Data collection report

1. Data obtained

We got 2 kind of data base on different use: testing data and real data.

- Testing data: First, we have a really small sets of data at a number of 100. This data set is created by ourselves and for testing the basic process of the code and help programming. Then, we have a small set of data that contains 50000 of data which is selected in the real data sets. This data set is use for test some situation that we may not covered with the data set created by ourselves.
- Real data: this is the real data set that collected by Memetracker (memetracker.org). The data collected the phrases/quote in the blogs and news media.

2. The size of the data.

- Testing data: Testing data is small, just in a number of 100, and a number of 50000.
- Real data: Real data have 96 million of memes from Memetracker. It contains 17 million of phrase which 54% of them comes from blogs and 46% comes from the news media. The whole data sets are split into several files which each file contain the data in a particular month.

3. Format of the data and preprocessing

The original data format is like this:

```
P	http://blogs.abcnews.com/politicalpunch/2008/09/obama-says-mc-1.html
T	2008-09-09 22:35:24
Q	that's not change
Q	you know you can put lipstick on a pig
Q	what's the difference between a hockey mom and a pit bull lipstick
Q	you can wrap an old fish in a piece of paper called change
```

L http://reuters.com/article/politicsnews/idusn2944356420080901?pagenumber=1&virtualbrandchannel=10112

L http://cbn.com/cbnnews/436448.aspx

L http://voices.washingtonpost.com/thefix/2008/09/bristol_palin_is_pregnant.html?hpid=topnews

(P for URL, T for time of the post,

Q for phrase extracted from the text of the document,

L for hyper-links in the document)

In our situation we don't need the Link part so the original data is preprocessing into this:

T 2008-09-09 22:35:24

C that's not change you know you can put lipstick on a pig what's the difference between a hockey mom and a pit bull lipstick you can wrap an old fish in a piece of paper called change

(T for time,

C for contents)

4. General idea.

We think about using Map-Reduce to doing this job. The original data is already like a map and we will be doing thing in following steps:

1) Get the meaningful words in the contents. Get rid of the words like: am, is, are, my, I, you. And get the words after like some verb word or after a special symbol like: "", ——, my Pokemon.
2) Calculate the frequency of the word that appeared in a time area (a day is the minimum time area)
3) Find the most frequently appeared 10 words in month.
4) Find the hot level (times that mentioned in a day) of the word day by day for a month or two.
5) Find the life cycle of the hot words.

5. Contribution

The section 1 and 2 is completed by yue.
The section 3 and 4 is completed by xueyang.