Project Proposal
for
CS 239

Yi Ding Ryan Hsu Lei Ding

April 13, 2016

1 Motivation

In social network analysis, it is useful to predict the future popularity of some subject, hot topic or event. Such analysis can be applied to real world social networks such as twitter, facebook, wechat and so on. If those social network companies can predict the trend of topics in advance, they can make better preparation to make their service friendly for the topic. Besides, they have incentive to do such analysis because it can bring them with commercial benefits such as more advertisement profits.
With the rapid increase of user data within social network, new challenges merge. That is big data. How to do analysis in the case of a huge amount of social network data is a hot topic nowadays. Distributed systems and parallel computing is a key factor to solve this problem. So it encourages our team to utilize such methods to do social network analysis with big data.

2 Proposed work

Tweets are the user data in Twitter. It can show what users care about by their discussion. So what our team aims to do is that with current and previous knowledge of tweet activity with a hashtag, we predict the tweet activity in the future. We may choose a topic, for example, food and then based on current and previous tweets and predict whether some food will be popular in the future.
Our work may consist the following part:

- Acquire tweets data through the Twitter Streaming API, the size of which will be 100GB or larger.
- Build our distributed computing system over AWS or other clusters we can find.
- Design a predicting algorithm (we may use some regression method) and implement it in our computing system.
- Evaluate our system and algorithm's performance.

3 Timeline

$14^{st} - 24^{th}$ *Apr.* $- -$*Collect tweet data*

$20^{th} - 30^{th}$ *Apr.* $- -$*Build our computing platform*

$14^{th} - 30^{th}$ *Apr.* $- -$*Problem analysis and algorithm development*

$1^{st} - 10^{th}$ *May.* $- -$*Implement our project in our computing platform*

$10^{st} - 17^{th}$ *May* $- -$*Result analysis and complete report*

4 Dataset

Due to that most of tweet data online have been deleted. The data we use is mainly obtained by ourselves through Twitter Streaming API.

5 Software Tools or Libaries

We plan to use Spark as the distributed computing platform and use Python or Scala to program. The library we use may be the inner machine learning library MLlib.

6 Team

Two team members: Yi Ding, Ryan Hsu, Lei Ding.