Yelp Friend Finder

By: Angus Leung

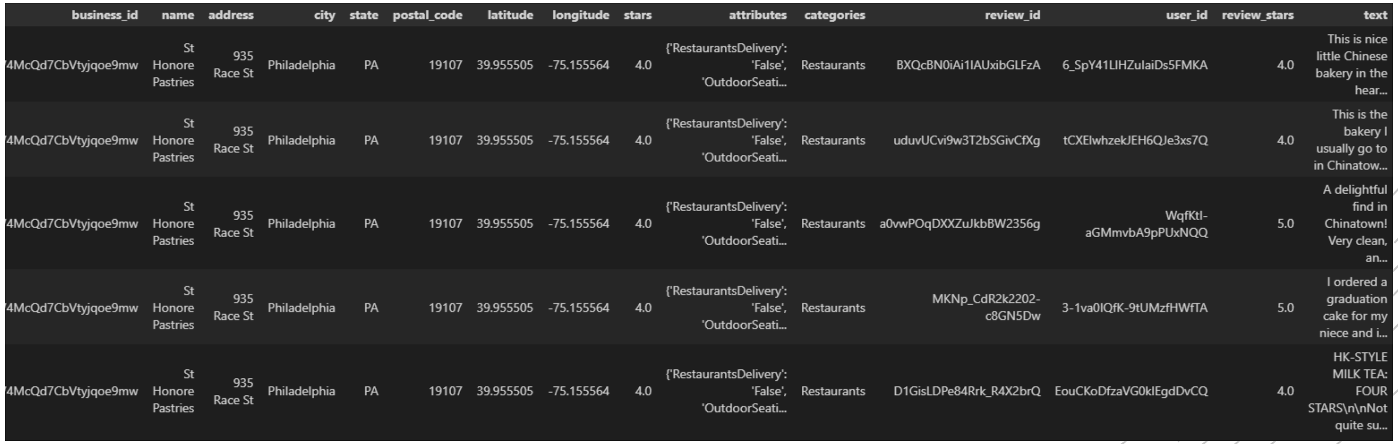
Client

- Continued work for the Yelp web application
  - Finding like minded individuals as well as restaurants
- Connect individuals with similar reviews

Data

- Yelp Open Dataset
  - Just under 7 million reviews
  - Just under 2 million users (tokenized)

- Final DataFrame after cleaning and merging
  - Kept Restaurants in Philadelphia
  - 168354 reviews



business_id	name	address	city	state	postal_code	latitude	longitude	stars	attributes	categories	review_id	user_id	review_stars	text
4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants	BXQcBN0iAi1IAUxibGLFzA	6_SpY41LIHZulaiDs5FMKA	4.0	This is nice little Chinese bakery in the hear...
4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants	uduvUCvi9w3T2bSGivCfXg	tCXElwhzekJEH6QJe3xs7Q	4.0	This is the bakery I usually go to in Chinatow...
4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants	a0vwPOqDXXZuJkbBW2356g	WqfKtI-aGMmvbA9pPUxNQQ	5.0	A delightful find in Chinatown! Very clean, an...
4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants	MKNp_CdR2k2202-c8GN5Dw	3-1va0IQfK-9tUMzfHWfTA	5.0	I ordered a graduation cake for my niece and i...
4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants	D1GisLDPe84Rrk_R4X2brQ	EouCKoDfzaVG0kIEgdDvCQ	4.0	HK-STYLE MILK TEA: FOUR STARS\n\nNot quite su...

Tools

- Pandas and Numpy for data manipulation
- Scikit-learn for mathematical analysis
  - NMF, TFIDF, CountVectorizer, DBSCAN
- NLTK for language processing
- Ipywidgets for topic exploration

Vectorizing

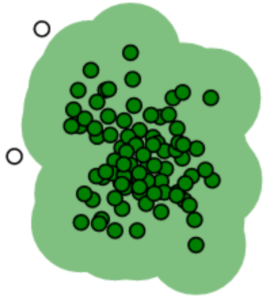
- CountVectorizer
- TfidfVectorizer
- HashVectorizer

- Ultimately picked results found with Count Vectorizer to capture any words TFIDF might normalize

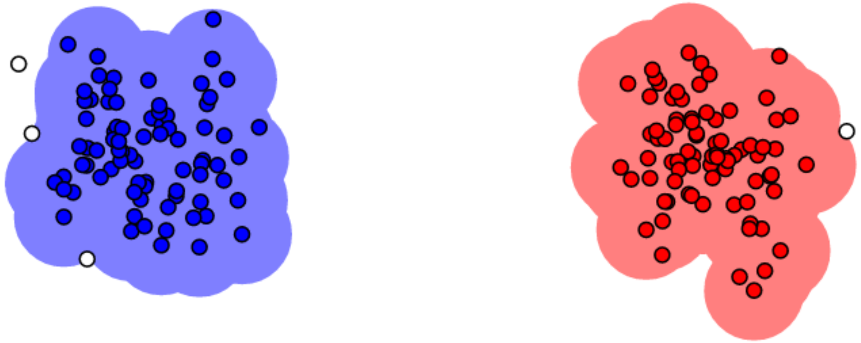
Dimensionality Reduction/Topic Modelling

- Non-Negative Matrix Factorization(NMF)
  - Landed on 3 topics
    - Negative reviews about the restaurant
    - Neutral reviews (some good things and some bad things)
    - Positive reviews about the restaurant

Clustering

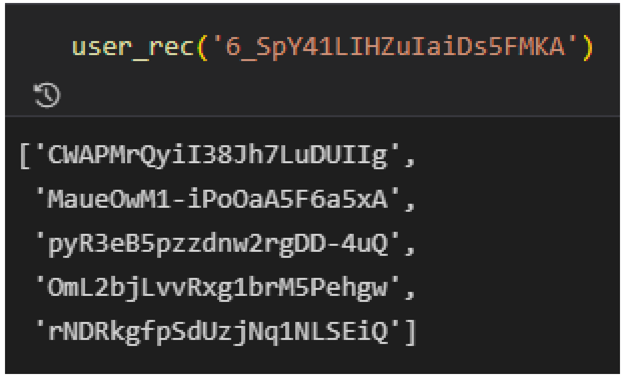


- DBSCAN was attempted but dataset was too large
- BIRCH algorithms to be used for future work to first reduce data size



Recommender

- Dot product between NMF Matrix vectors used to find similarity between reviews
- Top 5 users with the most similar reviews returned
- A Collaborative Recommender



```
user_rec('6_SpY41LIHZuIaiDs5FMKA')
```

```
['CWAPMrQyiI38Jh7LuDUIIg',
 'MaueOwM1-iPoOaA5F6a5xA',
 'pyR3eB5pzzdnw2rgDD-4uQ',
 'OmL2bjLvvRxg1brM5Pehgw',
 'rNDRkgfpSdUzjNq1NLSEiQ']
```

Future Work

- Expand dataset to include more locations
- Additional time on clustering to better visualize data
- Additional work to optimize run time of recommender
- A web-app to better visualize recommender