Location Prediction Using Sentiments of Twitter Users

Ritu Singh(✉) and Durga Toshniwal

Department of Computer Science and Engineering,
IIT Roorkee, Roorkee 247667, India
ritusingh2081992@gmail.com

Abstract. This study aims to predict the next location of a twitter user only by using his past tweets. Twitter is a very popular micro-blogging platform and a lot of people tweet about different topics varying from personal day-to-day activities to some global event. This provides us with the opportunity to perform sentiment analysis on their past tweets for prediction of their next visit. Sentiment analysis helps in revealing the opinion, desire or intentions of a person looking at the text that they write. In this paper, a new model called Sentiments based Labeled LDA model (SLLDA) is proposed to predict users' next location within a given geo-spatial range. This kind of prediction can be used by various establishment owners for the targeted promotions of their products. This can also be helpful for personalized recommendation. Various experiments have been performed to evaluate the performance of the proposed model. The proposed model outperforms in every set of experiments and is better than each baseline model considered in the study. The accuracy comparison has also been done for different window lengths of past tweets and different radii of query. The performance of the proposed model turned out to be better for each set of experiments.

Keywords: Location prediction · Sentiment analysis · Topic modelling
Twitter

1 Introduction

With the increase in trend of using various social media platforms, humongous amount of unstructured textual and pictorial data is generated daily. Various micro-blogs have recently captured peoples' attention where they can share their interests and opinions via short posts [1–3]. One of the most famous and widely used micro-blogs is Twitter. People from all around the world use Twitter frequently to share the details about their day-to-day activities or to express their views on some social event using tweets.

Sentiment analysis or opinion mining is also one of the trending techniques in the field of data analysis to classify the user-generated text into positive, negative, neutral or more sentimental classes. Sentiment analysis is generally applied over the user-generated text like reviews, surveys, social media posts etc. It has been used widely to predict election results, stock market etc. However, no work has been done to predict the next location of users by analyzing their sentiments.

C. Ordonez and L. Bellatreche (Eds.): DaWaK 2018, LNCS 11031, pp. 98–108, 2018.
https://doi.org/10.1007/978-3-319-98539-8_8

The popularity of twitter has led to huge amount of raw data which can be used in many applications like prediction of stock market [4] and election results, friend recommendation, location prediction, home location identification etc. This study deals with the prediction of next location that a user might visit by seeing only his past tweets. The location here means the category of location i.e. restaurant, gym, pub etc. Various establishment owners for their promotional activities can use this kind of prediction. They can focus and send their promotional messages only to a group of people who are likely to visit their kind of business in their proximity in near future, thus saving a lot of money. Knowing someone's location beforehand can also be used by many recommendation systems like location or movie recommendation. This can also be helpful for the common users in filtering out the best available options for them, as they would be getting personalized recommendations and targeted promotions based on their future interests.

In this paper, next visit of a twitter user at a given time and within a given geospatial range is predicted by combining topic modeling and sentiment analysis techniques. First, the past tweets of the user has been analyzed sentimentally to get a glimpse of what the user actually wants to say and then to predict the next visit of the user, topic modeling and sentiment analysis has been combined. For example, if a user writes I am not hungry, then topic modeling alone might predict the next visit of the user to be a restaurant by focusing on the word hungry. However, what user actually wants to say here is completely different.

The main contributions of this paper are:

- Proposal of a new model that performs topic-sensitive sentiment analysis for the next location prediction.
- Performance comparison of the proposed approach on different set of experiments.

The rest of the paper is organized as: Sect. 2 describes previous related work in the field of sentiment analysis and location prediction. Section 3 presents in detail the proposed model for the prediction of next visit of a twitter user. Section 4 talks about the dataset used, various experiments, and their results obtained. Finally, Sect. 5 concludes the paper.

2 Related Works

Sentiment analysis is a trending technique in the field of data mining and data analysis. Speriosu et al. [5] applied a label propagation method for sentiment analysis on twitter. The proposed approach leveraged the follower graph of twitter. Tweets, users, hashtags, unigrams, bigrams and emoticons were used as the nodes for constructing the graph.

Cui et al. [6] came up with an algorithm for sentiment analysis on twitter. The algorithm was based on label propagation by analyzing the emotion tokens. Wang et al. [7], instead of using emotion tokens, proposed a model based on graphs that leveraged co-occurrence of the hashtags to find out the sentiment of other hashtags.

The other sentiment analysis technique is lexicon based which uses external lexicon, like SentiWordNet [8], LIWC lexicon [9] etc. to train data. Thelwall et al. [10, 11]

proposed sentiStrength sentiment analysis technique that applies various lexical rules to the sentences to identify its sentiment. Lexical rules adopted by the algorithm are spelling correction algorithm and consideration of booster words to increase or reduce the sentiment strength of the subsequent words. The major drawback of this technique is that it can't be effectively applied to tweets as the word list it relies on is very small in length.

Ortega et al. [12] proposed a three step unsupervised sentiment analysis technique in which preprocessed tweets are detected for polarity. A rule-based classifier was applied as the last step. The classifier and polarity detection were based on SentiWordNet and WordNet. SentiCircle [13] proposed by Hassan Saif et al. captures the contextual sentiment of a word and hence the tweet. A word's actual sentiment is expressed by the context in which it is talked about. To calculate the overall sentiment of a word, they considered all the context words of the main word and made a term-context vector out of it.

Mathew et al. [14] predict the movement of an individual by clustering the location histories visited by an individual according to the locations' characteristics i.e. the time in which they were visited. They train the Hidden Markov Model for each such cluster formed and hence capture the sequential relations between places visited in given time. Bao et al. [15] utilized the growing popularity of location based social networks in understanding a user's preferences based on his location history. They presented a recommendation system that is location-based and is aware of the user's preference.

Yuan et al. [16] proposed a point-of-interest (POI) recommendation system based on geographical and temporal influences aware graph to model geographical influences, check-in records and the corresponding temporal influences. The approach proposed by them works only if the location histories are available for the users.

Arun et al. in [17] put forward a novel approach to predict the next location visited by twitter users only by looking at their past tweets. They considered two feature sets in order to do so. First, one was the personality traits of the user and second one being the users' past tweets. They applied a famous topic modeling technique, Labeled LDA [18], along with time factor on the past tweets of user. However, their major drawback was not considering the mood or opinion users expressed through their tweets. This gave us the opportunity to also consider the sentiments of a person towards a particular place type by looking at their tweets and then make the predictions.

However, the technique proposed in this study is based on a famous topic modeling approach, Labeled LDA (LLDA). It integrates sentiment analysis and LLDA for the purpose of topic modeling i.e. predicting the next visit of a twitter user.

3 Proposed Approach

3.1 Label Tweets with Location Categories

The raw tweets obtained from Twitter are not labeled with the location information. So to label the tweets with location information, all the tweets having location information i.e. geo-coordinates and location indicative words like @, I am at, I am @, I'm at, I'm @, i am at, i'm at, i'm @ are filtered out.

If similarity, as given by Eq. (1), between three words written after the location indicative words and place names returned by Google Places API [19] is greater than or equal to 75% then the corresponding tweet is labeled with the location category information of the matched entry of GPA.

$$Sim(w_1, w_2) = \frac{|w_1 \cap w_2|}{|w_2|} \quad (1)$$

Where $|w_1 \cap w_2|$ is the number of words common between w_1 and w_2, $|w_2|$ is the number of words in w_2.

3.2 Build Training and Test Datasets

The next visit of a user is predicted here by looking at his past tweets. For this, users are divided into two groups, one with greater than 60 location tweets in their timeline and rest of the users are put in other group. Training dataset is made from all the location tweets (except latest 50) of the users belonging to group-1. Test dataset-1 is made from rest of the tweets i.e. latest 50 location tweets of group-1 users and test dataset-2 is completely made from users belonging to group-2.

3.3 Sentiment Analysis

The sentiment analysis technique applied in this study is based on the Contextual semantics for sentiment analysis done by Hassan et al. in [13].

The overall sentiment analysis technique is explained in Fig. 1. For context dependent sentiment analysis, first term-context vector is generated from the whole corpus which is the number of times each context word appears with a particular word. Then prior sentiment scores are calculated for each word with the help of an external sentiment lexicon, SentiWordNet. A list of negation words is also used to invert the prior sentiment score of the word. Then term degree of correlation (TDOC) scores are calculated for each word in the tweet and its context terms present in the tweet, as done in [13]. Finally, to reassign the sentiment related to a term, median of its context terms present in that tweets is identified.

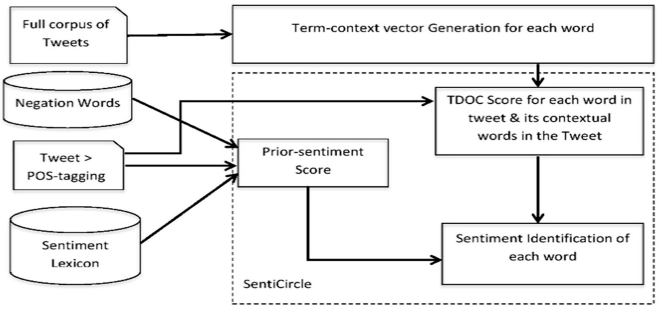

Fig. 1. Sentiment analysis workflow

3.4 Location Prediction Using SLLDA

The purpose of this study is to predict the next most probable location to be visited by a twitter user. The number of such locations is limited i.e. 99 in our case. Therefore, the model developed in this study i.e. Sentiments based Labeled LDA model is based on a famous topic modeling technique called Labeled LDA. In LLDA, there's a multinomial distribution of each label over all words and multinomial distribution of each label over a document. Also, each word in the document is derived by the preference of document for a label and how much a label prefers the word. Here, labels are the categories of the location to be predicted.

This study proposes a new technique for location prediction that is based on LLDA and sentiment analysis discussed in previous part. Assume D is the set of documents i.e. D = {d1, d2… dM} and each document is a set of words that comes from vocabulary V of distinct terms from the whole corpus. To learn new topics for each document and the words present in it, following steps are carried out:

- Choose some fixed number of topics.
- For each document d, randomly assign each word present in the document to one of the topics.
- Assign sentiment score to each word based on the contextual sentiment analysis technique, discussed in previous section.
- For each word wi in the document d and each topic zi, compute: P(zi | d), P(wi | zi, l).
- Reassign the topic zi to word wi with the probability of P(zi | d) * P(wi | zi, l).

During estimation, the topic-word-sentiment distribution for a particular word is obtained by considering different topics assigned to it at different places and the sentiment score of that word at those places. Thus, in each iteration a sentiment threshold, range of sentiment values for each topic-word combination is obtained for a particular word-topic combination. If the sentiment score of the word at a particular position lies within the range of the threshold in the next iteration, then its sentiment value at the place is considered to be positive or neutral else negative. During inference, the threshold is obtained by looking at the training set.

A new topic is assigned to a word w as per the Eq. (2).

$$P(T) = P(z|d)P(w|z,l) = \frac{P(z|d)P(w|z)P(w|l)}{P(w)} \tag{2}$$

Where, w is the word under consideration, z is the topic assigned to it and l is the sentiment label and d is the document.

$$P(w) = \frac{N_w}{\sum_{j \in V} N_j} \tag{3}$$

Where Nw is the number of times w appears in the corpus.

The first term indicates the chances of new topic z to belong to the document d at position w, as per Eq. (4).

$$P(z|d) = \frac{Number\ of\ words\ assigned\ to\ topic\ z\ in\ doc\ d}{Total\ number\ of\ words\ in\ doc\ d} \tag{4}$$

Second term comes from the word-topic distribution given by Eq. (5).

$$P(w|z) = \frac{Number\ of\ instances\ where\ w\ is\ assigned\ to\ z}{Total\ words\ assigned\ to\ topic\ z} \tag{5}$$

Third term, defined by Eq. (6), is where the word-topic-sentiment distribution comes into play. Number of sentiment labels considered is three i.e. positive, negative and neutral.

$$P(w|l) = \frac{Number\ of\ words\ assigned\ label\ l}{Total\ words\ assigned\ label\ l} \tag{6}$$

3.5 Additional Constraints

Additional constraints considered here is the query radius. The intuition behind this constraint is that people tend to travel short distances more.

4 Experimental Results

4.1 Dataset Description

We crawled the publicly available Twitter data using Twitter Search API [20]. The dataset contains tweets of randomly chosen 4,606 twitter users belonging to New York City who tweet at least 20 times a day collecting approximately 3,200 tweets from each users' timeline. In the dataset, out of 1,05,28,618 total tweets, only 21.34% of tweets are geo-tagged.

Total 99 location categories were found in our dataset, including subway_station, department_store, embassy etc. Out of 4,606 users, 1331 users were categorized in dataset_1 and rest of the users (3275) were put in dataset_2. Also, 2,12,219 tweets were found in dataset_1 and 59,339 tweets belonged to dataset_2. Total number of tweets in training dataset were 1,50,749 and size of test dataset_1 was 61,470 and that of test dataset_2 was 59,339.

4.2 Experiments and Results

We have conducted three sets of experiments to show the effectiveness of our proposed model. Accuracy is measured as per Eq. (7).

$$Accuracy = \frac{Number\ of\ instances\ correctly\ predicted}{Total\ number\ of\ test\ instances} \tag{7}$$

4.2.1 Comparison with Baseline Model

The baseline models considered for the first set of experiments are:

- Labeled LDA.
- Baseline-1: Nearest distance model.
- Baseline-2: Google popularity model.

Figures 2 and 3 compare the results of various baseline models considered in this study. It can be seen from the Figs. 2 and 3 that the proposed approach outperformed the various baseline models. Figure 3 also shows that the proposed approach is better than the baseline models for the users that are not shown to the model during its training phase.

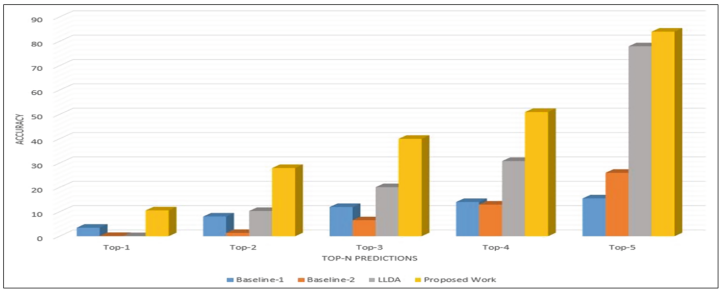

Fig. 2. Comparison of baseline models with proposed model for test dataset_1

4.2.2 Comparison with Different Window Sizes

Past tweets of a user are taken with window slots of 24 h, 12 h, 6 h and 3 h. Initially, experiments were conducted by taking a window slot of 24 h. Figure 4 shows the performance results of the proposed approach for this window slot.

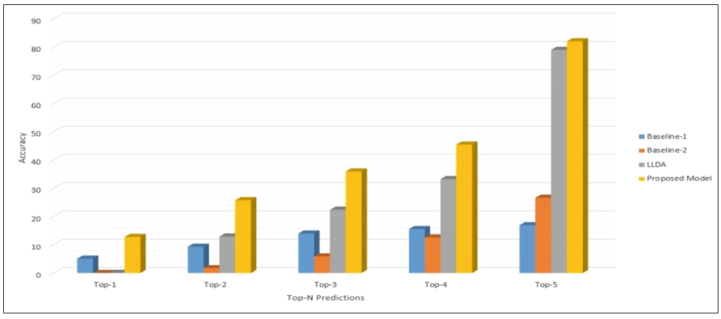

Fig. 3. Comparison of baseline models with the proposed model for test dataset_2

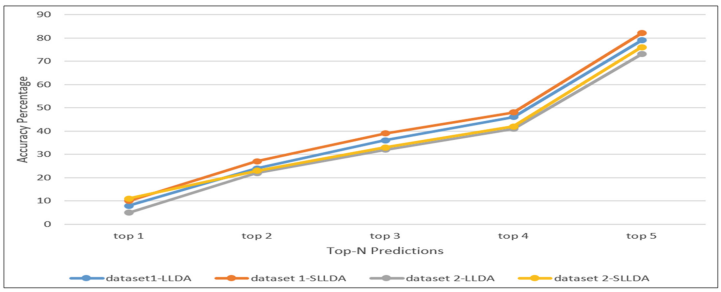

Fig. 4. LLDA vs Proposed model for window size of 24 h

To see the performance of the models in a more specific and precise environment, the size of the window slot was changed to 12 h. Figure 5 shows the comparison of the models for the past 12 h tweets.

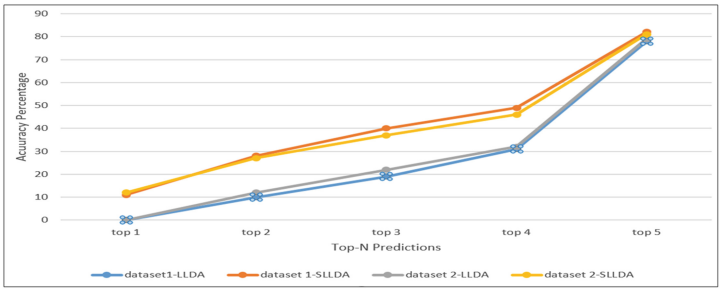

Fig. 5. LLDA vs Proposed model for window size of 12 h

Figures 6 and 7 shows the performance variations of LLDA and proposed approach for the past 6 h and 3 h tweets respectively.

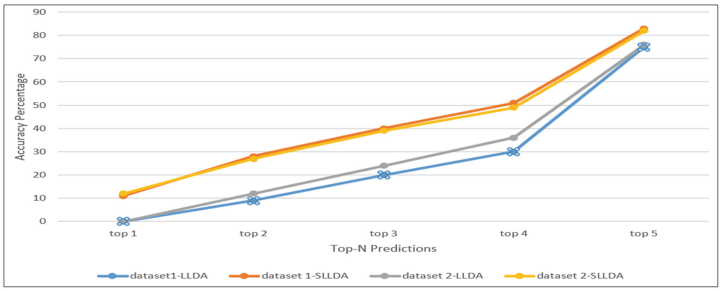

Fig. 6. LLDA vs Proposed model for window size of 6 h

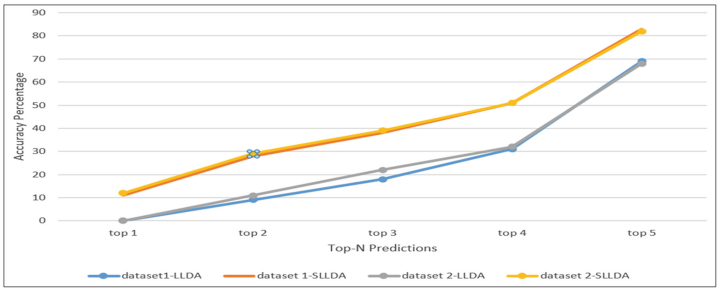

Fig. 7. LLDA vs Proposed model for window size of 3 h

It can be seen from the results of these figures that the proposed model is better than LLDA at any time slot. As the window size of the past tweets decreases, the accuracy difference between the proposed model and the LLDA model increases.

4.2.3 Comparison for Different Radii

To see the effect of query radius, experiments are done on the places at different distances of 300 m, 1000 m and 2000 m from the query point for the window size of past 12 h tweets.

Figure 8 compares the LLDA model and the proposed model for different radii. As it can be seen from the graph that the proposed model performs better than LLDA at different radii too. This shows that including sentiment analysis on the peoples' tweets actually helps in getting predictions that are more accurate.

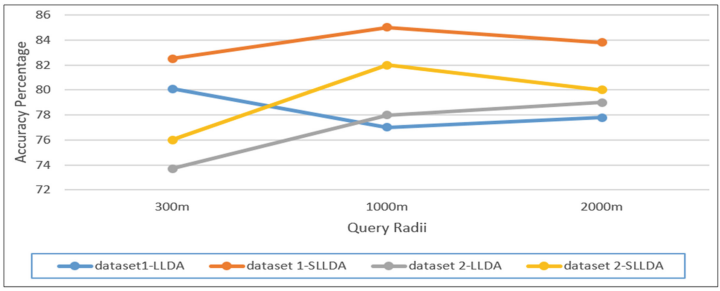

Fig. 8. LLDA vs Proposed model for different radii

5 Conclusion

This study is dedicated to predict the next location of a user within a geo-spatial range. The location considered here are the location categories like restaurant, shopping complex etc. In past, researchers have solved approximately the same problem using either just the text of tweets or by analyzing the user behavior. However, none have considered to use the sentiment of the words.

Therefore, this study extends the previous topic modeling technique by combining it with sentiment analysis model. It proposes a new topic modeling technique, SLLDA, which predicts the next place of visit based on the word-topic sentiment distribution. Various set of experiments were performed to compare the accuracy of the proposed model with the existing models. All the experiments proved that the proposed model is better than the existing ones.

One of the immediate extensions of the proposed model can be to consider the time of visit of a particular place. Not each place is as popular as the other at the same time.

References

1. Bhattacharya, P., Zafar, M.B., Ganguly, N., Ghosh, S., Gummadi, K.P.: Inferring ser interests in the Twitter social network. In: Kobsa, A., Zhou, M.X., Ester, M., Koren, Y. (eds.) Eighth ACM Conference on Recommender Systems, RecSys 2014, Foster City, Silicon Valley, CA, USA, 06–10 October 2014. ACM (2014)
2. Bollen, J., Mao, H., Zeng, X.-J.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 18 (2011)
3. Budak, C., Kannan, A., Agrawal, R., Pedersen, J.: Inferring user interests from microblogs. Technical report MSR-TR-2014-68 (2014). http://research.microsoft.com/apps/pubs/default.aspx?id=217311
4. Li, Q., Zhou, B., Liu, Q.: Can Twitter posts predict stock behavior? a study of stock market with Twitter social emotion. In: 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, pp. 359–364 (2016)
5. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First Workshop on Unsupervised Learning in NLP (EMNLP 2011), pp. 53–63. Association for Computational Linguistics, Stroudsburg (2011)
6. Cui, A., Zhang, M., Liu, Y., Ma, S.: Emotion tokens: bridging the gap among multilingual Twitter sentiment analysis. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 238–249. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25631-8_22
7. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in Twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 1031–1040. ACM, New York (2011)
8. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Seventh Conference on International Language Resources and Evaluation, Malta. Retrieved May, Valletta, Malta (2010)
9. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. **54**(1), 547–577 (2003)
10. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. **63**(1), 163–173 (2012)
11. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: entiment strength detection in short informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12), 2544–2558 (2010)

12. Ortega, R., Fonseca, A., Montoyo, A.: SSA-UO: unsupervised Twitter sentiment analysis. In: Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval 2013), pp. 501– 507. Association for Computational Linguistics
13. Saif, Hassan, He, Yulan, Fernandez, Miriam, Alani, Harith: Contextual semantics for sentiment analysis of Twitter. Inf. Process. Manag. Int. J. **52**(1), 5–19 (2016)
14. Mathew, W., Raposo, R., Martins, B.: Predicting future locations with hidden Markov models. In: Dey, A.K., Chu, H.-H., Hayes, G.R. (eds.) The 2012 ACM Conference on Ubiquitous Computing, Ubicomp 2012, Pittsburgh, PA, USA, 5–8 September 2012, pp. 911–918. ACM (2012)
15. Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2012), pp. 199–208. ACM, New York (2012)
16. Yuan, Q., Cong, G., Sun, A.: Graph-based Point-of-interest recommendation with geographical and temporal influences. In: Li, J., Wang, X.S., Garofalakis, M.N., Soboroff, I., Suel, T., Wang, M. (eds.) Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, 3–7 November 2014, pp 659–668. ACM (2014)
17. Chauhan, A., Kummamuru, K., Toshniwal, D.: Prediction of places of visit using tweets. Knowl. Inf. Syst. (2016). https://doi.org/10.1007/s10115-016-0936-x
18. Ramage, D., Hall David, L.W., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6–7 August 2009, Singapore, A Meeting of SIGDAT, A Special Interest Group of the ACL, pp. 248–256. ACL (2009)
19. GoogleAPI (2015) Google Places API. https://developers.google.com/places/documentation
20. TwAPI (2015) Twitter streaming API. https://dev.twitter.com/docs/using-search