

# Coursera\_Capstone

## Introduction

Project idea is to compare data about venues from different social networks. There is rating of venues at Foursquare, which is average mark from visitors of this venue, also we may take posts from Instagram and perform sentiment analysis of posts about featured venues. Out of this we will have actual values to compare and see if there is correlation between those social platforms and if people at specific social platform are more "kind" to venues for some reason. Finally we can draw a map where we show not only Foursquare data, but sentiment analysis estimation to give more information to user if place is really good.

## Install

```
git clone git@github.com:old-entertainment/Coursera_Capstone.git
cd Coursera_Capstone
virtualenv -p /usr/bin/python3.7 cou_env
source ./cou_env/bin/activate
pip install -r requirements.txt
```

## Data

Data was gathered from three sources: Foursquare, Twitter, Reddit.

- Fousquare
  - dataframe with venues locations and ratings
  - every venue is one row
  - columns: venue, latitude, longitude, rating, instagram\_sentiment
- Instagram
  - dataframe with posts for every venue
  - every venue is one post
  - columns: Date, compound, neg, neu, pos, query, text
- Result
  - dataframe with all information merged together
  - every venue is one row
  - columns: id, name, categories, lat, long, num\_checkins, num\_likes, price, rating, num\_ratings, url\_venue, url\_foursquare, query, inst\_snt

Detailed columns description

- id: (str) - id of venue from FourSquare
- name: (str) - name of place in FourSquare
- text: (str) - post text from Twitter or Instagram about specific venue
- lat: (float) - the angular distance of a place north or south of the earth's equator
- long: (float) - the angular distance of a place east or west of the Greenwich meridian
- rating: (float) - FourSquares average mark for venue
- inst\_snt: (float) - value for an attitude towards venue


## Methodology


Data is taken from two sources: Foursquare API and Instagram API. Functions are implemented in files `/src/lib/foursquare.py` and `/src/lib/instagram.py`. Data was examied, so every venue would have sentiment value and rating value. Originally 1K venues were taken for analysis and 62 were left for futher investigation. Every posts about venue was evaluated via sentiment analysis using [nltk](#) library, then we calculated average sentiment for every venue. Dataset merging and aggregation was performed to fit venue name, latitude, longitude, rating, instagram sentiment in one [result dataset](#). Then we calculated multiple correlations of sentiments and ratings to avoid [Anscombe's quartet](#) confusion and performed visualization to check by eye if correlation values make sense. Then we plotted [folium](#) map of San Francisco with venues, ratings and sentiments.

# Results

## Correlation of Instagram sentiment vs Foursquare rating

- Pearson: 0.03
- Kendall: 0.59
- Spearman: 0.54

San Francisco interactive map with Instagram sentiment and Foursquare ratings for every venue embedded in popup labels. Can be plotted with corresponding cell from final\_assignment.py. 

Graph with Instagram sentiment and Foursquare ratings for every venue. Plotted in order to figure out visually if there is a dependency between ratings and sentiment. 

## Discussion

- Spikes on sentiments graph can be explained via low number of posts for specific venues
- Correlation is ambiguous because Pearson shows results close to zero. But if we adjust rating values and to be positive and negative like sentiments(minus average rating(5.0)), then we'll see that they sort of correspond to each other - somewhat more than zero.
- Fact that there is no single negative sentiment value suggest that Instagram is not the platform, where people prefer to complain(perhaps Twitter is) or San Francisco has only good venues(not very likely).
- At the center of San Francisco ratings and sentiments of venues are on average higher than in other areas, which is perfectly understandable.

## Conclusion

- Average sentiments for every venue were calculated.
- Correlation of Foursquare ratings vs Instagram sentiments was examined.
- Interactive map with ratings and sentiments values was plotted.

## License

This project is licensed under the MIT License - see the [LICENSE](#) file for details.