Project Proposal *

Yongjian Mu, u1010337 Shan Wei, u0974032

February 1, 2016

1 Group

This group contains two students: Yongjian Mu and Shan Wei.

2 Data Collection

We would like to do a project which can show the recent **hot news**.

The way to implement this is to collect the each day news text information from two portal sites one is USA Today (*http://www.usatoday.com/*) and the other one is Time(*http://time.com/*) for approximate 60 days. Based on those information, we can use k-grams, and Jaccard to find out the similar news from those websites from two views. The first compare is taking a 7 days information sets from each website during the same period to see how much similarities they have for the instant news. The second is gathering all 60 days data into two sets, processing them respectively then collect the **hottest news** to do some compare.

We may use the tool "scrapy" to collect the news automatically from the website and store the data into the database or files. Then use k-grams by C/C++ or Python to process the data. Finally, use Python or Javascript to display the hot news.

3 Data Structure

The data may contain some basic information like:

- Title
- Source (like where the news come from)
- Date (when was the news published)
- News link

For example:

*CS 6140; Spring 2016 Instructor: Jeff M. Phillips, University of Utah

```
struct INFO
{
        char* title;
        char* source;
        char* date;
        char* link;
};
```

4 Motive

By browsing one website, we could know some latest news, but each website has distinguish different and some news in common. This may lead readers to spend extra but not necessary time to obtain news that they really care about. Besides, we may not know which one is more important. The benefit of the doing this project is that we want to know if they show people same news just in different angles or they have mostly different news on their own. Moreover, for the seven days compare we are trying to explore if both websites editors "value" the instant news from same viewpoint because instant news generally vary a lot.

5 Knowledge

From this project, we can have a better understanding about the algorithms about the similarity, like k-gram and Jaccard, we may choose other algorithms along with this course going through.