Data Collection Report

Yongjian Mu, u1010337 / Shan Wei, u0974032

1. Data Source, Size

By using scrapy spider I obtained those major news text data from two websites: {www.nytimes.com, www.time.com}.

At first, I selected and marked all the data that I need on the webpage, then made records of all their's xpath. Through those xpath and other specific features like special id, class, which I can use to figure out those useful data from noise, I build two specific spider.py for each website. Then I get the raw data. Data will be get from the following kind of sentences:

```
ItemLoader.add_xpath(
        'topnews' , '//*[@id="article"]/ div[1]/section/div/article[*]/div/p/text()'
                )
```

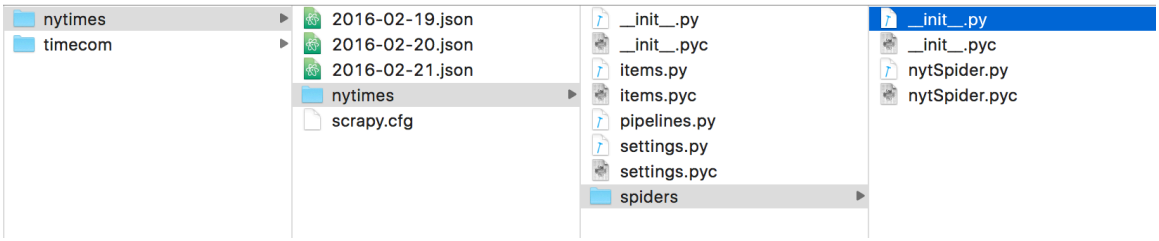
"topnews" is a list name, and also the key of dict(json's object), all text value recognized as this kind data will be the content of this list, also we could say the value of this key.

After saved those data in to lists, there will be a dictionary to store all those lists.

But, after I get those data, I find out there are a lot Unicode and useless whitespace among those text. Then I make some process to the list before they stored into the dictionary like that:

```
list.append(str(t.encode('ascii','ignore')).strip())
```

Finally, I this I initially get over some problems and collect the data I do want. And after three days collect, generally each json file's size is around 4kb. I am planning to collect one month's data. So finally I may have 30*2*4=240kb text.



2. Data Storage and Verify

The Original data from websites are HTML. To make it easy and clearly to input to the program, we can also use scrapy to convert the format to CSV, XML, JSON or even database format. Since database is a heavy weight component in our analysis program, we would not use this format. Among the CSV, XML and JSON format, we would like to choose JSON format since it has less redundancy and is quickly to process by C/C++ or Python.

At present my planning is to create one JSON file for each day, each website as mm-dd-year.json. In the JSON file we have some keys like: date, topnews, sectionsnews. And values will be text data (string), which we want to compare to analyze the similarity.

To verify the validation of the results of the data processing, we can do it in this way. Every time when we use scrapy to grab the data from the websites, we can also save the HTML files in local disk. If we want to verify our results, we can use script to check whether the key words or the items exist in the coordinate HTML files.