Amitesh Mahajan

CS 6775

HW2

Discussed only general strategies to solve problems for task 1 with Astha Tiwari.

Task 1:

Task 1: Data Download and Descriptive Analysis In this task, extract business IDs from yelp_academic_dataset_business.json if a business's city/location is either "Pittsburgh" or "Charlotte". By using each business ID as a unique key, extract corresponding checkins and reviews from yelp_academic_dataset_checkin.json and yelp_academic_dataset_review.json. Then, answer the following questions:

• Which types of restaurants are the most popular in each city? Define and describe what "popularity" means.

Answer: Popular restaurant: A popular restaurant is defined as the one having most check-ins.

Steps I followed to find popular restaurant categories in each city:

1. I Used Python script **(ScriptTo_GetBusinessIds.py)** to extract business_id, name, city, stars, review_count, categories, latitude, longitude about each entry in the yelp_academic_dataset_business.json file and saved these details to a separate file named as **businessIds.csv** .I used Pandas library for this task.
2. After fetching the above stated fields from yelp_academic_dataset_business.json file, I used another python script**(ScriptTo_Find_popularcatagories.py)** which uses the business_id fetched from step 1 to find the number of checkins for that restaurant category.This script generates a csv file containing checkins per category.
3. Used the excel to sort the results in descending order.

Top popular restaurants in *Charlotte:* *(I have removed the restaurant category from the list to make more sense as all of these can also be classified as a restaurant)*

Category	Checkins
Food	374
American (Traditional)	334
Nightlife	331
Bars	318
Sandwiches	316
Fast Food	304
American (New)	284
Burgers	212
Pizza	209

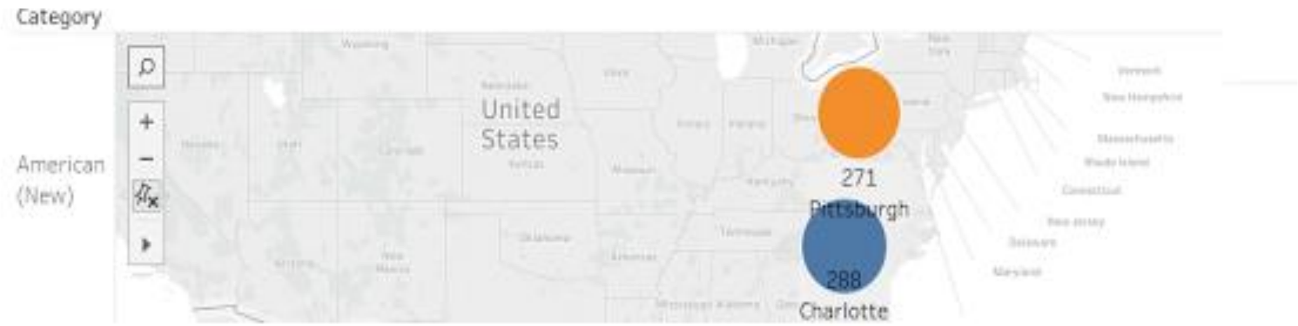
Top popular restaurants in *Pittsburgh:* *(I have removed the restaurant category from the list to make more sense as all of these can also be classified as a restaurant)*

Category	Checkins
Nightlife	349
Bars	344
Food	333
Pizza	312
American (Traditional)	297
American (New)	266
Sandwiches	261
Italian	189
Fast Food	138
Burgers	135

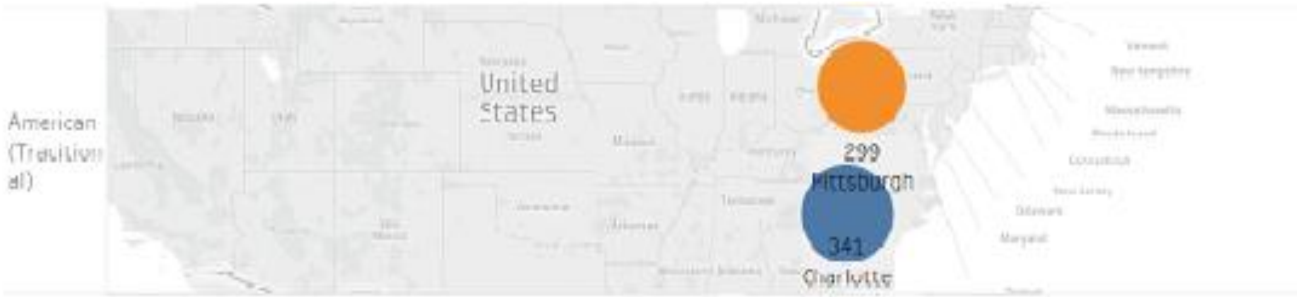
- In order to understand the popularlity of each city, visualize distribution on the map using tools like Tableau, Google maps API, basemap, D3, etc. Report your findings including some figures like snapshots of the maps.

Top 5 Category wise popularity comparison of the two cities:

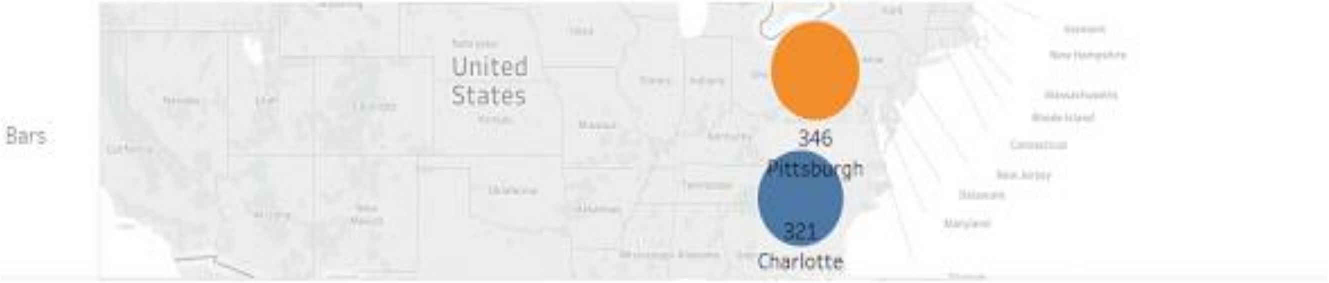
American(New):



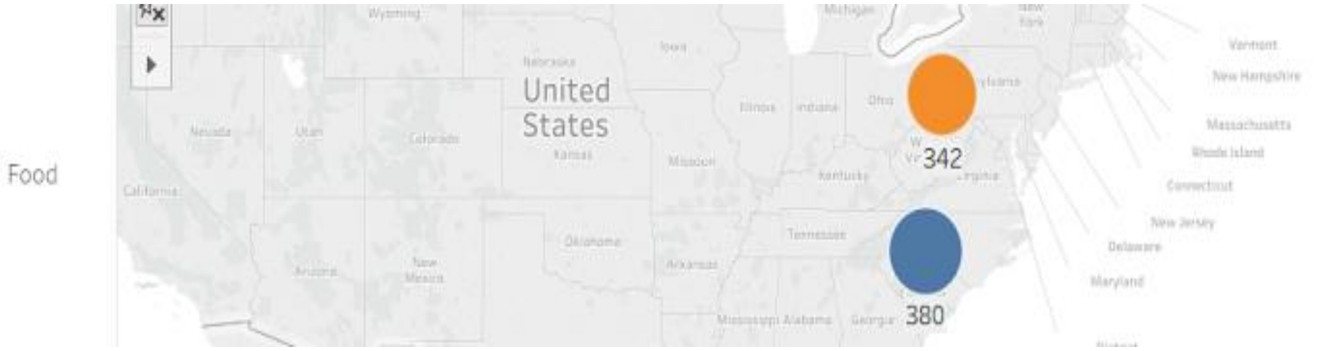
American traditional:



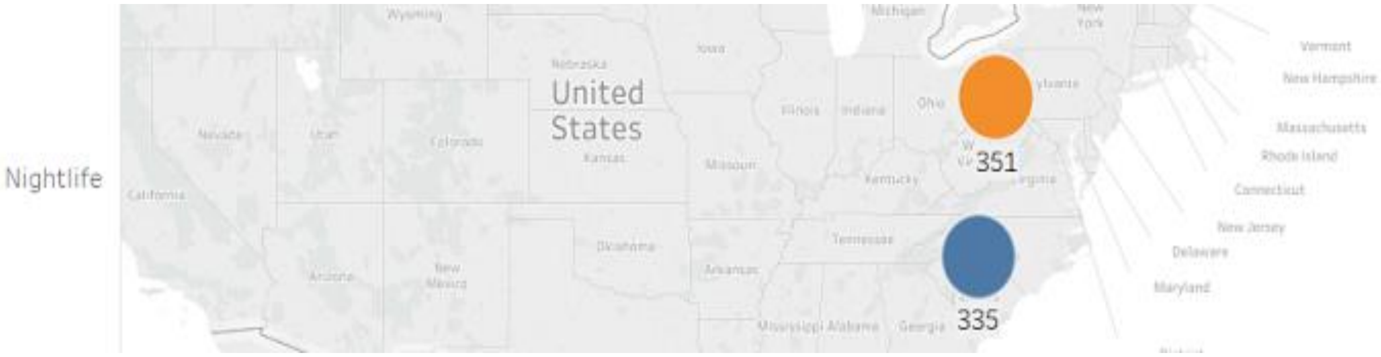
Bars



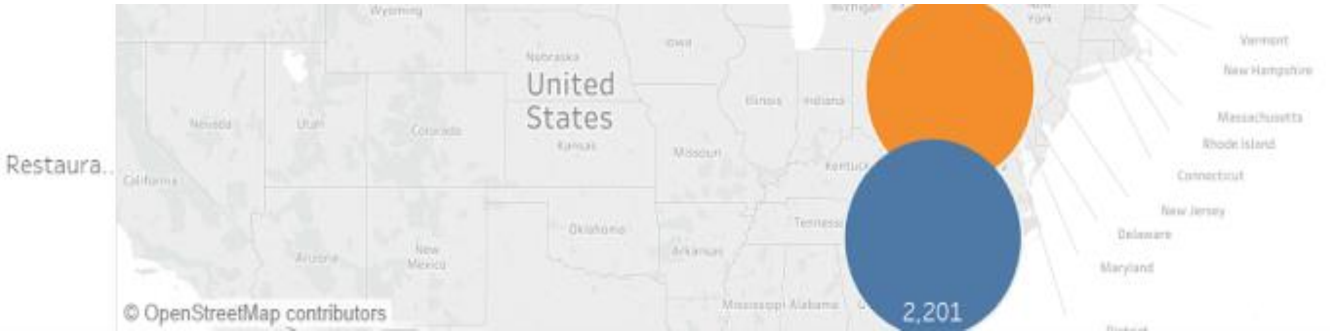
Food



Nightlife



Restaurants



Q3. Report one more interesting finding through descriptive analysis:

- As illustrated in the above maps, charlotte has more popularity as compared to Pittsburgh restaurants.
- Population of charlotte is 792,862 and that of pittsburg is 305,841, from this data we can derive an obvious reason for above observation.
- Popularity of bars and nightlife in Pittsburg is slightly higher than charlotte.

Task 2: Done using Mallet.

Commands:

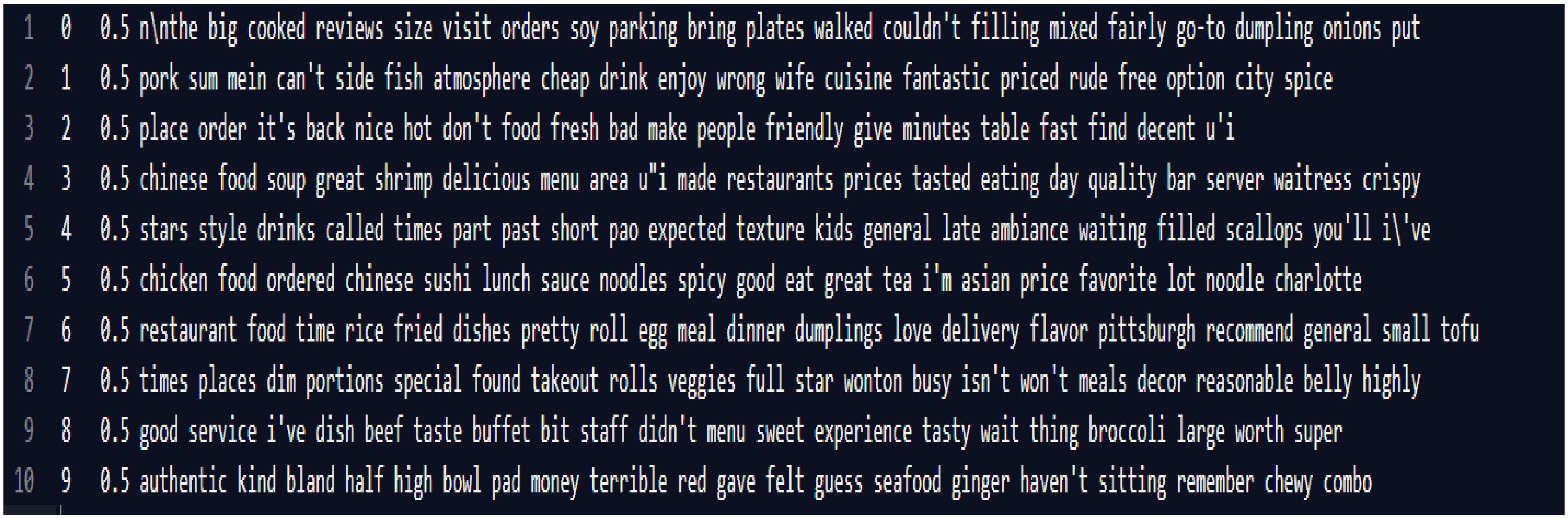
1.To convert input txt file into mallet file format:

```
bin\mallet import-dir --input C:\mallet\sample-data\datachinese --output output.mallet --keep-sequence --remove-stopwords
```

2. To create a topic model (train-topics), and save the output.

```
bin\mallet train-topics  --input output.mallet --num-topics 10 --output-state topic-state.gz --output-topic-keys tutorial_keys.txt --output-doc-topics tutorial_compostion.txt
```

Q1.Following are the most frequently used words to describe Chinese restaurants:



```
0   0.5 n\nthe big cooked reviews size visit orders soy parking bring plates walked couldn't filling mixed fairly go-to dumpling onions put
1   0.5 pork sum mein can't side fish atmosphere cheap drink enjoy wrong wife cuisine fantastic priced rude free option city spice
2   0.5 place order it's back nice hot don't food fresh bad make people friendly give minutes table fast find decent u'i
3   0.5 chinese food soup great shrimp delicious menu area u"i made restaurants prices tasted eating day quality bar server waitress crispy
4   0.5 stars style drinks called times part past short pao expected texture kids general late ambiance waiting filled scallops you'll i\'ve
5   0.5 chicken food ordered chinese sushi lunch sauce noodles spicy good eat great tea i'm asian price favorite lot noodle charlotte
6   0.5 restaurant food time rice fried dishes pretty roll egg meal dinner dumplings love delivery flavor pittsburgh recommend general small tofu
7   0.5 times places dim portions special found takeout rolls veggies full star wonton busy isn't won't meals decor reasonable belly highly
8   0.5 good service i've dish beef taste buffet bit staff didn't menu sweet experience tasty wait thing broccoli large worth super
9   0.5 authentic kind bland half high bowl pad money terrible red gave felt guess seafood ginger haven't sitting remember chewy combo
```

2.major themes/topics in the reviews of Chinese restaurants:

1. **food**
2. **noodles**
3. **good**
4. **chinese**
5. **general**