CS 226 - Final Project Report

What America Likes To Eat

Varun Sapre
Dept. of CSE
University of California, Riverside
vsapr002@ucr.edu

Karthik Harpanahalli
Dept. of CSE
University of California, Riverside
kharp009@ucr.edu

Sarthak Jain
Dept. of CSE
University of California, Riverside
sjain050@ucr.edu

Abstract

This project aims to visualize popular food across major metropolitan areas in the US. It aims to help travelers, visitors, immigrants and even locals view the local interest in food and its popularity. Internet companies such as Yelp, Foursquare are pioneers in aggregating information about businesses and restaurants in the US. They collected extremely minute details on every business such as the category of the business, hours of operation, foot traffic at a particular business, etc. Apart from this, they also aggregate reviews of the businesses submitted by real users. These reviews are a good source of information to understand if the restaurant is being enjoyed by its customers, or not. Apart from just this, with the rise of social media, we can easily get data from websites such as twitter to check how many people are talking about these restaurants and what they are talking about.

Section 2 introduces the problem and challenges. Section 3 gives our analytical framework including four parts: data processing, natural language processing, ranking algorithms, and web application. Section 4 shows the evaluation of our project. Section 5 provides the future scope of the project.

1. Problem Statement and Challenges

With vast amounts of data available to food-based websites and free to use social media platforms where people voice their unadulterated opinions, we decide to combine these datasets to provide a deeper insight into the reviews of restaurants across the United States Of America.

With the help of data from Yelp and Twitter, we use it to develop a ranking algorithm to identify the popular food categories in an area. Key challenges include getting good quality datasets, preprocessing different structured and semi-structured datasets, running NLP algorithms on 8.6+ million reviews and 4.6 million tweets, deploying the analyzed data on a web site.

2. Literature Survey

The literature survey was undertaken to provide an extensive list of papers published that are relevant to our work and summarize the findings in these papers. The papers have been

chosen to study our project by components viz., data collection from FourSquare, Yelp and Twitter and combining them, ranking locations, performing Sentiment Analysis etc, and each of the following papers provides insights to one or more of the above mentioned components of our project.

The methodology to procure data from foursquare and then combine them using the fusion method is highly relevant to our project. The steps followed by authors provides a direction with which we can approach the data collection and data combination problem. Fusion method to combine data, tensor factorization methods to classify and SEALs framework are key insights on how to solve the core problem of our project. The performance MT-RTF, PopularK, Relevance+PrefU, HOSVD and PITF provides an insight to what is the best use case for each. We can draw several inferences from the paper that guide our project to extract a highly meaningful analysis. [1]

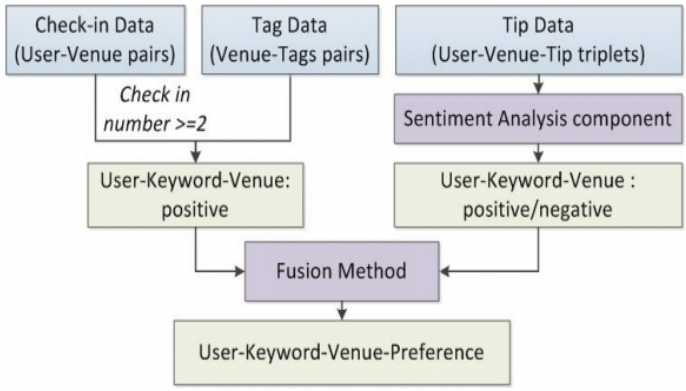

Fig 1: User-Keyword-Venue Modeling from source [1]

The methodology in the source [2] procurement of the dataset from Yelp, a platform for crowd-sourced reviews about businesses. The paper divulges on how it pre-processes and filters the data specifically for restaurants. It extracts the features to access insights on the preprocessed data. Information Gain Ratio provides a weight for different features that are taken into consideration in the prediction models.

The predicted values are then compared against the actual ones and the performance percentage for each of the models are calculated.

3. Analytical Framework

I. Data Collection

Yelp Dataset

We downloaded the Yelp Dataset which contains 8.63 million reviews of 160k businesses spread over the top 8 metropolitan areas in the USA. We use that dataset to filter reviews from only restaurants.

Twitter Mentions

To further expand our review dataset we include relevant tweets for the selected restaurant. We use a powerful web scraping tool built for twitter called twint and we use the location data of the restaurants from the Yelp dataset to fetch tweets from around that location pertaining to that restaurant.

II. Data Storage

We use HDFS (Hadoop Distributed File System) to store our data. We set up instances of HDFS in our local machines, and used the hadoop cli to send data into the hdfs master node. Since we were running this on our laptops and we were working with large files, we set the replication factor to 1.

III. Data Integration

The yelp dataset is split into 3 files - business.json (125mb, 160k rows), review.json (6.94GB, 8.63M rows), check in.json (400mb, 132k rows). These three files had to be combined using the business_id column that is common to them all. The same business_id column was used to join the scraped data from twitter. We used the name of the restaurant to scrape data from twitter and joined it along with the business_id from yelp. This way, we could join the reviews dataset with the twitter mentions as well.

IV. Natural Language Processing

We used the “Flair” pre-trained NLP library to run sentiment analysis on the reviews from yelp and the twitter mentions. The output of the sentiment analysis was a score of the accuracy of the prediction and the “positive”/”negative”/”neutral” sentiment. We aggregated the sentiments for the reviews and have a single sentiment assigned to each restaurant. This sentiment is used as part of the ranking algorithm.

V. Ranking Algorithm

For the ranking algorithm, we looked for several models that can handle data that are in varying ranges. The following are the parameters used to generate the score for each restaurant.

- Ratings : Ratings left by patrons for a business location normalized over a single scale.
- Foot Traffic : The number of visits at a business location collected over a time period.
- Number of reviews : The number of reviews left for a business location
- Sentiment analysis
  - Twitter mentions
  - Yelp reviews

For this, we studied feature scaling and three ways of normalization were considered. The first one is sum normalization, where the ratio of column value is taken to the sum of values. Second one is max normalization, where the ratio is between value and the maximum value of the column data, and for the third one rescaling (min-max normalization). The normalized values were computed for each and the restaurants were ranked using each normalization to understand the different output.

$$\underline{\text{Maximization}}$$

$$\text{Let A be an attribute which we want to maximize,}$$
$$\text{and its elements are: } [\,a_1 \quad a_2 \quad \dots \quad a_n\,];\ 1 \le i \le n$$
$$\text{Then, } maximize(a_i) = maximizeFunc(a_i, A)$$
$$\text{where, } maximizeFunc(a_i, A) = \begin{cases} \frac{a_i}{sum(A)}, & \text{sum normalization} \\ \frac{a_i}{max(A)}, & \text{max normalization} \\ \frac{a_i - min(A)}{max(A) - min(A)}, & \text{max-min scaling} \\ \dots \end{cases}$$

$$\underline{\text{Minimization}}$$

$$\text{Let A be an attribute which we want to minimize,}$$
$$\text{and its elements are: } [\,a_1 \quad a_2 \quad \dots \quad a_n\,];\ 1 \le i \le n$$
$$\text{Then, } minimize(a_i) = minimizeFunc(a_i, A)$$
$$\text{where, } minimizeFunc(a_i, A) = \begin{cases} \frac{1}{maximizeFunc(a_i, A)}, & \text{inverse} \\ 1 - maximizeFunc(a_i, A), & \text{subtract} \\ \dots \end{cases}$$

Fig 2: Image credits: towardsdatascience.com

Sentiment analysis was performed for the reviews posted for the restaurant and the tweets collected for the particular restaurant within 100 kilometer of radius. A score on the range of -1 to 1 was assigned. -1 being negative, 1 being positive. The weights were assigned to these three columns and the product of these were taken to provide the score. Once the scores were computed, the places were sorted by the ranks. Top restaurants were chosen from 8 cities to be displayed on the map.

```
df_rankingDataFrame.show(10)
```

```
+--------------------+--------------------+------------+--------------+-----+--------------------+-----------+-----+-----------+-------+-----+------------+
|         business_id|                name|    latitude|     longitude|stars|          categories|       city|state|postal_code|checkin|senti|review_count|
+--------------------+--------------------+------------+--------------+-----+--------------------+-----------+-----+-----------+-------+-----+------------+
|6iYb2HFDywm3zjuRg...| Oskar Blues Taproom|   40.0175444|  -105.2833481|    4|Gastropubs, Food,...|    Boulder|   CO|      80302|    184|   -1|          86|
|tCbdrRPZA0oiIYSmH...|Flying Elephants ...|45.5889058992|-122.5933307507|    4|Salad, Soup, Sand...|   Portland|   OR|      97218|   1180|   -1|         126|
|jFYIsSb7r1QeESVUn...|     Boxwood Biscuit| 39.947006523|  -82.997471|  4.5|Breakfast & Brunc...|   Columbus|   OH|      43206|      7|    1|          11|
|HPA_qyMEddpAEtFof...| Mr G's Pizza & Subs|    42.541155|   -70.973438|    4|Food, Pizza, Rest...|    Peabody|   MA|      01960|     36|    1|          39|
|ufCxltuh56FP4-ZFZ...|      Sister Honey's|   28.5132647|  -81.3747072|  4.5|Restaurants, Amer...|    Orlando|   FL|      32806|    246|    1|         135|
|GfWJ19Js7wX9rwaHQ...|Everything POP Sh...|   28.3504984|   -81.542819|    3|Restaurants, Amer...|    Orlando|   FL|      32830|     63|   -1|           7|
|dmbbf3AqeG6l_QHRZ...|           RaceTrac|   28.4503025|  -81.3805873|  3.5|Automotive, Ameri...|Pine Castle|   FL|      32809|     26|    1|           5|
|ynTjh_FdhbG5hY69H...|  Cascade Restaurant|   28.3819454|   -81.510327|  3.5|Hotels, American ...|    Orlando|   FL|      32836|     78|    1|          18|
|hcRxdDg7DYryCxCoI...|   Longwood Galleria|    42.338544|   -71.106842|  2.5|Restaurants, Shop...|     Boston|   MA|      02215|    242|    1|          24|
|jGennaZUr2MsJyRhi...|     Legal Sea Foods|   42.3634422|  -71.0257812|  3.5|Sandwiches, Food,...|     Boston|   MA|      02128|   3262|    1|         856|
+--------------------+--------------------+------------+--------------+-----+--------------------+-----------+-----+-----------+-------+-----+------------+
only showing top 10 rows
```

```
df_rankingDataFrame.show(10)
```

```
+--------------------+-----------+-----+-----------+-------+-----+------------+-------------------------+-------------------------+-------------------------+
|          categories|       city|state|postal_code|checkin|senti|review_count|finalMaxNormalizationScore|finalSumNormalizationScore|finalMinMaxScalingScore|
+--------------------+-----------+-----+-----------+-------+-----+------------+-------------------------+-------------------------+-------------------------+
|Gastropubs, Food,...|    Boulder|   CO|      80302|    184|   -1|          86|       0.055823382798639534|      0.000666705054364...|       0.055823382798639534|
|Salad, Soup, Sand...|   Portland|   OR|      97218|   1180|   -1|         126|        0.35799778099127527|       0.00427560850081524|        0.35799778099127527|
|Breakfast & Brunc...|   Columbus|   OH|      43206|      7|    1|          11|       0.002123715649948...|      0.000025363779242...|       0.002123715649948...|
|Food, Pizza, Rest...|    Peabody|   MA|      01960|     36|    1|          39|       0.010921966199733821|      0.000130442293245...|       0.010921966199733821|
|Restaurants, Amer...|    Orlando|   FL|      32806|    246|    1|         135|        0.07463343569818111|       0.00089135567050894|        0.07463343569818111|
|Restaurants, Amer...|    Orlando|   FL|      32830|     63|   -1|           7|        0.01911344084953419|      0.000228274013179...|       0.01911344084953419|
|Automotive, Ameri...|Pine Castle|   FL|      32809|     26|    1|           5|       0.00788808669980776|      0.000094208322899...|       0.00788808669980776|
|Hotels, American ...|    Orlando|   FL|      32836|     78|    1|          18|       0.023664260099423278|      0.000282624968697...|       0.023664260099423278|
|Restaurants, Shop...|     Boston|   MA|      02215|    242|    1|          24|        0.07341988389821069|      0.000876862082370...|       0.07341988389821069|
|Sandwiches, Food,...|     Boston|   MA|      02128|   3262|    1|         856|         0.9896514928758813|       0.011819521126829928|        0.9896514928758813|
+--------------------+-----------+-----+-----------+-------+-----+------------+-------------------------+-------------------------+-------------------------+
```

Fig 3: Tables showing the final output

4. Evaluation

a. Ranking Algorithm

For data normalization, 3 different normalization techniques were used and each had a different set of output, which led to insights into how a particular parameter such as foot traffic had more weight than average reviews. As not everyone who visits a business leaves a review.

```
+-------------------------+-------------------------+-----------------------+
|finalMaxNormalizationScore|finalSumNormalizationScore|finalMinMaxScalingScore|
+-------------------------+-------------------------+-----------------------+
|      0.055823382798639534|      0.000666705054364...|   0.055823382798639534|
|       0.35799778099127527|       0.00427560850081524|    0.35799778099127527|
|      0.002123715649948...|      0.000025363779242...|   0.002123715649948...|
|      0.010921966199733821|      0.000130442293245...|   0.010921966199733821|
|       0.07463343569818111|       0.00089135567050894|    0.07463343569818111|
|       0.01911344084953419|      0.000228274013179...|   0.01911344084953419|
|       0.00788808669980776|      0.000094208322899...|   0.00788808669980776|
|      0.023664260099423278|      0.000282624968697...|   0.023664260099423278|
|       0.07341988389821069|      0.000876862082370...|   0.07341988389821069|
|        0.9896514928758813|       0.011819521126829928|    0.9896514928758813|
+-------------------------+-------------------------+-----------------------+
```

Fig 4: Tables showing the final output

b. Sentiment Analysis on reviews and tweets
    - Building Model from scratch
        - A custom model for sentiment analysis would required us to build a universal sentence encoder which is different from embedding based model
        - We did not go with this approach because there are better pre-trained models available that are made to work with reviews and social media
    - TextBlob
        - Textblob is a simple python library that offers processing on text data such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
        - We ran textblob on a sample of the review dataset and saw that the accuracy of textblob in predicting the sentiment was not consistent. It would often predict the wrong sentiment for the review.
        - It reports the polarity and subjectivity of the text
        - Textblob also ignores a lot of the words that it doesn't have data for
    - VADER (Valence Aware Dictionary for sentiment reasoning)
        - VADER is a rule-based sentiment analyzer
        - It requires a list of words that are individually marked as positive or negative words
        - It then uses the semantic orientation of those words to analyze the sentiment
        - The problem with VADER is that it only aggregates based on the words individually and ignores the context in which they are presented
    - Flair NLP
        - Flair is an embedding based sentiment analysis model
        - Text embedding makes sure that synonymically similar words are represented similarly
        - It uses this text representation to predict the sentiment
        - Embedding based models give a much higher accuracy in predicting the sentiment
        - Usage of flair is very simple as it provides pre-trained models for multiple languages which were trained on the IMDB dataset
        - We found flair to be the easiest and most accurate model to work with
c. Performance of Spark
    - Spark dataframe caused memory overflow issues for very large datasets that involved several record changes
d. Challenges faced:
    - Loading of data from HDFS was significantly slower than local filesystem
    - Processing large dataframes in spark such as the one for the review dataset causes spark standalone to crash multiple times due to out-of-memory issues and extreme bottlenecking.
    - Working with rate-limiters from twitter to download mentions slowed down the whole process

5. Conclusion
    a. Analysis

We analyzed big data for business check-in, reviews, ratings, location to get insights into US eating habits. We performed sentiment analysis on the reviews and tweets, ranked the restaurants by the above parameters and displayed them on the map.

    b. Web Application

We use ArcGIS to deploy our data and visualize it on the map. We develop a website that hosts the ArcGIS application that shows the locations of all the restaurants and its information.

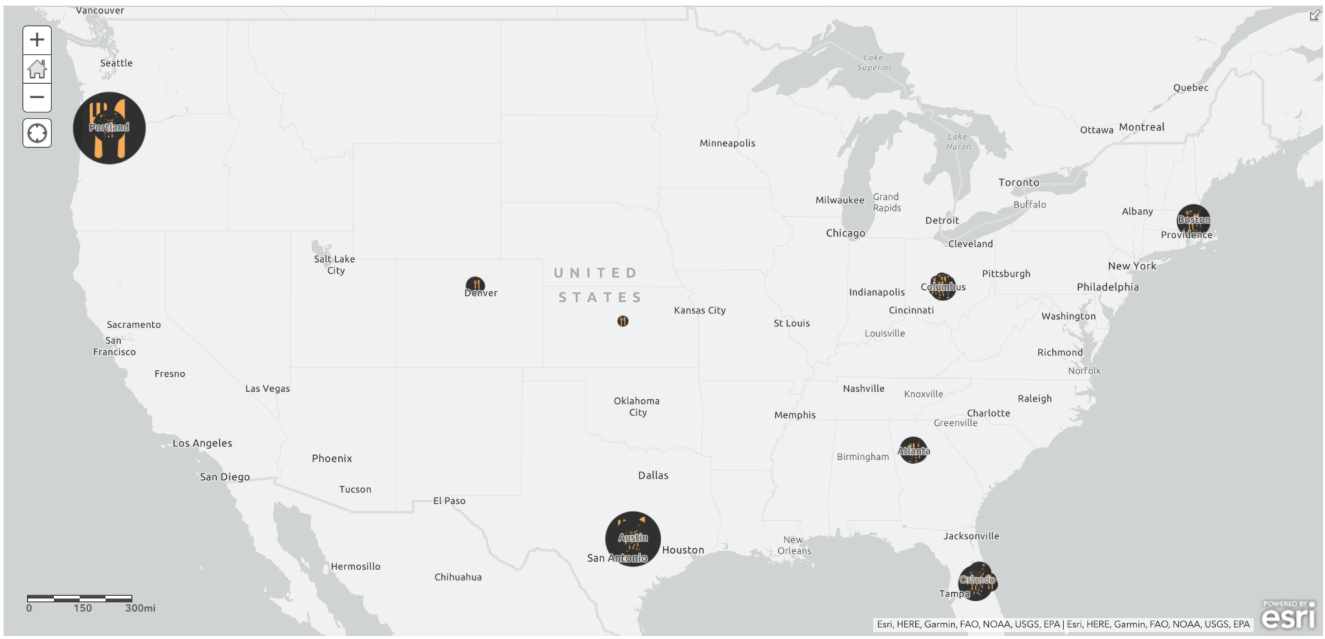

Fig 5: Visualization of restaurants by their ranking

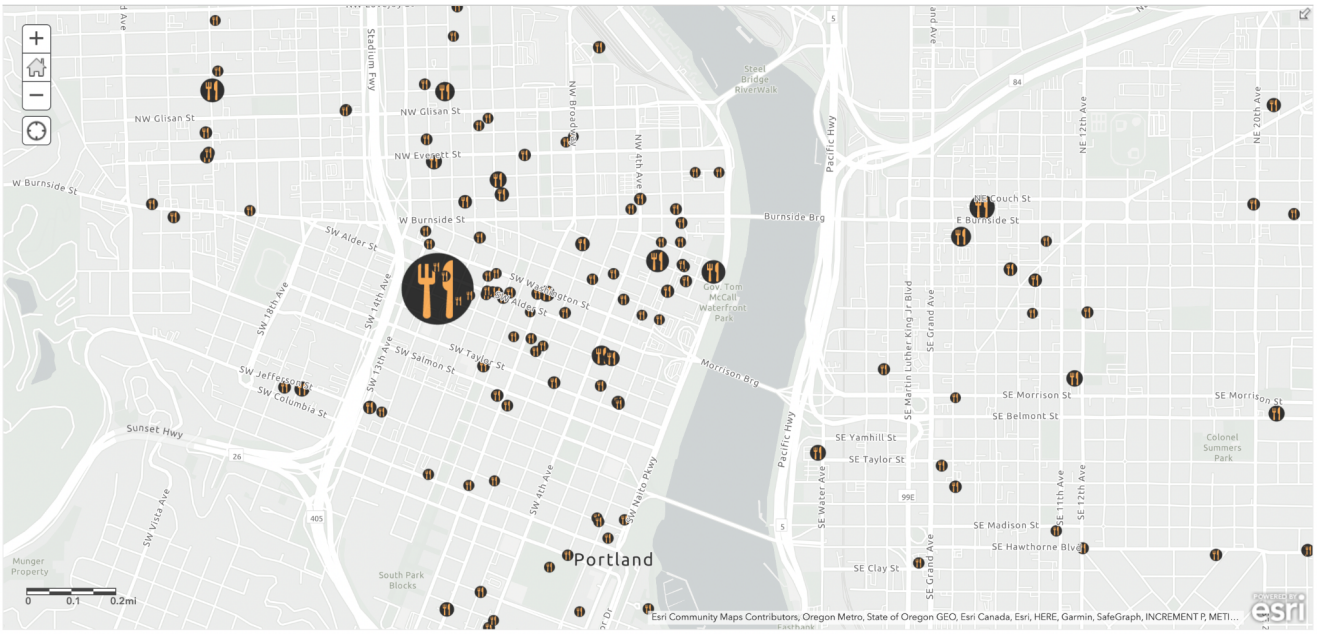

Fig 6: Visualization of restaurants in Portland, Oregon

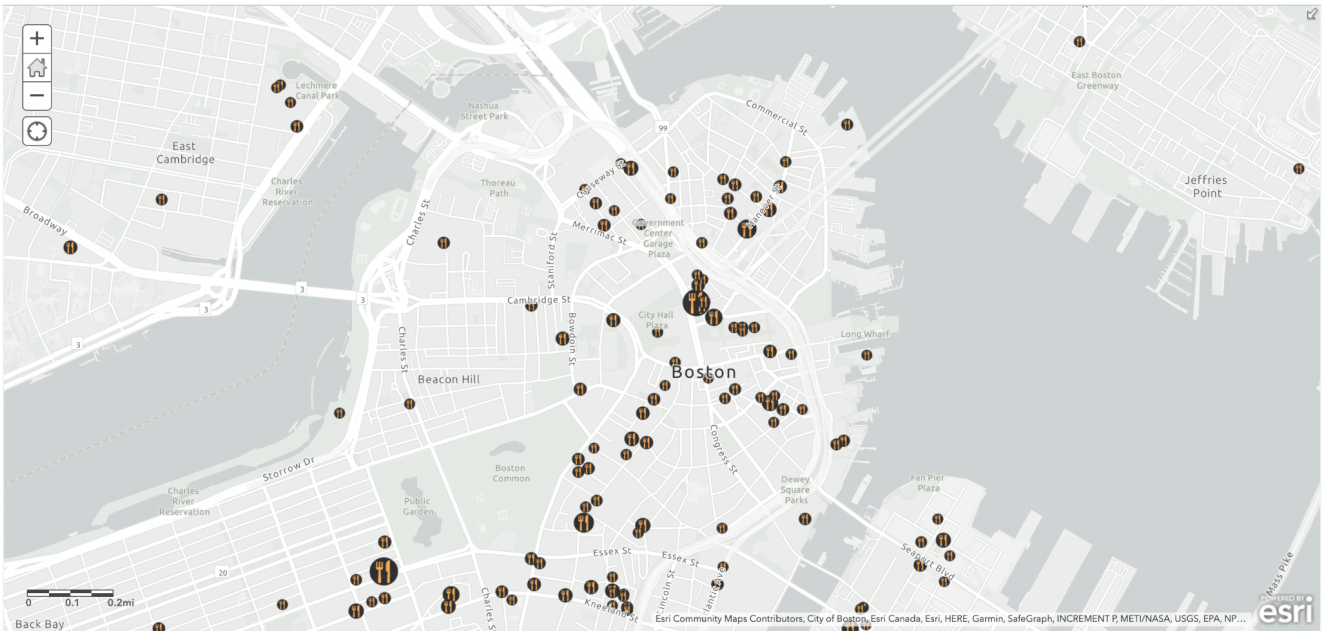

Fig 7: Visualization of restaurants in Boston, Massachusetts

6. Future Work

a. Setup a complete spark cluster with a driver node and worker nodes. Use this cluster to fully realize the effectiveness of using Spark.
b. Improve ranking algorithms, use machine learning to extract features for the algorithm and compare the performance.
c. Provide more levels of interaction in the web interface to view each locality's generalized food choice along with data for each business location.
d. Detecting fake reviews and discarding them

7. References

[1] Dingqi Yang, Daqing Zhang, Zhiyong Yu and Zhiwen Yu, Fine-Grained Preference-Aware Location Search Leveraging Crowdsourced Digital Footprints from LBSNs. Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), September 8-12, 2013, in Zurich, Switzerland. [PDF]

[2] Y. Chen and F. Xia, "Restaurants' Rating Prediction Using Yelp Dataset," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA), 2020 [PDF]

[3] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare., ICWSM 11 (2011) 70–573. [PDF]

[4] Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In Proc. of ACL'02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. Association for Computational Linguistics, 63–70.

[5] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. In DS'10, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol., 60(11):2169–2188, Nov. 2009.

[7] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. In Proc. of LREC'10. 2200–2204.

[8] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare., ICWSM 11 (2011) 70–573.

[9] Dingqi Yang, Daqing Zhang, Zhiyong Yu and Zhu Wang, A Sentiment-enhanced Personalized Location Recommendation System. Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT 2013), 1-3 May, 2013, Paris, France. [PDF]

[10] Dingqi Yang, Daqing Zhang, Zhiyong Yu, Zhiwen Yu, Djamal Zeghlache. SESAME: Mining User Digital Footprints for Fine-Grained Preference-Aware Social Media Search. ACM Trans. on Internet Technology, (TOIT), 14(4), 28, 2014. [PDF]

[11] Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch - Neptune.ai