Project Proposal *

Yongjian Mu, u1010337 Shan Wei

January 31, 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 recent news from some portal sites like USA Today (*http://www.usatoday.com/*) and Time(*http://time.com/*). Based on these information, we can use k-grams, and Jaccard to find out the similar news from those websites, which are the **hot news**.

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:

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

4 Motive

By browsing one website, we could know some latest news, but we may not know which one is important. The benefit of the doing this project is that we can quickly gather the information from several websites and find out which news are the most important ones.

5 Knowledge

From this project, we can have a better understanding about the algorithms about the similarity, like k-gram and Jaccard.