———————————————————————————————

Not like traditional recommender which only gives a list of classic top 10 sights of a city, we want to show you the recent popular sights based on data analysis using online social network.

We will recommend city sights by using sentiment analysis and machine learning to classify events and this recommendation has a strong timelines.

———————————————————————————————

First approach: we tried to collected 2000 tweets for each sight, but failed due to the slow searching process. We also collected 5000 tweets with location but no keywords assigned, but the valid data is too sparse.

Second approach: we search for 100 tweets at most with keywords and location for every sight. Then based on the mean scores from the sentiment analysis, we could get the result with convincing rank.

Third approach: we manually pre-labeled all the tweets with a binary bit indicating the visit/ events and used them to make a classification by using logistic regression. And we are able to get a result with an acceptable mean accuracy.

———————————————————————————————

We use a boot up list of pre-collected 30 sights' official names.

Search for 100 tweets at most for each sight within 7-days with the geo info, totally about 2500.

Also we manually label these tweets by giving a binary bit to represent whether the tweet indicates a visit or an event related.

==
First approach: only less than 20 tweets out of the 5000 are valid tweets with the keywords of the sights. So this approach is not applicable for a small amount of data.

Second approach:

For the positives, the Grant park has a popular ice ski activity. The Field Museum is holding an event of Greek History

For the negatives, the University of Chicago had an FBI warning of gun threat recently. The Michigan Avenue had protest at the Black Friday.

The tweets were collected before the news of retirement of 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.

Third approach: by using the pre-labeled tweets and logistic regression we get an acceptable mean accuracy

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

The 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.

There are also some 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.

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.