City Sight Recommendation System

Chengnan Zhao: A20288353 Xin Su: A20338792

Recommendation

- Instead of giving top 10 classic sights, we recommend with timeliness.
- Also, we recommend city sights by using sentiment analysis and machine learning to classify events.

Approach

- First approach: collecting tweets as much as possible (2000 tweets for each sight). Failed due to high volume of the data and slow searching process. Also tried to collect 5000 tweets with location but no keywords assigned. Able to retrieve the data, but valid data is too sparse.
- Second approach: search for 100 tweets at most with keywords and location for every sight. Then based on the mean scores calculated by using sentiment analysis of the tweets from the AFINN lexicon. Could get result with convincing rank.
- Third approach: manually pre-label all the tweets with a binary bit indicating the visit/events and use them to make a classification by using logistic regression. Able to get a result with an acceptable mean accuracy.

Data

- City sights boot up list: pre-collect 30 sights' official name.
- Tweets: search for 100 tweets at most for each sight within 7-days with the geo info (this is only for test purpose), totally about 2500.
- Labels of the tweets: manually label these tweets by giving a binary bit to represent whether the tweet indicates a visit or an event related.

Results

- First approach: only less than 20 tweets out of the 5000 are valid tweets with the keywords of the sights. Proved this approach is not applicable for a relatively small amount of data.
- Second approach:
    - Grant park: a popular ice ski activity
    - The Field Museum is holding an event of Greek History and Culture
    - University of Chicago: a FBI warning of gun threat recently
    - Michigan Avenue: protest at the Black Friday

Top 10 positive sights	Top 5 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

    - Also the tweets were collected before the news of retirement of Kobe Byrant spreads out. For the new results, the United Center is going to clime up due to the mention of the Lakers games at the Bulls.
- Third approach: by using the pre-labeled tweets and logistic regression we get a mean accuracy of 0.707

Conclusion

- Lessons
  - The amount of data matters at all.
  - Sentiment analysis 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.
- Problems to be solved
  - All the approach are based on the list of the 30 sights' name collected. This may cause some famous sights are omitted. We are still trying figure out a way to build on a dynamic data set.
  - The only search term for searching the tweet for every sight is its official name. This may cause some sights get lower mentions because people intends to use other names.
  - Manually pre-labeling is time consuming and would cause subjective bias.