Most Prospering Business Prediction Based on Big Data Analysis

Jinglin Wang

Department of Computer Science, Courant Institute of Mathematical Sciences, New York University

New York, United States

jlwang@nyu.edu

Yicong Tao

Department of Computer Science, Courant Institute of Mathematical Sciences, New York University

New York, United States

taoyc@nyu.edu

Xi Xiao

Department of Computer Science, Courant Institute of Mathematical Sciences, New York University

New York, United States

xx565@nyu.edu

***Abstract—In this project, we target at offering meaningful information and suggestions on the most prospering business in a given location. By combining data from the following sources: Yelp reviews, government sales census and Twitter tweets, we aim to analyze best business choice for companies to turn their effort to. Technologies such as Hadoop, Spark and HBase will be used to store and analyze large datasets. Analytical methods such as natural language processing, N-Gram analysis, and machine learning are applied to draw insightful conclusions.***

***Keywords—Big data analytics, Natural Language Processing***

# Introduction

Online platform reviews are becoming more influential than ever. Numerical ranking is an explicit way of deciding which store is better than another but it’s not comprehensive. According to Pavlou and Dimoka’s research[[1]](https://paperpile.com/c/RLNPu9/mPbj), text comments have more influence on a seller’s reputation than ranking and will further influence the price premiums. Thus reviews seem to be more important if we are to decide features that are crucial for a store to become competitive. However, it’s been estimated that 16% of Yelp reviews are fake[[2]](https://paperpile.com/c/RLNPu9/3z0c). And those fake reviews are written in an organized way. That is to say, the fraud practitioners have worked collaboratively and form coordinated campaigns[[3]](https://paperpile.com/c/RLNPu9/cSjV)[[4]](https://paperpile.com/c/RLNPu9/Z3IZ). Researchers have proposed both supervised and unsupervised models to either rank reviewers or identify fake reviews directly[[3]](https://paperpile.com/c/RLNPu9/cSjV)[[4]](https://paperpile.com/c/RLNPu9/Z3IZ)[[5]](https://paperpile.com/c/RLNPu9/7FRp). In this paper, we’ll use the one of the established models to identify the fake reviews and discard them for better feature extraction. (details to be completed…)

Social networks, which is a social structure consisting of a combination of social actors, emerge dramatically in recent years due to the popularization of smartphones. Twitter, for example, is one of the most important contributors in this trend. With a simple 140-character message called ‘tweet’, everyone can share their happiness and express their opinions with others instantly. Tremendous information is embedded in such ‘tweets’ and they serve as a huge database from which we may be able to answer almost any questions related to society. For example, using Twitter’s data to predict United States presidential election results[[6]](https://paperpile.com/c/RLNPu9/b08x) becomes a research hotspot in recent years, and its precision usually outperforms traditional media polls.

Though videos, photos, and even audios are becoming more and more popular on social media[[7]](https://paperpile.com/c/RLNPu9/ofJy), text is still the main source of public opinions and user status. Therefore, natural language processing seems to be a plausible choice in analyzing social media data. In this paper, we aim to apply opinion mining and sentiment analysis on ‘tweets’ in order to gain insights about the most popular businesses in a certain place.

Our target here is to provide our client with a solid ranking of all prospering business. To define "prospering business", we think from three perspective: the first one is it should be of high score and review of people in recent years. So we choose Yelp dataset. It contains multiple features of business and reviews users, which is very informative. The second one is the business should be constantly discussed by people recently. A hot topic means new opportunities, so we would like to use real-time tweets on Twitter to help us build this one. The third one is it has solid background, mostly in sales perspective, which could be proven by the census data. 2012 sales census started in 2012 and was released recently. It is a very important resource. It could tell us a lot about economics status of a business through multiple features, such as geotype, geoid, business, number of establishments, number of non-employer establishments, employer number etc. I am planning dig more information from the geographical perspective and build the business ranking list. To sum these three points up, we could safely conclude that the top ranking business will have a brighter future than others.

# Motivation

(Write a couple of paragraphs describing why you think this analytic is important.

There are innumerous analysis on the Internet about recommendation of business location. A common factor that appear in almost every one of them is current prosperity and future developing prospects of the locations. This factor is hard to measure for someone who wants to start a business in a new city he has never been to. And even for a person who is familiar with the city, it’s difficult to give objective suggestions across the whole city. Except for need of individuals, for chain stores or big companies, it’s necessary to have a system to help choose candidate locations for a new division and then do further market research. Our project focuses on providing rankings of thousands of locations in a certain city

Recently, startups are emerging and business analysis is playing a more and more important role in decision making. We want to explore more business pattern within given city and provide a list of business choice ranking to support business choosing process for startup owners. Our goal is to evaluate multiple business from three perspectives based on location information. We choose location information to merge datasets because in this way we could evaluate whether a given area is a great choice to start a business. The first one of our dataset is Yelp dataset, which provides the reviews and stars of a business. The second one is tweets from Twitter. Tweets are real-time data, and could provide us with information about the hottest business recently. The last dataset is Google GTFS(General Transit Feed Specification) Real-time data in Phoenix. We use vehicle position data to evaluate the traffic status, which could help us infer the density of population and potential customers. These three datasets are concentrating on different aspects. We believe combining them together will help us get more comprehensive and convincing conclusion.

# Related Work

In paper “Mining Quality Phrases from Massive Text Corpora”[[8]](https://paperpile.com/c/RLNPu9/krDQ), the author presented a very interesting improvement of word mining from large datasets. A concept called “Quality Phrases” is used to distinguish meaningful frequent words with false positives. The main characteristics of such “Quality Phrases” are popularity, concordance, informativeness, and completeness. The author presented several algorithms to tackle this problem. Raw frequent phrases are first generated using the Apriori algorithm. After several rounds of optimization and filtering, the original frequent phrases are ordered according to their “quality”. The result is very promising as it shows significant improvement of precision compared with traditional phrase mining algorithms. The time and space complexity of the algorithm is also very efficient and is linear to the corpus size, making it suitable for processing massive text corpora.

This paper[[9]](https://paperpile.com/c/RLNPu9/USCE) gives a comprehensive overview of common sentiment analysis tasks and opinion mining approaches. Opinion mining and sentiment analysis plays an important role in analyzing social network big data, and is a hot business research topic in many companies. Traditional sentiment analysis methods, such as keyword spotting and lexical affinity analysis, are supervised classification algorithms and are often biased towards the linguistic corpora’s source. Statistical methods use mathematical models and machine learning methods to learn the relationship between a statistical distribution and an affective tone, and can out-perform keyword spotting and lexical affinity analysis when analyzing passages, but have poor precision on smaller text units like sentences or clauses. Concept-based approaches utilize Web ontologies or semantic networks to accomplish semantic text analysis, which can detect subtly expressed emotions. However, it also heavily relies on a comprehensive knowledge system and is not flexible. In conclusion, current sentiment analysis methods all have their corresponding advantages and drawbacks, and a better approach is needed to comply with future demands.

In article "Business Location Decisions in the United States: Estimates of the Effects of Unionization, Taxes, and Other Characteristics of States"[[10]](https://paperpile.com/c/RLNPu9/lCmQ), the author examines how corporate decisions about the location for a new manufacturing plant in the U.S. are influenced by unionization, taxes, and other characteristics of states. So we decide to use the 2012 sales census data, based on resourceful location information, detailed industry taxonomy and other features like taxes to build our model.

Besides, in paper “Business data mining - a machine learning perspective”[[11]](https://paperpile.com/c/RLNPu9/dUkJ), the author states the importance of data mining and machine learning. Data mining is a fast growing application area in business. Machine learning techniques are used for data analysis and pattern discovery and thus can play a key role in the development of data mining applications. Understanding the strengths and weaknesses of these techniques in the context of business is useful in selecting an appropriate method for a specific application. We want to use some related techniques to explore more about the data sources.

Racherla and Friske[[12]](https://paperpile.com/c/RLNPu9/exAK) found papers on topics of review analysis hadn’t cover the social aspects of reviewers and consumers. So they based their research on three sets of factors: reviewer characteristics, review characteristics and consumers’ decision context. They used several factors to evaluate these three factors. And they applied their analysis to three different services: search-based services, credence services and experiential services. OLS regression was used to test the hypotheses and drew the conclusions. The results show that reviewers’ reputation is positively correlated with perceived usefulness while reviewers’ expertise is negatively correlated with perceived usefulness. And extreme reviews are perceived more useful than others.

Xu and Zhang[[4]](https://paperpile.com/c/RLNPu9/Z3IZ) reported a research focused on differentiating collusive fraud reviews from real reviews on online platforms such as Amazon and Yelp. There’re previous research trying to detect fake reviews in a supervised way which rely heavily on real fake reviews for model training. The authors found that today’s fraud reviews are not only conducted by individuals but rather by a collective group. So this paper found a variety of collective behaviors of reviewers and proposed an unsupervised model to rank reviewers. They thus developed homogeneity-based collusive behavior measures that inspect similar behaviors that a group of reviewers possess when they work for the same fraud team. 1) Their targets must be similar. If a group of users co-reviewed the same businesses, their reviews are likely to be fraud. 2) They give the businesses the same rates (high or low). 3) Many fraud reviews come from first-time reviewers. If a group of people use their account for the first time and review same businesses, their reviews are suspicious. 4) Because the fraud team will try to get things done quickly, the active moments of the participants should be similar. Based on these features, the paper construct a model to identify the reviewers who might involved in any collusive fraud attack.

After discreet discuss, our group decide to use Google GTFS(General Transit Feed Specification) Realtime data in Phoenix instead of government sales census data. We change the thought from using business name as common feature to location information. This could provide us with more location information, like the position of high density of population and potential customers, thus might lead to better choice for startup business location.

# Design

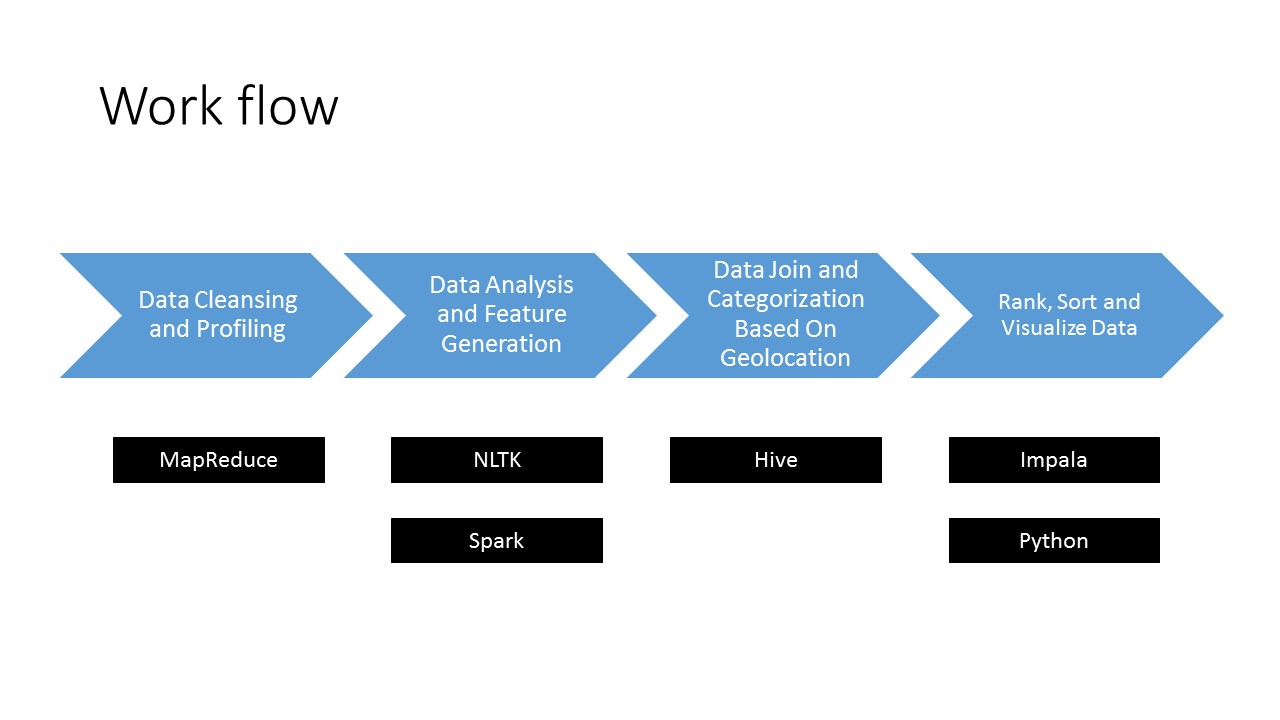


Figure 1. Project Workflow. The original data is first profiled and cleaned using Hadoop MapReduce. Then select features are generated from each dataset using natural language processing and machine learning package such as NLTK and MlLib. The separated dataset is then joined based on the geolocation metadata in each dataset, and is stored as Hive Metastore. Final ranking, sorting and visualizing are applied to the complete feature table using Impala and Python libraries.

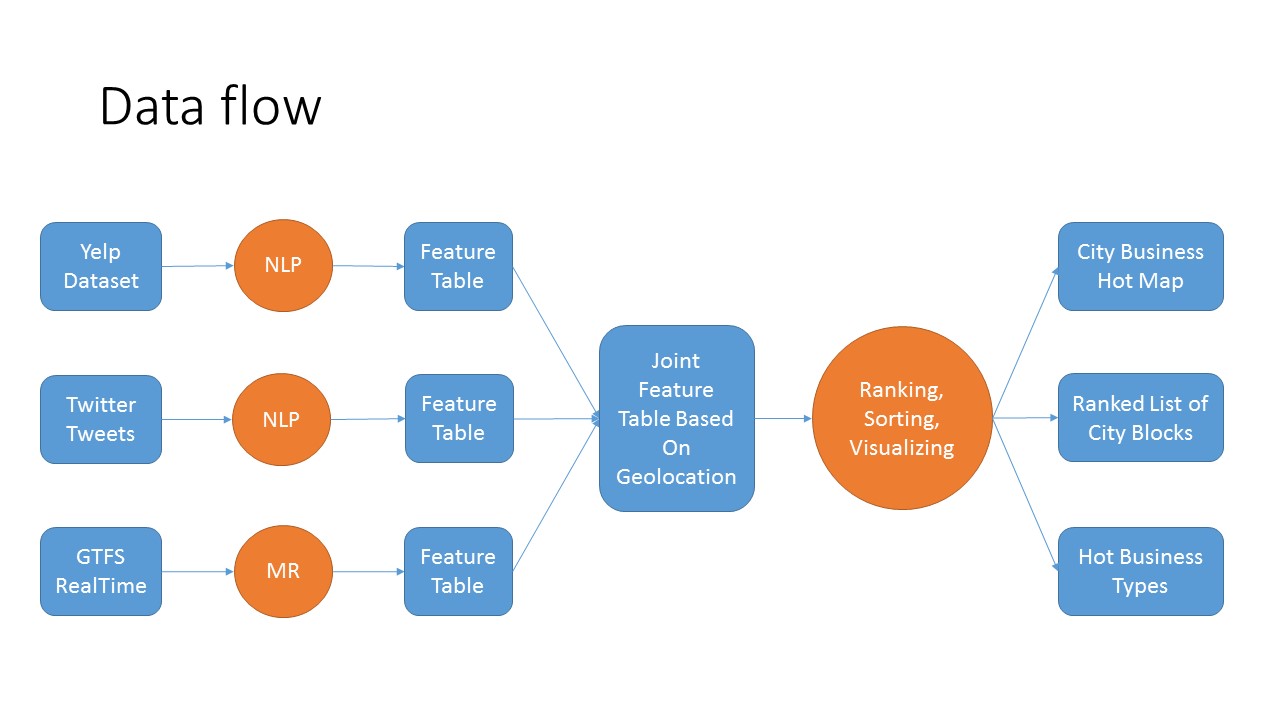


Figure 2. Project Dataflow. Three datasets are used in this project: Yelp Dataset from Yelp Dataset Challenge, Twitter Tweets fetched using Twitter Stream API, and GTFS (General Transit Feed Specification) Real Time Vehicle Positions. Business reviews in Yelp Dataset and Twitter tweets are then analyzed using natural language processing method, including sentiment analysis and phrasal frequency analysis. Selected features are merged based on common geolocations. The final result is generated by doing ranking, sorting and visualizing of the merged feature table.

# Results

(Future… In this section, can describe: Your experiment setup/issues with data/performance/etc.)

# Future Work

(Future… Given time, how would you expand your analytic? Could it be applied to other areas? Etc…)

# Conclusion

(Future… One or two paragraphs about the value/accuracy/goodness of your analytic.)

##### Acknowledgment

The authors thank professor Suzanne McIntosh for her valuable advice and high performance computing at New York University for providing the resources for the authors to work on hadoop and use other related tools. In addition, the authors thank Yelp Inc. and Twitter for their data.

##### References

[1] [P. A. Pavlou and A. Dimoka, “The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation,” *Information Systems Research*, vol. 17, no. 4, pp. 392–414, 2006.](http://paperpile.com/b/RLNPu9/mPbj)

[2] [M. Luca and G. Zervas, “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud,” *Manage. Sci.*, 2016.](http://paperpile.com/b/RLNPu9/3z0c)

[3] [A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in *Proceedings of the 21st international conference on World Wide Web - WWW ’12*, 2012.](http://paperpile.com/b/RLNPu9/cSjV)

[4] [C. Xu and J. Zhang, “Towards Collusive Fraud Detection in Online Reviews,” in *2015 IEEE International Conference on Data Mining*, 2015.](http://paperpile.com/b/RLNPu9/Z3IZ)

[5] [G. Wang, S. Xie, B. Liu, and P. S. Yu, “Identify Online Store Review Spammers via Social Review Graph,” *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, pp. 1–21, 2012.](http://paperpile.com/b/RLNPu9/7FRp)

[6] [F. Nooralahzadeh, V. Arunachalam, and C.-G. Chiru, “2012 Presidential Elections on Twitter -- An Analysis of How the US and French Election were Reflected in Tweets,” in *2013 19th International Conference on Control Systems and Computer Science*, 2013.](http://paperpile.com/b/RLNPu9/b08x)

[7] [Z. Zhang and K. Wang, “A trust model for multimedia social networks,” *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 969–979, 2012.](http://paperpile.com/b/RLNPu9/ofJy)

[8] [J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, “Mining Quality Phrases from Massive Text Corpora,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, vol. 2015, pp. 1729–1744, May 2015.](http://paperpile.com/b/RLNPu9/krDQ)

[9] [“New Avenues in Opinion Mining and Sentiment Analysis - IEEE Xplore Document.” [Online]. Available:](http://paperpile.com/b/RLNPu9/USCE) <https://doi.org/10.1109/MIS.2013.30>[. [Accessed: 16-Nov-2016].](http://paperpile.com/b/RLNPu9/USCE)

[10] [T. J. Bartik, “Business Location Decisions in the United States: Estimates of the Effects of Unionization, Taxes, and Other Characteristics of States,” *J. Bus. Econ. Stat.*, vol. 3, no. 1, pp. 14–22, 1985.](http://paperpile.com/b/RLNPu9/lCmQ)

[11] [I. Bose and R. K. Mahapatra, “Business data mining — a machine learning perspective,” *Information & Management*, vol. 39, no. 3, pp. 211–225, 2001.](http://paperpile.com/b/RLNPu9/dUkJ)

[12] [P. Racherla and W. Friske, “Perceived ‘Usefulness’ of Online Consumer Reviews: An Exploratory Investigation Across Three Services Categories,” *SSRN Electronic Journal*.](http://paperpile.com/b/RLNPu9/exAK)