Analyzing High-velocity Real-time data

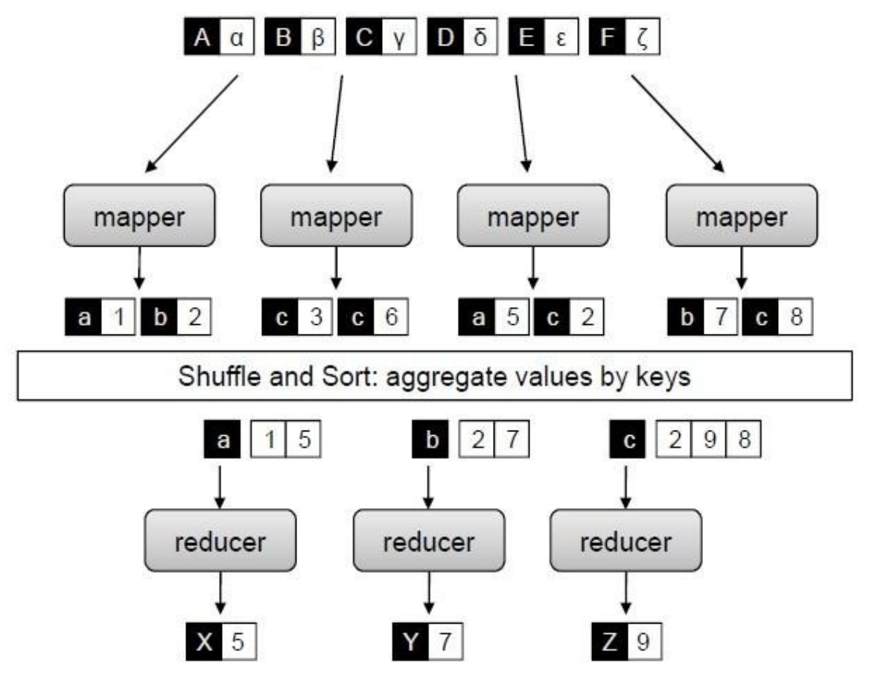
Jinglun Li, Xiang Lin, Bina Ramamurthy, Ph.D.

Computer Science and Engineering Department, the State University of New York at Buffalo NY, 14260

Purpose and Goal

Big-data analysis is getting more attention with explosive growth in Internet and mobile applications. Big-data is defined by four V's (high) Volume, Velocity, Variety and Veracity. We focus on the data generated at high velocity. An example of this kind of data is the tweets sent out by twitter users. Our goal is to analyze the real-time twitter data to find whether the words that from people's discussion can help us to predict the result of future events and the popular things in the society.

Introduction to MapReduce



Firgue1: View of MapReduce

1) Iterate over a large data set
2) Mapper: generated key-value pairs for all input
3) Shuffle and sort intermediate results
4) Reducer: All values with the same key are reduced together
5) Output final result

Implementation Detail

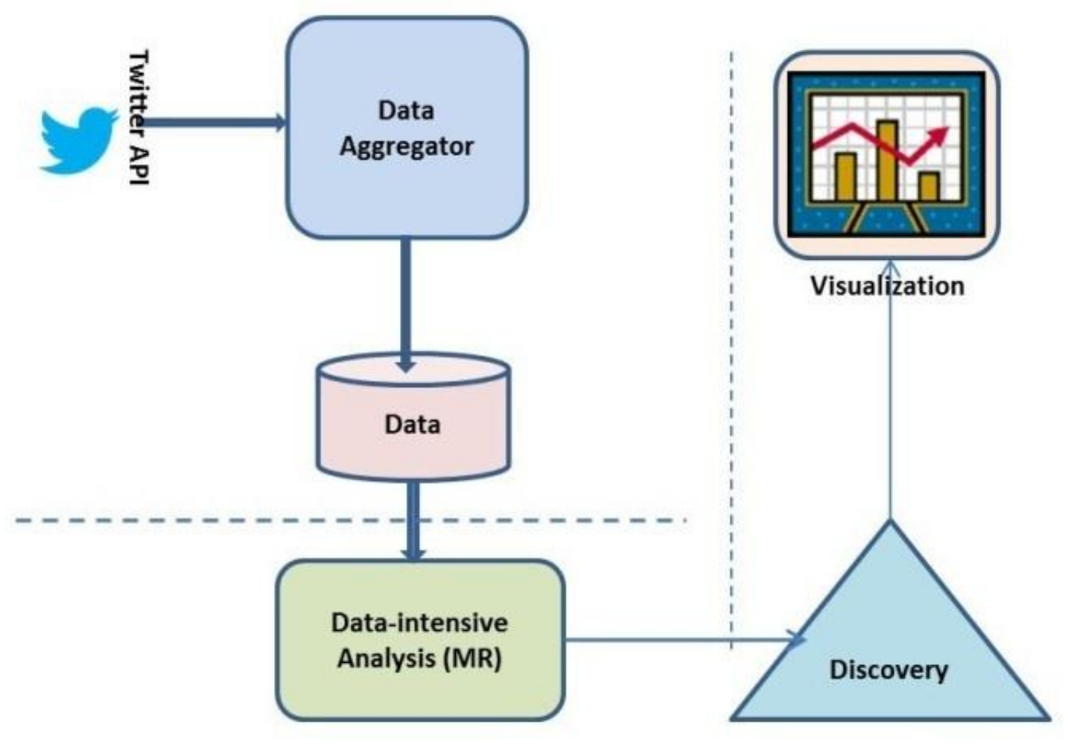


Figure2: System Architecture for Data-intensive analyzer

1) Collect twitter data by Python Twitter API.
2) Filter raw data by Stopwords list.
3) Run Wordcount in MapReduce and sort output data by frequency of words.
4) Discovery interesting words of persons or objects to analyze.
5) Visualize the result.

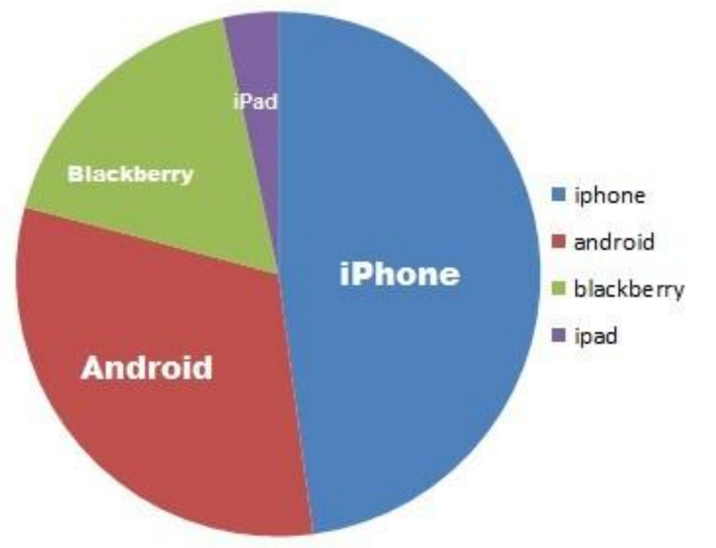
Issues and Challenges

- The raw data we get back will be encoded in JSON(JavaScript Object Notation) which enclosing complicated information in a platform-independent way. We need to re-process the JSON data to precisely extract what users post in each tweet.
- The words that twitter users post in each tweet contain many unnecessary stop words that need to be filtered out.
- Time complexity time is high because there are a lots of word pairs to sort and shuffle around. We need to improve the algorithm to reduce the complexity

Proportion of Mobile Devices in Twitter

Mobile Device	Feb 14	Feb 19
iPhone	370322	380251
Android	241635	219204
Blackberry	134956	117918
iPad	26454	27373

iPhone is the most popular mobile device in Twitter



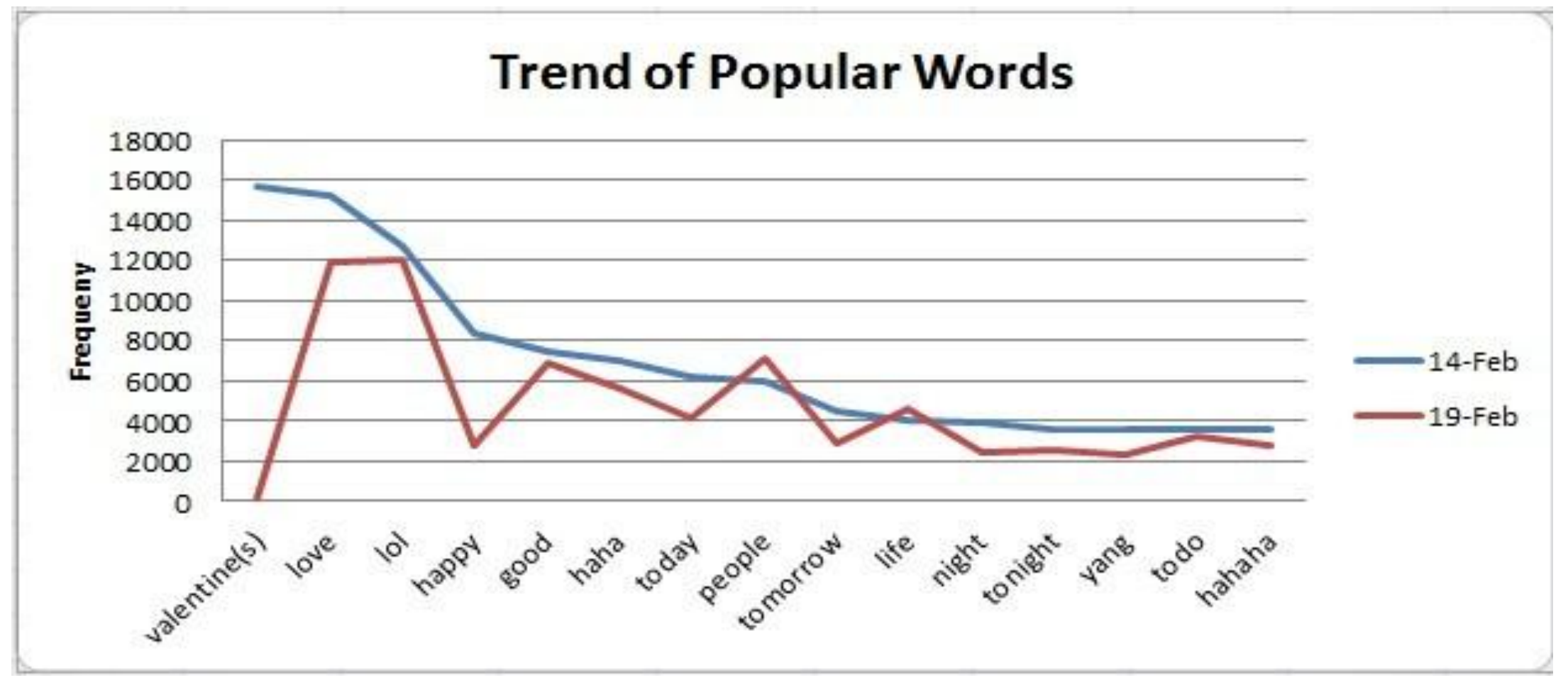
The List of Words that Being Tweet

Words\Data	14-Feb	19-Feb
valentine(s)	15686	214
love	15254	11928
lol	12752	12050
happy	8357	2747
good	7508	6933
haha	7000	5666
today	6255	4101
people	5989	7154
tomorrow	4543	2931
life	3999	4562
night	3927	2468
tonight	3603	2564
yang	3594	2349
todo	3582	3252
hahaha	3562	2813

Only analyze the data from each tweet that post by users instead of whole content, on14th valentines and 19th regular day.

hypothesis :

- The words of "love" "happy" and "valentines" would be most popular on the date of 14-Feb and the frequency of these words higher than 19-Feb.



Result:

- The words of "valentine(s)" , "love" and "happy " appear much more frequently on Feb 14th compare to Feb 19th 2013 because Feb 14th is valentines holiday.
- The "lol", "good" and "haha" are all modal and common words during conversation. Thus the frequency remain similar between these two days

Words Co-occurrence

- Set data as co-occurrence context
- Term of co-occurrence matrix
  - M is square n x n matrix
    - n=|V| (the vocabulary size)
    - Mij: number of times word wi co-occurs with word wj in a specific context, such as sentence, paragraph.
    - Purpose: Infer the interesting points about the events
- Two ways to approach
  - "Pairs" and "Stripes"

Implementation Detail

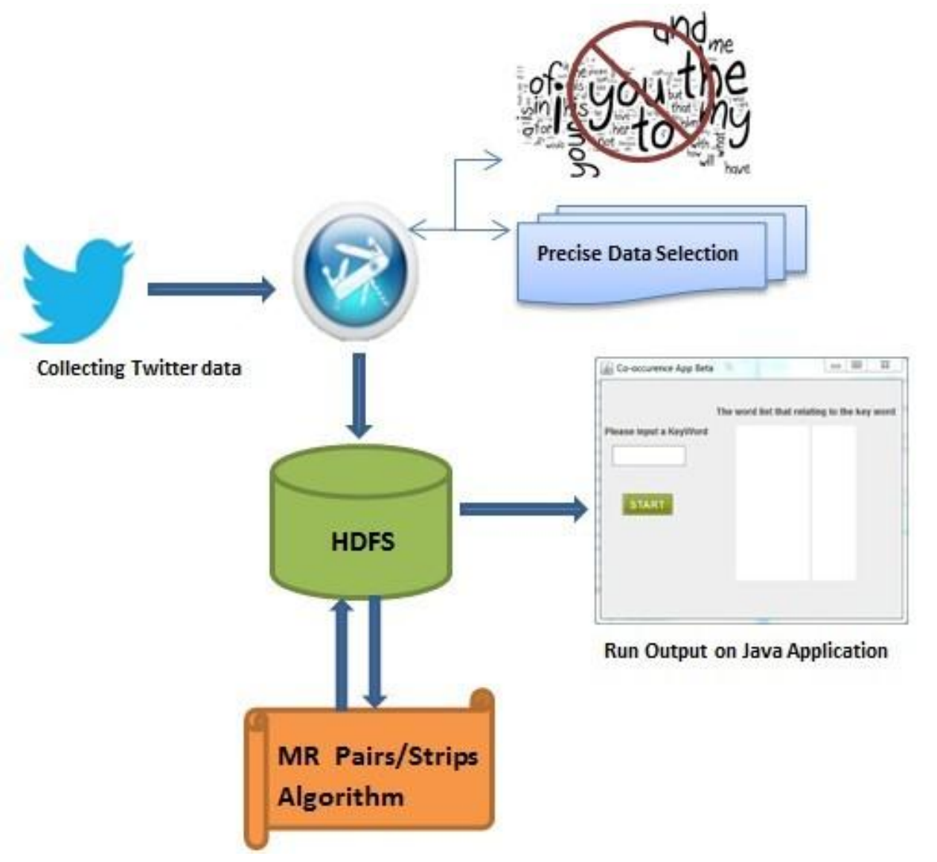
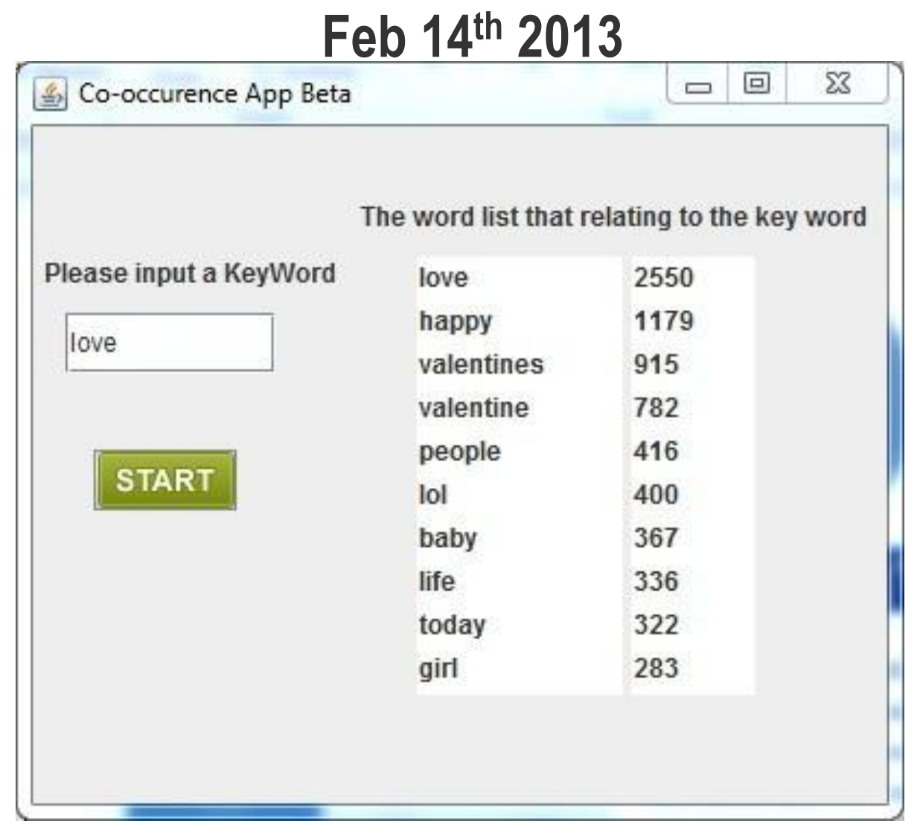

and me of in is you the to my
Collecting Twitter data
Precise Data Selection
HDFS
MR Pairs/Strips Algorithm
Co-occurence App Beta
The word list that relating to the key word
Please input a KeyWord
START
Run Output on Java Application

Figure3: Application Architecture

1) Collect data by twitter API
2) Filter data by Stopwords list
3) Run MapReduce co-occurrence by using "Pairs"/"Strips" algorithm.
4) Output frequency of keyword co-occur with others words.
5) Analyze relation among words that are of interest.
6) Run output on Java App

Java Application for Co-occurrence Output

Feb 14th 2013

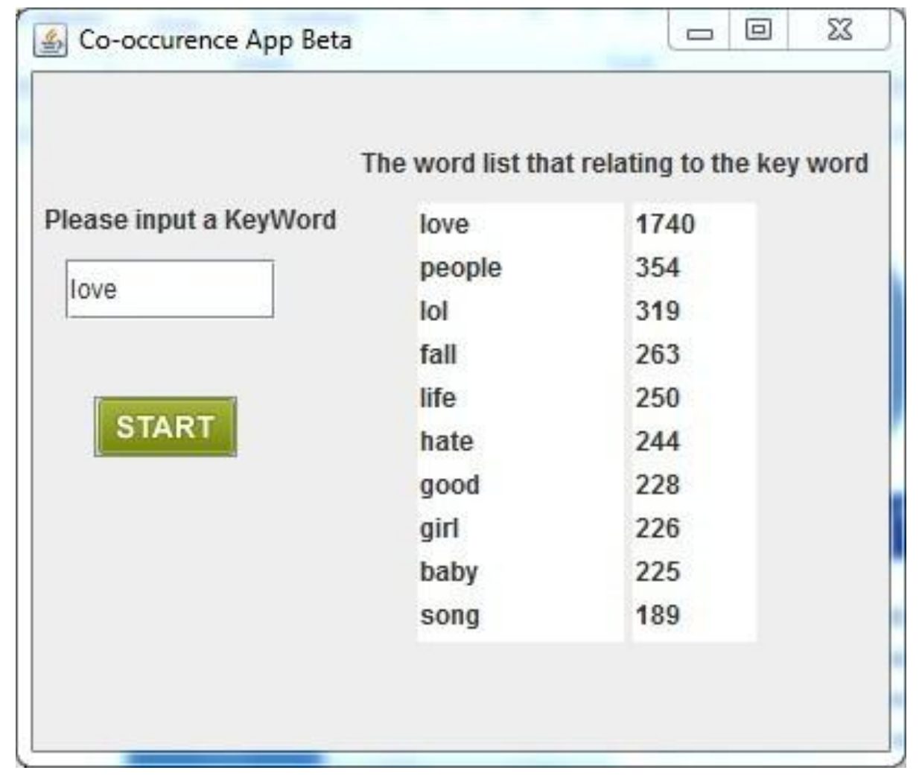


Please input a KeyWord: love

The word list that relating to the key word

love	2550
happy	1179
valentines	915
valentine	782
people	416
lol	400
baby	367
life	336
today	322
girl	283

Feb 19th 2013



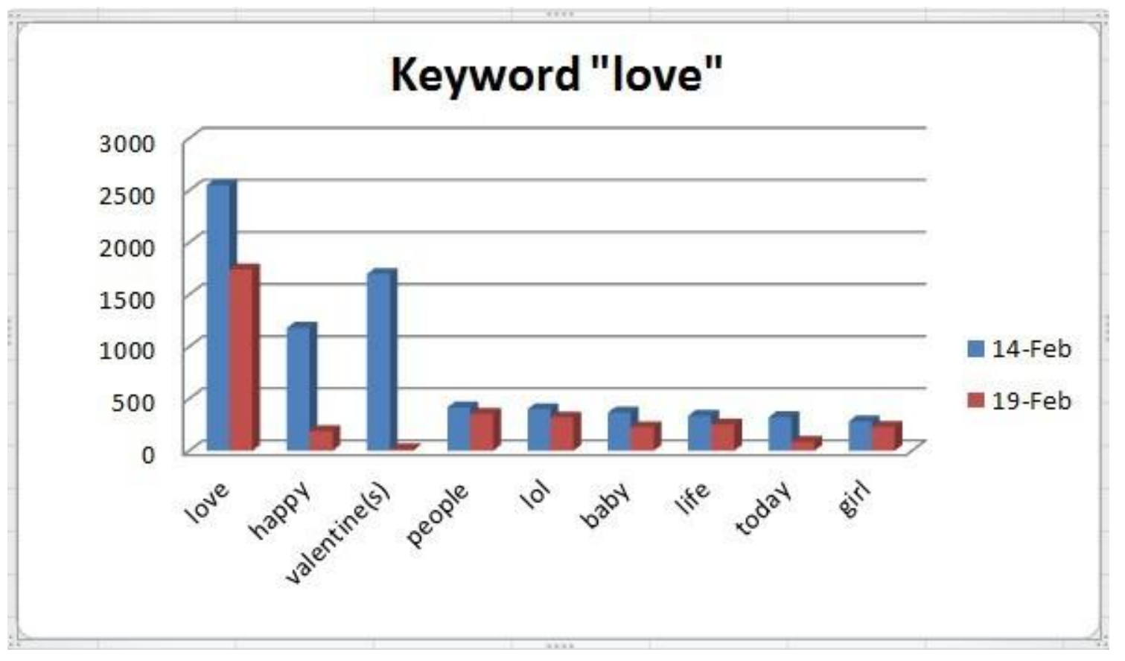
Please input a KeyWord: love

The word list that relating to the key word

love	1740
people	354
lol	319
fall	263
life	250
hate	244
good	228
girl	226
baby	225
song	189

Obviously, for the keyword of "love", the number of times "happy" and "valentine(s)" co-occur with "love" on 14-Feb much higher than 19-Feb

Keyword "love"	14-Feb	19-Feb
love	2550	1740
happy	1179	189
valentine(s)	1697	15
people	416	354
lol	400	319
baby	367	225
life	336	250
today	322	85
girl	283	226



Conclusion

- Through implementing words count and co-occurrence in MapReduce with Hapdoop, we are able to find the relation between words which will improve the precision of the prediction in future.
- Understand the sentiments and trends in the society by analyzing the real-time twitter data.

Acknowledgment and Reference

- Partially funded by the Collegiate Science and Technology Entry Program (CSTEP) UNIVERSITY AT BUFFALO http://cads.buffalo.edu/cstep/
- Partially funded by National Science Foundation Grant NSF-DUE-CCLI-0920335 http://www.cse.buffalo.edu/~bina/cse487/spring2013/
- Center for Computational Research (CCR) UNIVERSITY AT BUFFALO http://ccr.buffalo.edu/

- Four dimensions of Big Data: http://www-01.ibm.com/software/data/bigdata/
- Twitter Application Programing Interface (API): https://twitter.com/twitterapi
- Apache™ Hadoop: http://hadoop.apache.org/
- Lin, Jimmy, and Chris Dyer. Data-Intensive Text Processing with MapReduce. College Park: University of Maryland,, 2010. eBook.