Data Collection Report

By using scrapy spider I obtained those major news text data from two website: {[www.nytimes.com](http://www.nytimes.com), [www.time.com](http://www.time.com)}. After processing, those data will be saved as json format. 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, which we want to compare to analyze the similarity.

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 months data. So finally I may have 30\*2\*4=240kb text.

