City Sight Recommendation System

Xin Su (A20338792) Chengnan Zhao (A20288353)

Introduction

Most traditional online Recommendation Systems only give a list of classic top 10 sights of a city, no matter if the sight is holding an event or closed at the time a user searches for it. Our system would implement a mechanism to show the recent popular sights (holding events, visited most). Therefore this recommendation we provide has a strong timeliness, and it will be a much more practical reference then traditional ones especially for users who visits a new city as a tourist.

We implement our system by using based on data analysis using online social network (tweets from Twitter) sentiment analysis and machine learning to classify events/visits.

Data

Preparation: We create a boot-up list of pre-collected 30 sights' official names of the city "Chicago".

First approach: We tried to collected 2000 tweets for each sight by using the Twitter Streaming API with three combination of the official name as the keywords. For example, we would use "art institute of chicago", "artinstituteofchicago" and "art_institute_of_chicago" for searching the tweets for the Art Institute of Chicago. However this method failed due to the high volume of the total tweets data causing unreachable searching process. Then we tried to collect 5000 tweets with location info from Chicago yet no keywords assigned, but the valid data is too sparse: only less than 20 tweets with the keywords out of 5000 in total are selected.

Second approach: By using the Twitter REST API we retrieved 100 tweets at most with keywords and geo info for every sight. However since people send tweets about many topics, this method still cannot get enough tweets as we want. In the end we get rid of geo info and only search with keywords and get about 2489 tweets related to the 30 sights we choose. With the AFINN lexicon. After processing the sentiment analysis, we got:

- 816 positives
- 430 negatives

- 1243 neutrals

The neutrals is more than the positives and negatives because people intend to describe their activities without using too many words to express clear emotional angles.

Third approach: We manually pre-labeled all the 2489 tweets retrieved above by giving a binary bit to represent whether the tweet indicates a visit or an event related.

Methods

Second Approach: We treat the positive and negative scores like the scores rated by the user of the tweets. Let pos(i) and neg(i) be the positive score and negative score given by the user of tweet i. Let mean(s) be the mean scores of the sight s. Suppose n tweets of sight s are retrieved. Then we use the following formula to calculate the mean score of a sight:

$$mean(s) = \frac{\sum_{i=1}^{n} (pos(i) + neg(i))}{n}$$

Third approach: We used the pre-labeled tweets to make a classification based on the logistic regression with a 10-fold cross validation. And we are able to get a result with an acceptable mean accuracy at 70.75%.

Experiments

Second Approach: When tokenizing the tweets, we changed the txt to lowercase, kept the punctuation, collapsed the urls but left the mentions. Based on the mean scores of all sights, we could get the result rank listed below (by 11/29/2015 20:00):

Top 10 positive sights	Top 10 negative sights
1.57000 grant park	-4.18000 university of chicago
1.53000 field museum	-2.05000 michigan avenue
1.34000 wrigley field	-1.00000 cadillac palace theatre
1.21212 buckingham fountain	-0.67000 tribune tower
1.18000 united center	-0.29000 garfield park
0.97000 shedd aquarium	0.00000 rockefeller memorial chapel
0.88000 navy pier	0.20000 u.s. cellular field
0.75581 maggie daley park	0.21000 cloud gate
0.71000 museum of science and industry	0.27273 holy name cathedral
0.71000 chicago history museum	0.34000 john hancock

For the positives, the Grant Park has a popular ice ski activity and the Field Museum is holding an event of Greek History and Culture. For the negatives, the University of Chicago received an FBI warning of gun threat recently, and the Michigan Avenue had a protest on Black Friday last week.

The tweets were collected before the news of retirement of the NBA basketball star Kobe Byrant spreads out. For the new results, the United Center will clime up due to the mention of the Lakers game at the Bulls. We collected the data again on 11/29/2015 10:00, and the United Center was ranked at the second positive place.

Therefore all these information explained the result and proved the accuracy that the rank has a timeliness and related to the events held and visits by people.

Third approach: We are able to get a result with an proper mean accuracy at 70.75%, which we treat as that the labels we give to the tweets are within the acceptable area.

Conclusion and Future Work

From the implement of the project we've learnt that the amount of data matters at all.

Some may choose to use others lexicon dictionaries that contains a larger set of word list. However, after we figured out all the tweets during the labeling process, we decided to choose AFINN which we considered more suitable. The reason for this is that the AFINN lexicon dictionary covers most of the common words people use in the usual life, and most tweets in our retrieved set that people talk about the sights only include these common words. Also, from the reading of the tweets we collected over and over for at least twenty times, we found the

ambiguous phrases are infrequent compared to the normal sentences used by people. Therefore a better lexicon dictionary may not bring much enhancement on our system.

The sentiment analysis may give a emotional orientation, but it cannot distinguish whether a mention of a sight name is an event or not. Therefore machine learning comes to help, but only if we have a pre-collected training data set.

There are also some problems to be solved.

First all the approaches are based on the list of the 30 sights' name collected. This may cause some famous sights are ignored. We may combine the traditional recommendation method of using ratings from users and typical content-based and collaborative filtering to form the boot-up list and give the sights of a city more features and profiles. This would be in our future work of research.

Second the only term for searching the tweet for every sight is the official name. This may cause some sights get lower mentions because people intends to use other names. Like the Museum of Modern Art in New York is famous for its abbreviation "MoMA".

The possible solution may be pre-collecting frequently-used names of a sight and setting weights for them. The third concern is that this tweets content and sentiment based analysis would be challenged by automatically creating many positive tweets for growing one's own interest, or spreading fake negative contents against one's counterparts. From the tweets we collected, there were quite few ads selling tickets for theaters and games which indicated there is going to be some events. As theaters always have shows on, this may keep them on the list most of the time. One way dealing with this situation is like handling junk mails, we can build another model to classify ads from general comments.

One additional thought that comes up is besides listing the rank with a timeliness, we can, at the same time, provide the most frequently mentioned words, which may point to to the activities of the reason why people pay a visit. Again, we need to build machine learning models for this to focus on the effective words. This may also need a dataset to use to differentiate a particular set of words from other general phrases. This is also handed over to the future work.