Final Report

Liu Willow Jiang

IST664

Professor: Nancy McCracken

Date:05/10/2016

The Reporting Preference and Bias Analysis of New York Times

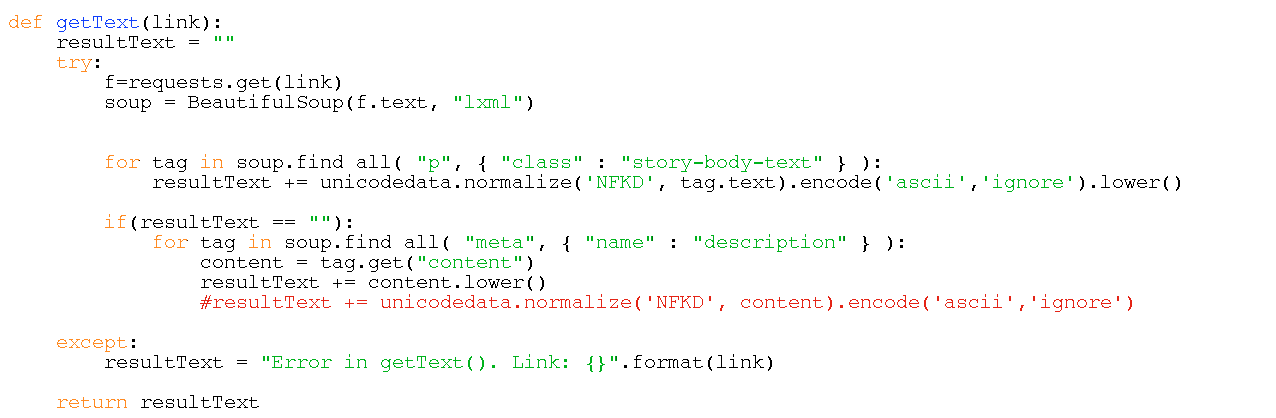
My project is to collect all the articles on New York Times from January to April in 2016 which were related to China issues and calculate the word frequencies to see its reporting preferences, focus and even bias. My data source is New York Time archive, and the tools are NYT api, python2.7 and other libraries of python.

First of all, I scraped all the urls that meet the requirement and read the urls. Next, I cleaned all the text, including getting rid of html structures and tags and only keeping the article content. Last, I used stop words to filter as many meaningless words as possible and NLTK will calculate the word frequency. The ideal result will be printing out the most frequent 100 words in NYT in the first four months, from which we can tell what kind of issues are most popular and if NYT holds any sentimental bias toward Chinese issue.

Now let us see the program.

The first steps were to find all the articles related to China and get the web urls via api.



and then I figured out the structure of the html and got the certain text.

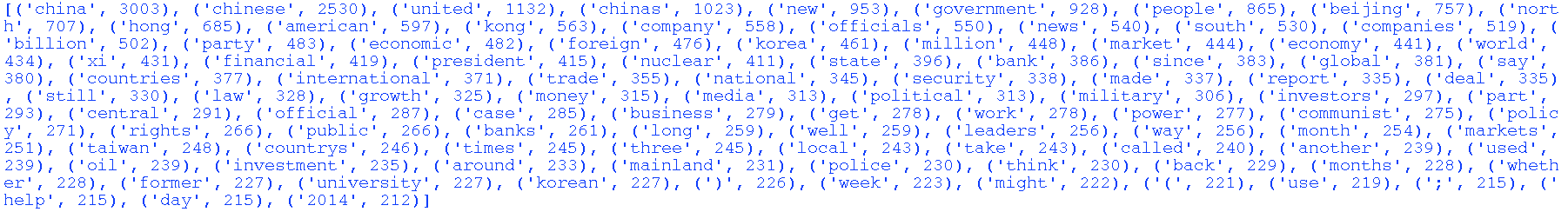
next is to get the frequencies of all the text. Since there’re more than 1000 articles to process, It’s such a gigantic work that the only way I can think about is to use for loop to combine all the frequency of each page together.(note that we say page1, it contains articles from 1 to 10, that’s the NYT style to contain articles within a page.)



The last snippet of code is to refine to use stop words to refine the result. But I have to admit that it’s really hard to get a perfect corpus if key words, even if I have extend the stop words there’re still some meaningless marks and words.



Here is the screenshot of the result, from which we can draw some conclusions.



* The popular topics that NYT focused about China in the first four months are Hong Kong, Taiwan, South and North Korea, what surprised me is that, as a long-time enemy of China in respect of WWII and Diaoyu islands, Japan wasn’t on the list as a reporting object. It seems that NYT didn’t prioritize China-Japan issues this year, instead, it focused on other complicated political issues in east Asia, such as the political tension between Hongkong and mainlad China, the debate regarding Taiwan’s sovereignty, North Korea’ nuclear threat to China, etc. It is difficult to assume whether or not it is the political agenda set by NYT, nonetheless, this result really more or less provides a glimpse of east Asia political situations.
* Among all the topics, ‘economy’ is the most popular one. “company”, “market”, “economy”, “financial”, “trade”, “bank” are always highlights. With the U.S. presidential selection is approaching, the discussion of economy and trade between two countries will become more and more tense.
* No strong preferred adjectives occurred in top 100 tokens, and it proves that NYT is relatively impartial and professional amid other U.S. media. However, ‘communist’, which implied the autocracy and red terror, occurred a lot, and this usuage to some degree reflected how NYT or even U.S. media think about China.