Term project report

Shuchang Liu (SUID: 968892838)

Introduction

The idea is setting up a timely sentiment analysis engine of social media posts. Users can enter a certain topic, then analysis engine retrieves the latest social media posts about that topic and engages a sentiment analysis. The analysis will compare results from different social media and will analyze the categories of both positive and negative opinions.

The reason for doing this project is to help me keep track of the social media voice. For a certain topic, especially hot topic, societies' opinions are actually the majority and they rapidly change. As an instance, artificial intelligence has been a popular debate issue since two decades ago. While robot threat has been a prevalent topic in Hollywood movies from last century, Google has recently launched AlphaGo that can win Lee Sedol, a professional Go player of 9-dan rank, in a way that like a human learner. However, professional ideas are only a small part of the worlds opinion. Social media opinion is, though not complete, a reasonable complement that may determine the peoples' vision of future.

Requirement

Set up a timely sentiment analysis engine that is easy to understand and implement. Then the project is divided into two parts: the engine and the user interface. The engine provide certain functionalities:

- Social media data retrieval and preprocessing
- Sentiment analysis of the data
- Find Interesting name entities
- Discover relations
- Text summary of the collected data
- Most common topic ever searched

There are also requirements for user interface:

- Query entry point like a search engine
- Change the social media for the query
- Analysis result visualization

Project construction

Design:

Basic structure of the project is a Django website, shown in Figure 1. There are two web pages as user interfaces.

- index.html: search entry point, most common query
- result.html: analysis results

The analysis engine in the middle layer provides sub-services that satisfy functionalities mentioned in the previous section:

- sentAnalyzeFacebook.py: analysis engine for twitter
- sentAnalyzeTwitter.py: analysis engine for facebook
- Miner.jar: provide a function retriveData(topic, number, api_token) to retrive Facebook posts.
- views.py: facade of engines, provide response to query

Finally, at the bottom, lies the database that keep query record and associated statistics:

- QueryRecord (text: CharField, media: CharField, time: DateTime)
- QueryResult (posCount: IntegerField, neuCount: IntegerField, negCount: IntegerField, query: foreighKey QueryRecord)
- MostCommonQuery (queryListJson: CharField, time: DateTime)

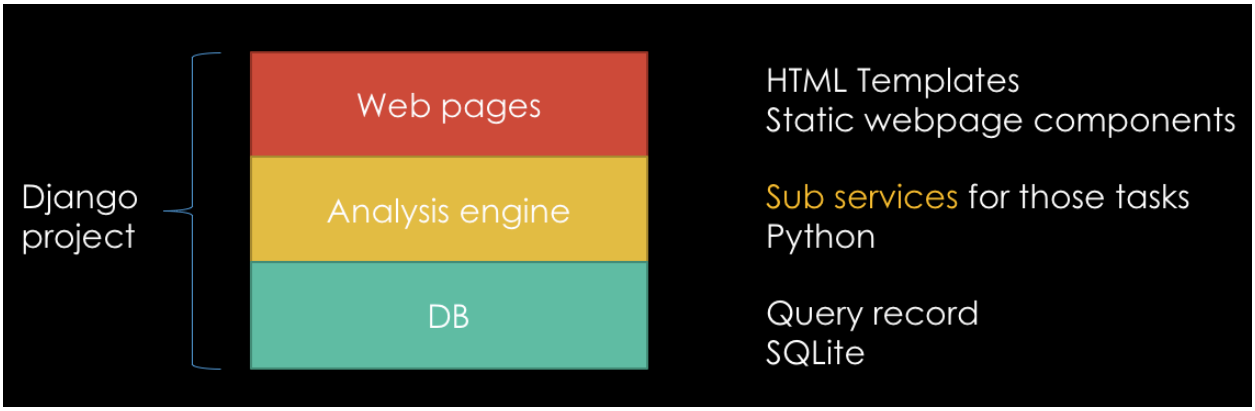


Figure 1 - Basic structure of the website

Implementation detail

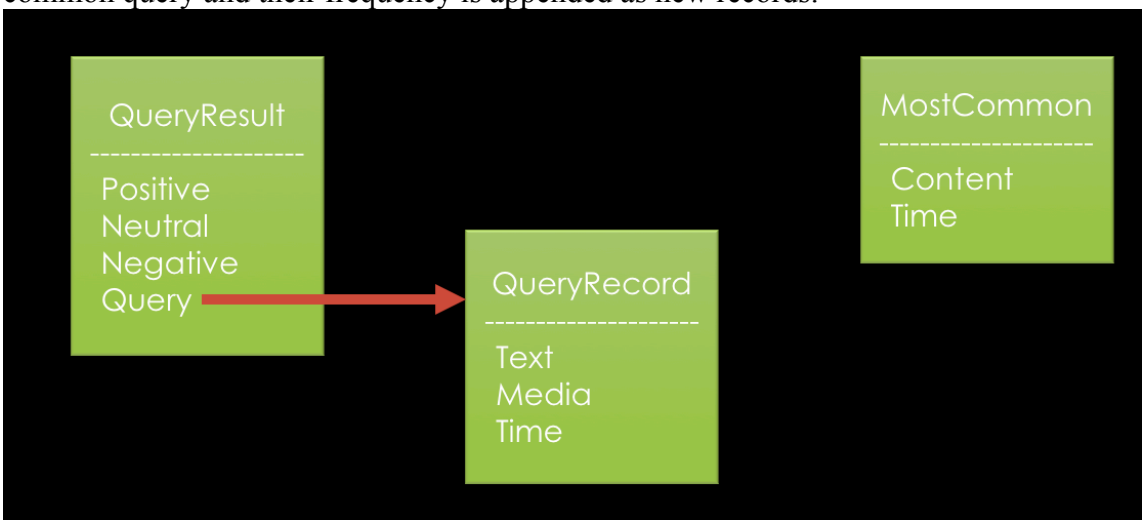
The project uses Python and several framework tools to realize sub-services:

1) **Social media retrieval and preprocessing**
   There are two social media considered in this project: Twitter and Facebook. While the twitter data can be easily obtained by Search API and Streaming API, Facebook API for python is currently not available. So I used RESTFB API in Java and added a python wrapper by Pyjnius.
   Then the data is preprocessed. Since tweet can be retweet, post can be forward, this redundancy is removed. Then for simplicity, I filtered out texts in languages other than English. At last the output data from different social media is uniformed.
2) **Sentiment analysis**
   This is based on the data retrieved from the first sub-service. The data is then analyzed using existing tools:
   Sentiment 140 for Twitter.
   Datum Box analysis API for Facebook.
3) **Finding interesting name entities**
   Similar to the approach in the course, it tokenize, POS tag, and do NER for the data set, using NLTK package.
4) **Discover relations**
   Sem_rel() function in NLTK package is used. Then the relation is determined by co-existence.
5) **Text summary**

Tweets are short, but for a certain query as a topic, it is proper to combine them as a single document. Facebook posts can be relatively longer, but still it may be not enough. So I first combine all retrieved texts and then generate summary using simple summarization algorithm that based on most frequent words.

6) **Most common queries**

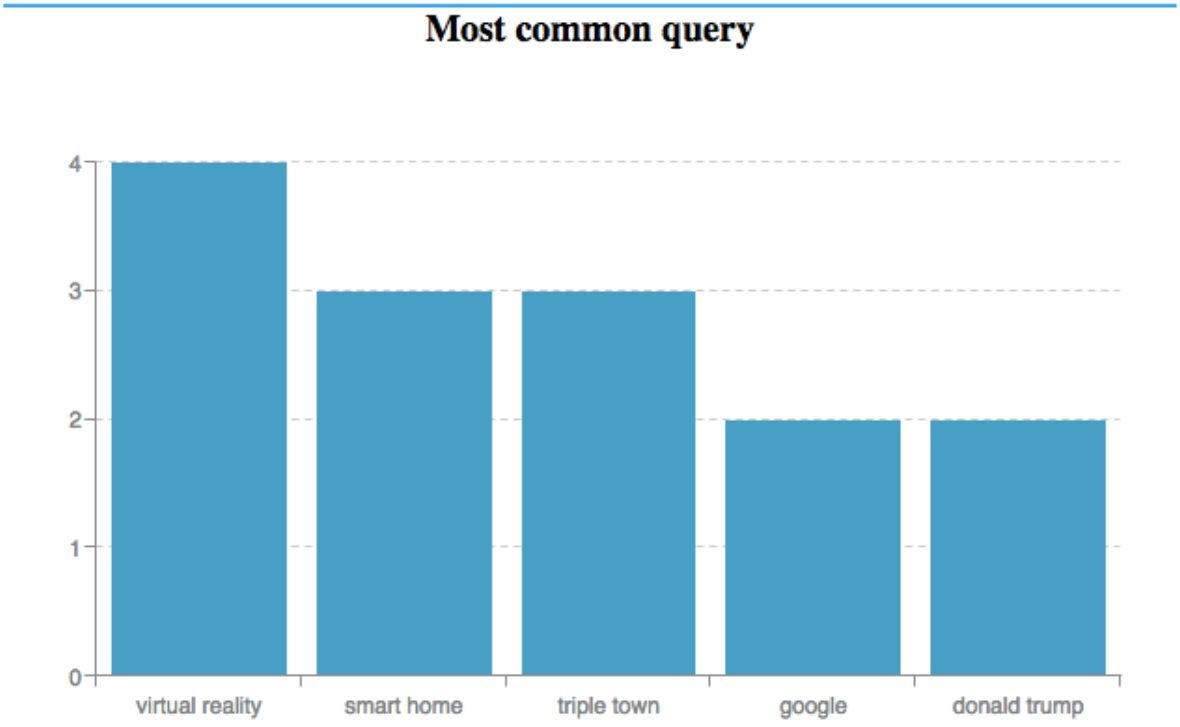
Each query is recorded into the database, so does its results. Moreover, the most common query is also recorded as data record, this means whenever a query is saved the most common query and their frequency is appended as new records.

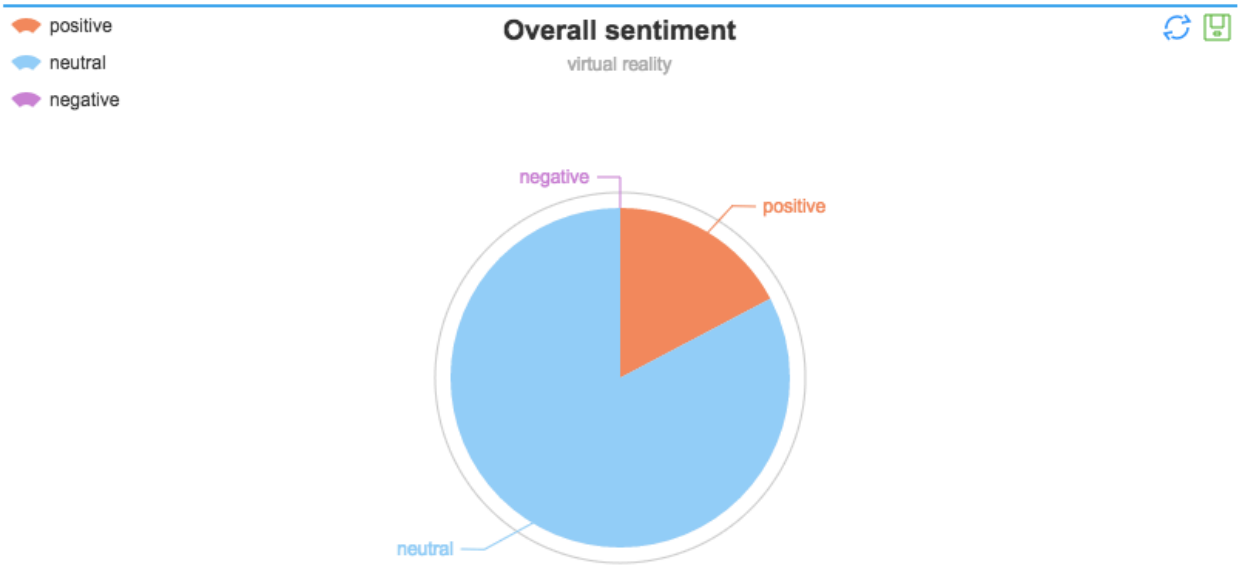


Tools and references

Techniques	Usage
Django	Project framework
Twitter search API & streaming API	collect tweets
Java RESTFB API	collect Facebook posts
Pyjnius	Python wrapper of java class
NLTK	Text to number, NER, text summarization, relation extraction
Sentiment 140 API	Twitter sentiment analysis
DatumBox API	Facebook sentiment analysis, language detection
SQLite	Database, default in Django framework
Echarts, Zingcharts	Javascript visualization tools

Appendix A


Most common query
4
3
2
1
0
virtual reality
smart home
triple town
google
donald trump


positive
neutral
negative
Overall sentiment
virtual reality
negative
positive
neutral

Reasons: related name entities

Summary

RT @AubreyHruby: Eric Schmidt- best way to predict the future is to invent it. RT @InsideRealityVR: Why #VirtualReality Arcades & Theme Parks Will Be a Billion Dollar Industry - https://t.co/qBoxVQn8V6 https://t.co/Y4m…. RT @StanfordHealth: 360° virtual reality tours of the new Stanford Hospital #HealthMatters 5/14. RT @We_Tech_News: Inside the Burgeoning World of Chinese #VR Headsets and HMDs: https://t.co/1JRmc12KtU @RtoVR #VirtualReality #IoT https:/….

List of documents

Positive list:

@mcozma Awesome! We love VR too, could you join and post also at https://t.co/wZUQdxXWmm ?
@CandyCodeApps We're a community of people who love VR, would you join and post also at https://t.co/wZUQdxXWmm ?
RT @AubreyHruby: Eric Schmidt- best way to predict the future is to invent it. 3D printing, virtual reality #MIGlobal https://t.co/D742ogLc…
Now You Can Slack In Virtual Reality (yay?) https://t.co/a5XFJqDkY9 via @Futurism
RT @BillGates: Making my first virtual-reality video with a special guest star... https://t.co/EnCZ5XA9Tj
@themanwho66 It would be amazing if u could join our growing VR community, and post also at https://t.co/wZUQdxXWmm ?
@gafferongames Hey, will you join our growing VR community, and post also at https://t.co/wZUQdxGkXM ?
I liked a @YouTube video https://t.co/bLk5WAWypL 3D Sound - WEAR HEADPHONES - Virtual Reality Audio - WWI
RT @BC_FilmIndustry: The future of filmmaking? New ways to tell stories? #VR #VirtualReality @oculus @VRNewsNow @nofilmschool @VFXSociety h…
@4LTRPress_MIS I think this type of virtual reality combined with actual reality is a great combo for thrill and entertainment.
@inMtHood Awesome! We love VR too, could you join and post also at https://t.co/wZUQdxXWmm ?
@amandalicastro We're a community of people who love VR, would you join and post also at https://t.co/wZUQdxXWmm ?
I liked a @YouTube video from @thiojoe https://t.co/X7Z0rjjAOy Top 5 Awesome Uses for Virtual Reality
Here's an early read on the public interest in VR from @CED. @NewzooHQ https://t.co/s2eSkOdWtI

Negative list:

Discover relations

sentiment	relations
positive	(ORGANIZATION)VR ;
positive	(PERSON)Eric Schmidt ;
positive	(GPE)Virtual ;
positive	(ORGANIZATION)VR ;
positive	(ORGANIZATION)VR ;
positive	(GPE)Sound ; (ORGANIZATION)WEAR ; (ORGANIZATION)HEADPHONES ; (PERSON)Virtual Reality Audio ; (ORGANIZA
positive	(PERSON)Awesome ;
positive	(ORGANIZATION)VR ;
positive	(PERSON)Virtual Reality ;
neutral	(GPE)Re ;
neutral	(GPE)Healthcare ;
neutral	(ORGANIZATION)Rift ;