Using permutation test under assumption that Zagat data has similar distribution to that of Yelp, accept or reject hypothesis 1 under 95 percent confidence interval in two-tail test.
    - Count the number of reviews in Yelp with sentiment score/ rating that deviates more than 2.0 $\times$ Standard Deviation
    - Calculate the percentage of deviating reviews in Yelp dataset and if frequency of outling reviews are higher thatn 5%, Yelp reviews are less credible than Zagat's
    - Visualize Yelp's rating distribution using Bar graph with x axis being rating scale and y axis being frequency of rating with corresponding rating range. Also, visualize the distribution result of permuation test using dot graph, x axis being permutation test number and y axis being the actual difference of the test.
- **Task List**
    - Time Plan : Finish by 28 APR 15
    - Hourly Estimate : 25 hours
    - Deliverable Stages
        - B: Validate hypothesis using Yelp dataset and parsed Zagat dataset and conclude which of Yelp / Zagat is more credible source as restaurant reference.
        - A : Validate the two hypothesis and visualize the resulting data with distrubution chart, showing number of outling reviews.
        - Extra credit : Conclude which of Zagat and Yelp is more credible source, and demonstrate creative visualization. Also, exclude the Yelp reviews that deviate from average more than 2 $\times$ standard deviation, and compare the new calculated average Yelp rating and Zagat rating.