Project Proposals

IST664

Liu Willow Jiang

Professor:Nancy McCraken

The Reporting Preference and Bias Analysis of New York Times

I want to do a text scraping on New York Times. By scraping the most frequent words in the first four months’ articles which are related to China this year, I want to figure out the most popular issues that NYT focused on and whether or not there’re sentimental biases.

I prefer to work individually, but I’ll ask for professor’s help if necessary. In the end, the final project will be presented in a form of report and programming codes.

The ideal database is the NYT archive, which can be scraped by NYT api. The tools will be python2.7 and some libraries like BeautifulSoup, urllib2, json, etc., which are helpful in text parsing. I tried to filter the articles from the archive and did an estimation of the data, approximately, it’ll be 1000 articles to scrape.

The main tasks of the project will be:

* Creating an api account in NYT;
* Filtering the correct url with api;
* Reading all the data and figuring out the data structure;
* Getting the content;
* Doing word frequency calculation;

Potential problem:

* Api console can’t provide specific filters for getting data;
* There’re different tags for content in html, which can be hard for parsing with the same code;
* Too much content for parsing and the laptop can’t handle that;
* The requests returned by api is not consistent and sending varied results.