Intermediate report

1. Progress so far

1.1 data collection and structure

We are doing some prepare work on the data set that we have already collected.

- Real Data:

We are using the data from real data set that collected by Memetracker (memetracker.org). The data collected the phrases/quote in the blogs and news media.

- Testing Data

We also generate a testing data that contains TWO main key words: "Pokemon" and "Olympic" and try to find these two words in 100 lines of data. The time stamp is from 2016-07-10 to 2016-07-19. We are trying to find the decreasing trend of "Pokemon" from 2016-07-10 to 2016-07-15 and find the increasing trend of "Olympic" from 2016-07-13 to 2016-07-19 in the testing data

Data Example:

T	2016-07-10
C	when u catch a weaker duplicate in pokemon go

1.2 Algorithm

We are using Map-Reduce to solve this problem. Right now we are using two sets of Map-Reduce. The first round is focus on finding the keywords which is the hottest. The second round is focus on finding the trend of the different keywords.

- First Map-Reduce:

The Map process is using to finding out the useful key words out of the plain text. We transfer the data into the following format and send out:

[keywords, timestamp]

The first reduce will collected all the keywords that found and get the amount of all of them and find the top 5 keywords.

- Second Map-Reduce:

The Map process is trying to find out the appeared times of a hot words in a time area day by day. The data from the first reduce will be like:

[keyword, timestamplist]

And we are going to transfer the data into the flowing format:

[timestamp, keyword]

The timestamp will be put in order in the reduce part and the data will put into the structure like:

[timestamp, keywordssets]

Keywordsset is a hashmap describe as below:

Keyword→timestamp

After the two round of Map-Reduce we have got the appeared times of hottest words in different days. Then we can find the trend in it.

Output Example:

```
T    2016-07-10
K    Pokemon
N    5
T    2016-07-13
K    Pokemon
N    2
K    Olympic
N    3
```

2. Work analysis

The work right now has mainly two part: the first Map-Reduce and second Map-Reduce. The first round is not working well right now comparing to the second round.

- First Map-Reduce:

Finding out the keywords in the plain text is much more difficult than we thought before. The testing data is copy from the blogs in twitter we search for pokemon. But the keywords are not just like "Pokemon" but also "PokemonGO", "Pokémon". After we get rid of some useless words we need to find the keywords that have the same meaning but appeared in different format. These things happened in the testing data and will cause much bigger problem when we deal with the real data. Get rid of the useless words is also a big work, now we have a list of the most useless words and compare each words in this big list will cause a lot of time and this can't avoid the author's misspell and some other situation.

- Second Map-Reduce:

After finding the keywords the second part is much easier, but still have a lot of problems. The structure of the data is hard to see the trend because of it is showing with the day not the keywords. Transfer the data set to structure like [keywords, timestamp, times] is still a challenge. Right now we are just performing a new program to working with the output of the Map-Reduce program to get the final result that showing the trend.

3. Future improvement

We are considering to add another Map-Reduce part in the process. The first is to find the keywords, second to get the keywords that have the same meaning merge into one, the third is to find the trend. Also we are trying to showing the trend in the same program but not in another program and have a much easier view of the trend. GoogleChart maybe a good choice to use. We have some knowledge of it but not sure if it can suit for this project but we will find a good way to do that.

4. Experiments

We have run the code with the testing data, it not working well, showing with the same keywords with different format but same meaning and takes a lot of time.

We will run this code with the real data sets at the end of this semester. This run may with a stand along mode in a single computer but we will show the segments number that we send out and give a predict running time if we have a lot of machine to run it distribute to show the effectiveness of our code.

5. Contribution

The section 1.1 and 2is completed by Yue Zhang(a1682285).

The section1.2,3 and 4 is completed by Xueyang Wang(a1690260).