Final Project

Computation for Public Policy

Yuxin Jin

**Introduction**

In this project, I set out to explore the landscape of reports on the environmental issues by New York Times over the time span of 1990 to 2015. The project is of exploratory nature that I hope to build into my understanding of how might media reports on certain issues over time feed into people’s perception and attitude toward them. Coupled with the phenomenon of rapidly rising environmental activism in urban and rural China, which has been documented by a number of scholars, it is an intriguing to examine how might the framing of NYT on China’s environmental issues has over time transformed the public understanding of the problems, though the actual environmental situation remains relatively the same. Below, to elaborate on the puzzling situation, I present my methodology and some of the preliminary findings from the text data compiled using the NewYorkTimes Article API and historical Air Quality Index (AQI) data gathered from China’s Ministry of Environmental Protection (MEP).

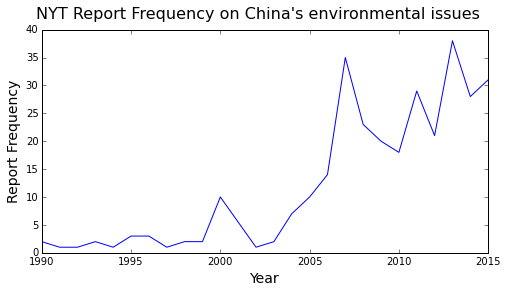
**Methodology & Result**

1. **Article Search and Text Extraction**

I utilized the NYT API to perform article searches and to extract basic information such publish data, headline and url links for each related articles. The build-in search function of the API allows me to search through all the extant NYT articles including news, blogs, opinions and reviews. I used both ‘environment’ and ‘pollution’ for the key query word, and refined the search to all articles that have ‘China’ or ‘Beijing’ or ‘Shanghai’ in their headlines. The searches returned nearly 1,275 distinctive articles published between 1990/01/01 and 2015/12/31.

However, for most of the times, resulted articles are not the ones that I targeted for. For example, in a great number of occasions, the word ‘environment’ is mentioned in the context of addressing economic environment or investment environment, rather than the natural environment. Without better techniques to efficiently filter out the unrelated articles, I examined all the 1,275 article titles and took out the seemingly unrelated ones manually. I realize that only looking at the article titles for filtering criteria is not a terribly reliable measurement. To validate my filtering method, on a smaller set of data, I extracted the entire article bodies and examined the contexts that the word ‘environment’ and ‘pollution’ appears in. The results from the two methods did not differ much. Therefore, given the time and resource constraints, the simpler screening process using only article headlines proved to be strategically efficient. The screened articles have a broad coverage of environmental issues in mainland China, ranging from Dam construction, trash smuggling, water and soil pollutions to the often-heard air quality problems. The manually filter process preserves only 305 articles from the original data set.

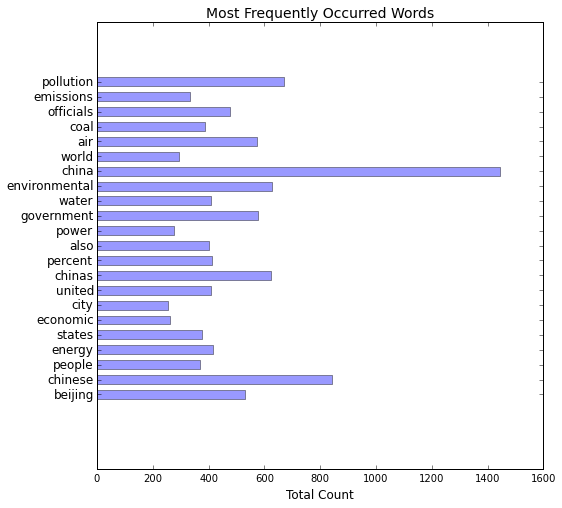
With the dataset of 305 articles on environmental issues of China, gather via the above stated measurements, I then grouped them into the year when they were first published and counted the total number of reports of each year. All the remaining 305 articles were first published by New York Times, with the vast majority of them being categorized as news and few of them as Op-Ed , brief and blog. Below is a graph presenting the report frequency of each year. We can see that the number of reports on China’s environmental issues published by the NTY has been growing dramatically since 1990, with a small peak in 2000 and rapid growing since 2003, reaching major peaks in 2008 and 2013.



Looking at the article contents of those published in 2000, I found that the increasing attention on China’s environment was a result of Beijing being elected as the hosting city for the 2008 Olympic Games. The game had turned out to be a phenomenon showcase for China to the world audience, to which the Chinese government invested tons of resources to turn Beijing into an exemplar of Chinese cities by addressing issues such as air and water pollution. Similarly, the 2008 peak resulted from numerous report on Beijing’s effort to clean up its air during the time of the games, and reports on some of the authorities’ aggressive measurements such as closing down hundreds of factories around the city.

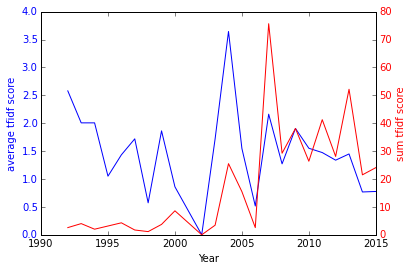
Beyond coverages related to the Olympics, the number of reports has been driving up by the increasing incidents of citizen protests on the environmental issues, including movements against constructions of potentially dangerous chemical-handling factories, raising number of cancers that have been connected to degrading environmental qualities, and air pollutions and health concerns.

For a closer observation of the coverage on China’s environment, I then gathered all the article bodies from the gathered url, tokenized the texts and cleaned up the tokens. Handling the article bodies from all 305 articles aggregately, I then countered the total number of occurrence of each word token. The most frequently occurred words are presented in the chart below, after taking out of irrelevant tokens such as ‘mr’, ‘today’, ‘would’, ’said’ and ‘year’.



The above chart clearly shows that China’s environment has been frequently associated with pollution, air and water problems, and problems derives from coal burning and energy emissions. While Chinese government’s role has received a lot of attention, what the Chinese people have to say about the environment has received possibly equal attention. There is also a curious link between China’s environment and the United States, since both the words ‘united’ and ‘states’ appears on the top list. Admittedly, without further digging into the texts, the exact nature of NYT’s depiction of China’s environment is not clear at this point.

I also attempted to incorporate the TF-IDF score into the analysis. I did this by defining function to calculate the TD-IDF score for the selected word in each tokenized article. As much as I invested a great amount of time in the coding, the actual finding does not seem to offer further insights regarding the text corpus. For example, below I present a graph that captures the TF-IDF score of the word ‘china’, which is the most frequently occurred word in the corpus. The blue line traces the average TF-IDF score of all the articles of the year, and the red line indicates the accumulated TF-IDF score for all articles of that year. While it is reasonable to expect ‘china’ to be roughly equally important in all articles across time, the TF-IDF score received by ‘china’ fluctuates in an unpredictable manner. I did not explore further with the TF-IDF score for this project.



1. **Air Quality data from MEP**

Starting summer of 2006, the Chinese government started to publicize air quality data that they monitor. Since then, information transparency on air quality has dramatically improved within the country. The AQI data for all Chinese cities that are currently being monitored for air quality collected from the official data center of MEP website. I took out only Beijing’s data because 1) the city has the most complete data from 2000 June to present and 2) the city has attracted most of NYT’s attentions in China’s environmental issues, probably partially due to hosting the Olympic Games.

The historical AQI data is presented below in the graph, both daily and monthly average. Simply put, the higher the AQI, the worse the air quality. An AQI below 50 indicates good air quality. Beijing’s air is apparently alarming that its average AQI is about 100 and occasionally have days of extremely hazardous air. The seasonal pattern is probably due to large amount of coal burning during winters for heating. Looking at the general trend, it should be noticed that over the past decade and half, according to the official record, Beijing’s air quality remains much the same, except for a moderate trend of decreasing air pollution from 2000 to 2013 and a visible increase of AQI scores starting 2014. Such sudden worse of the air quality maybe a result of rapidly increased industrial emissions, but may due to improved monitoring technique and more stringent regulation on reporting and publicizing the environmental data. The missing data and the confusing trends posts serious questions on the credibility of the data source, since it is possible that the past records were not honestly reported to the public, and only due to the public pressure that the MEP started to report more reliable data. In general, again, without further exploring the issues, no definite conclusion could be drawn.

