Term project Progress Report #1

Tasks I've accomplished so far:

- Firstly, I analyzed the Yelp! Reviews dataset which contains approximately 160000 entries. The file is in JSON format and there are 7 attributes to this dataset, in which there are 4 continuous attributes and 3 categorical attributes. I also analyzed other datasets of Yelp! Including the business, user and tips.

- During the process, I found another idea which I find more interesting and relevant than my current project goal.

- I would like to restate my term project goal as follows:

  The main purpose of the project will be to extract the most popular and least popular dishes from the text reviews for a particular restaurant. Maybe, generate a list which tells the top 5 dishes to try in a particular restaurant/café/business. Also, the dishes which are not so good can be put into another list as the least popular dishes. The process can then be further extended to other businesses like theme parks (top 5 rides) and so on.
  These lists will help customers look for the best dishes to eat as well as give a chance to restaurant owners to tweak and change their least favorite dishes on the list.

- First, I took a chunk of data from the dataset to perform the preprocessing tasks:
  - As the dataset contains businesses other than restaurant/café, my first task was to remove other businesses.
  - Remove the reviews of other businesses also.

- Now, I have to use <business-id> attribute present in the reviews dataset to group the reviews for the same business/restaurant/café together.

The tasks which I plan to tackle next:

- Make a set of words containing these frequently occurring words: {try, awesome, delicious, tasty, mouthwatering, appetizing} and so on. Keep adding words to these sets.
- Extract the sentences that contain these words, probably these sentences will contain a dish name.
- Try to extract the dishes name from the sentences.
- Generate a list of these dishes according to the most frequent dish.