Name Entity Recognition Analysis for Global Times Opinion Articles

October 28, 2018

Fiona Fan ### I. Introduction Global Times is one of the official voices of the Chinese government. Its narratives have a limited set of audience in western countries, as opposed to the hundreds of millions of readers that it has in China. Many renowned scholars and governmental officials are regular contributors to the opinion column of Global Times. All articles published in the column would undergo a strict review process both within the parent news agency, and from the external Party Committee. Thus, the articles within the global opinion column can serve as a good reflection of China's governmental take on a lot of heatedly debated issues, both domestic and global. This analysis performs name-entity recognition analyses on the corpus of all available articles within the column, both in English and Chinese. It identifies the most frequently mentioned objects in both corpi. Hopefully it can allow the readers to identify what the topics of China's interest are to promote better understanding of the Chinese narrative. It can also tell if the column has different areas of focus for its English and Chinese readers. Some sample titles of the articles are as followed:

`China, US will inevitably come to 3rd alternative Xinjiangs soft landing to peace, stability deserves respect Confidence key to coping with Chinas downward pressure Stocks fall augurs trade war uncertainty China-Japan ties move beyond past setbacks Governance in Xinjiang stands on righteous side Trade war lingers despite US Treasury report Khashoggi case triggers wider implications China-US competition could be driving force for human society Taiwans silly folly in aiding trade war Chinas offshore areas not stage for US unilateral show of force`

0.0.1 II. Program Setup

Producers (Website Crawling) The program consists of two producer processes that crawl the websites in parallel, constructing one corpus of English and one corpus of Chinese. Both corpi contain all crawlable articles from "http://www.globaltimes.cn/opinion/editorial/index" and "http://opinion.huanqiu.com/hqpl/" respectively. The corpi are passed to two consumer processes in forms of a multiprocessing queue `q_en \ q_ch` and a multiprocessing manager dictionary `en_dict \ ch_dict`. During the crawling processes I store the corpi into mp dictionaries and store them in pickle files, just in case I need further analyses in the furture.

Consumers (NER Analysis) After the producer processes are complete, I use the `mp.Queue().get()` to pass the results to the consumer processes for Name-Entity Recognition

analysis (NER) performed by Stanford CoreNLP. The NER results returned by the CoreNLP wrapper are stored in another multiprocessing manager dictionary `ners_dict_en \ ners_dict_ch`. There are two consumer processes, one for English and one for Chinese. They run concurrently. The result dictionaries `ners_dict_en \ ners_dict_ch` are dumped into pickle files for further analyses.

Word Cloud Generation Running on local dictionaries from pickle files `ners_dict_en \ ners_dict_ch`, I generate word clouds for the most frequently mentioned name-entities within different categories with sciprt `ner_analyze.py`. Results are stored in the `\output` folder. The pictures are the word clouds created in proportion to the frequencies with which words are mentioned. The text files list the 15 most frequently mentioned entities within the category.

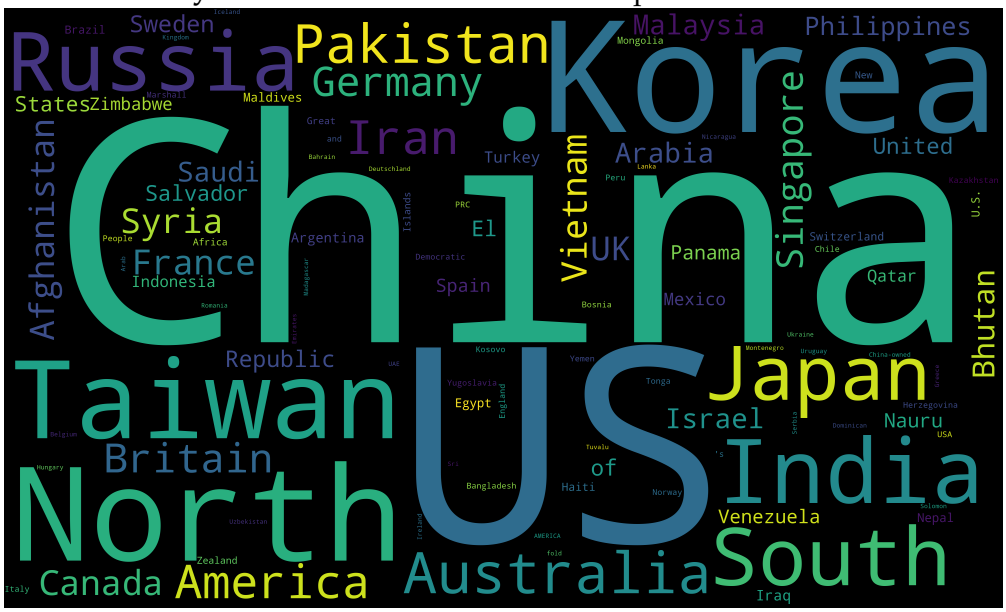
Sentiment Analysis For sentiment analysis, I also run from the local pickled dictionaties `en_dict \ ch_dict`. Based on the results from word cloud generation, I identify and sample the most frequently mentioned words within each category, and sift the English and Chinese corpi for articles mentioning specific words within each category. Then I feed the corpi of pertinent articles to the CoreNLP server and get a sentiment value for each sentitence within the pertinent articles. I store a numpy array for the sentiment sentiment values for each key word in the categories. Calculating the mean and standard deviation of the array, I can get a sense of the sentiments attached to the keyword.

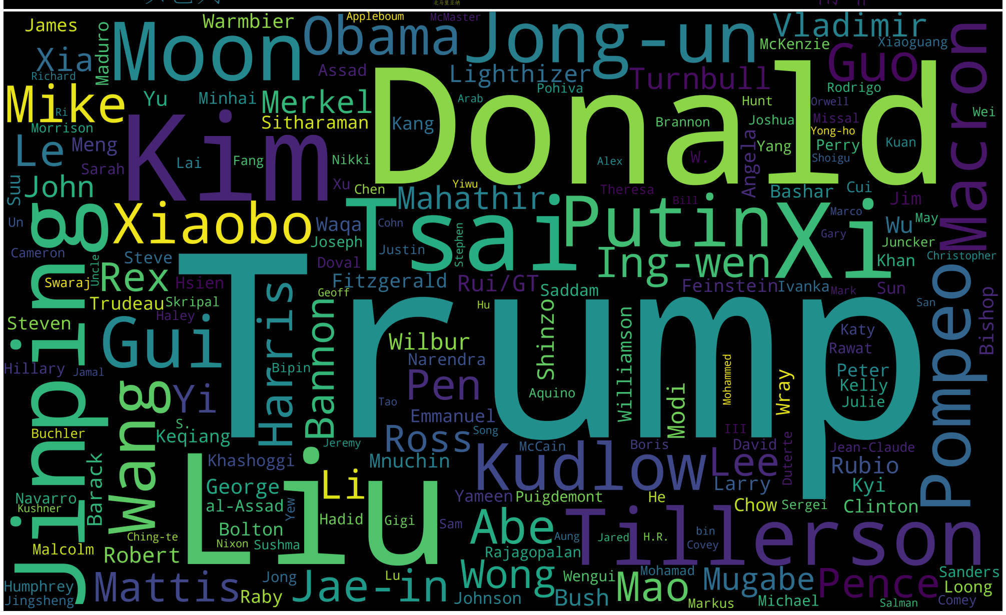
Before running the code, please launch a new AWS Unbuntu instance to host Stanford CoreNLP and change the host address accordingly. Note I modified CoreNLP.sh to include a Chinese analyzing `.jar` within the CoreNLP directory.

The command line I'm using to launch the AWS instance is: aws ec2 run-instances –image-id ami-0f65671a86f061fcd –count 1 –instance-type t2.large –key-name 10-15 –security-group-ids sg-0ec9d69401e65fd3c –user-data file://CoreNLP.sh

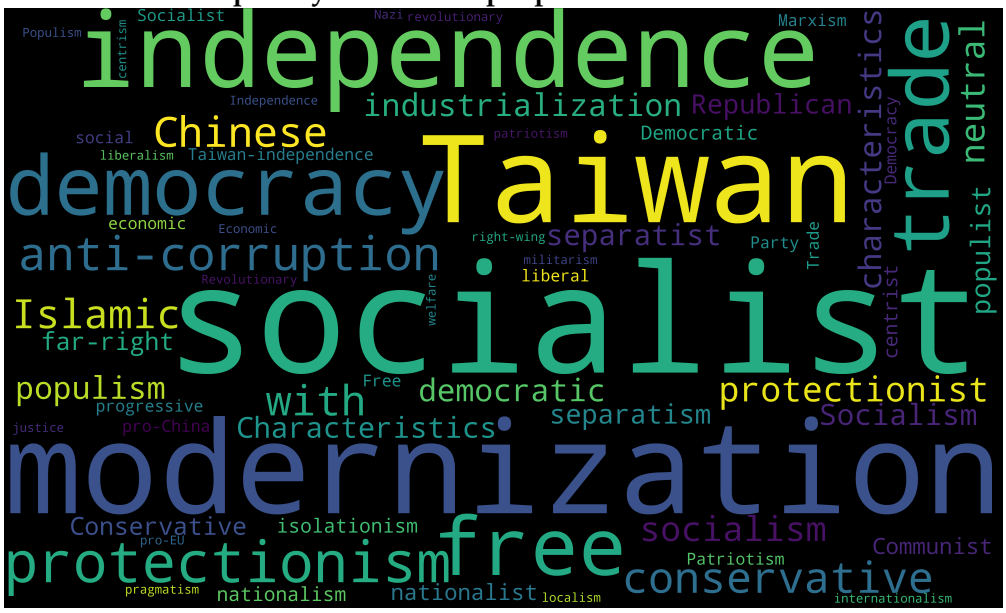
0.0.2 III. Results

Word Cloud Analysis of Name-entities Some sample results are as followed:

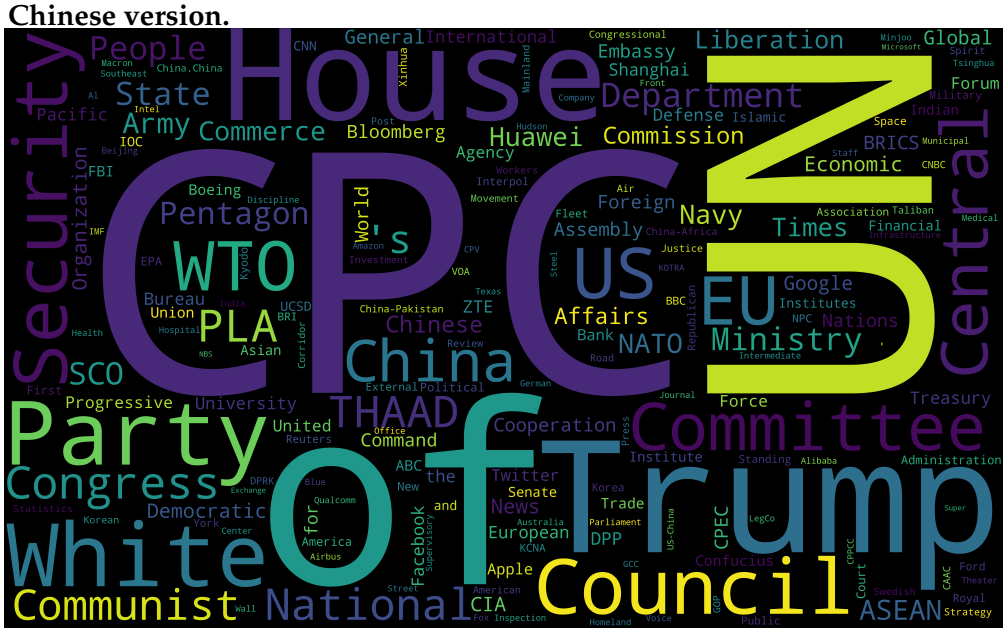


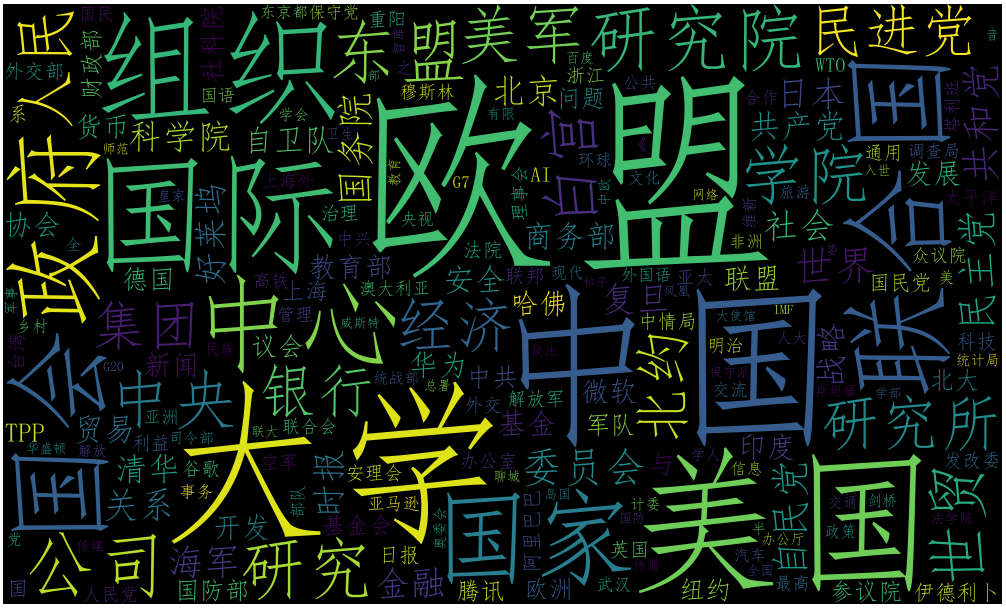

中国
美国
日本
俄罗斯
印度
朝鲜
韩国
德国
英国
澳大利亚
新加坡
墨西哥
巴基斯坦
俄
泰国
加拿大
菲律宾
巴西
法国
越南
土耳其
伊朗
乌克兰
马来西亚
叙利亚
华
印韩
阿富汗
伊拉克
柬埔寨
新西兰
委内瑞拉
西班牙
共和国
美
澳
以色列
南非
津巴布韦
伊
法
巴
英
希腊
意大利
瑞士
中华
波兰
关岛
荷兰
斯里兰卡
人民
德
也门
卡塔尔
吉布提
埃塞俄比亚
利比亚
爱尔兰
尼泊尔
缅甸
埃及
几内亚
瑞典

Trump
Donald
Kim
Moon
Jong-un
Xi
Putin
Tsai
Jinping
Liu
Tillerson
Macron
Pompeo
Xiaobo
Obama
Vladimir
Ing-wen
Kudlow
Wang
Gui
Harris
Bannon
Mike
Abe
Mattis
Jae-in
Guo
Turnbull
Merkel
Rex
Lee
Ross
Pen
Pence
Mugabe
Wong
Mao
Modi
Shinzo
Lighthizer
Mahathir
Williamson
Le
John
Xia
Yi
Li
Rubio
Clinton
Bolton
George
al-Assad
Robert
Malcolm
Mattis
Wilbur
Narendra
Emmanuel
Mnuchin
Khashoggi
Keqiang
Barack
Steven
Hillary
Trudeau
Steve
Fitzgerald
Rui/GT
Saddam
Feinstein
Ivanka
Bashar
Angela
Kelly
Peter
Katy
Julie
Kyi
Jean-Claude
Chow
Sergei
Larry
David
Boris
Jared
Covey
Rajagopalan
Puigdemont
Yameen
Hadid
Gigi
Sushma
Bush
Johnson
Raby
Jong
Humphrey
Jingsheng
Navarro
Kushner
Buchler
Richard
Morrison
James
Warmbier
Maduro
Sitharaman
Minhai
Kang
Brannon
Hunt
Joshua
Yang
Perry
Kuan
Wei
Rodrigo
McKenzie
Xiaoguang
Assad
Yu
Meng
Sarah
Lai
Fang
Nikki
Alex
Theresa
Xu
Chen
Waqa
Joseph
Justin
Doval
Stephen
Swaraj
Skripal
Haley
Hsien
Wu
May
Juncker
Khan
Christopher
Sun
Jim
Cui
Marco
Gary
Bishop
Rawat
Wray
Duterte
Aquino
Mohammed
Loong
Sanders
Comey
Salman
Michael
Markus
Mohamad
Wengui
Aung
McCain
Jeremy
Song
Tao
Bipin
Jamal
Ching-te
Nixon
Cameron
Suu
Un
Uncle
Hu
Geoff
Cohn
Arab
Pohiva
Appleboum
McMaster
Orwell
Missal
Yong-ho
Shoigu
Gary
Mark
Yew
Saddam
H.R.
bin
He
Lu
Sam
III
S.

As expected, US (Trump) and China (Xi) are the most frequently mentioned name-entities within the Country category, both in English and in Chinese, followed by countries like Russia, India, Japan and North/South Korea. Leaders of these countries (and also Taiwan) make up most of the most frequently mentioned people.

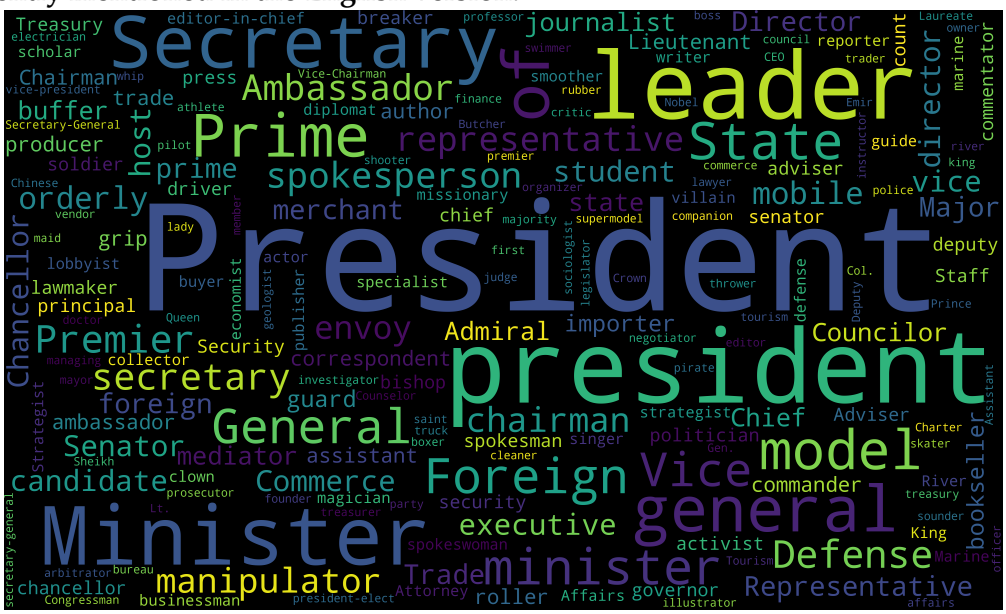


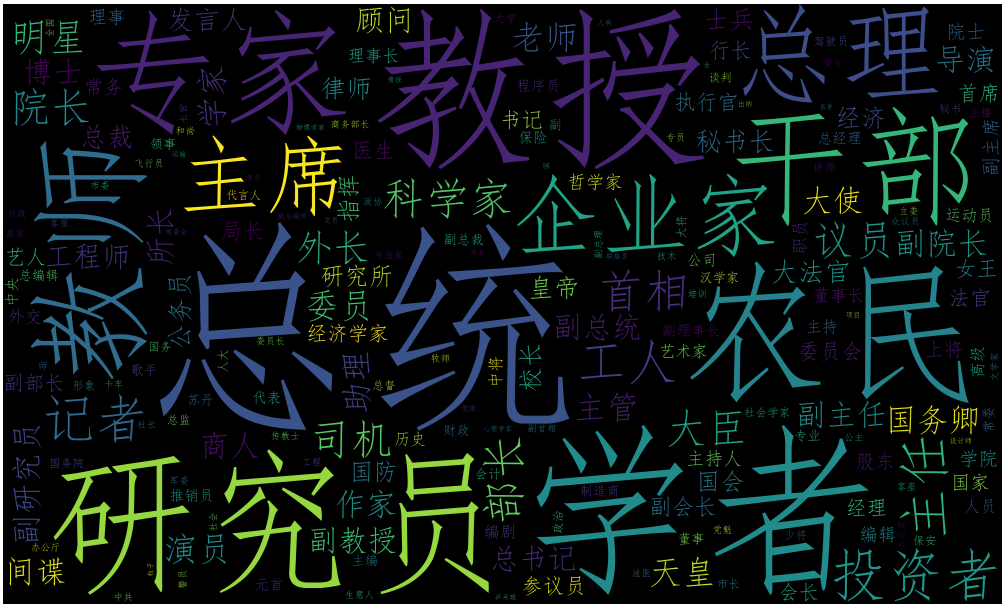
In terms of ideologies, both English and Chinese versions have frequent mentions of socialism and modernization. There are more mentions of nationalism, Marxism and communism in the Chinese version.



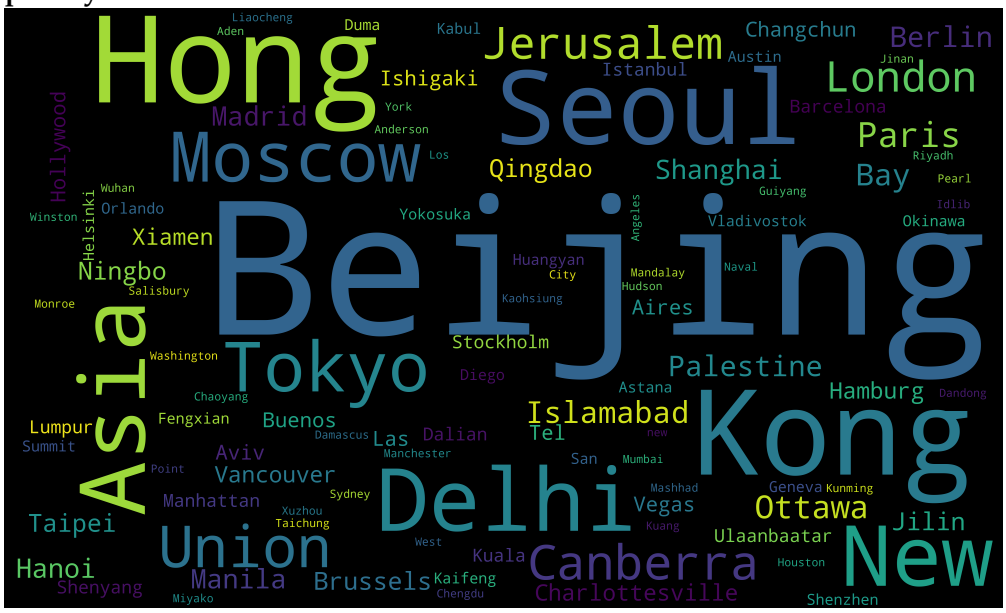


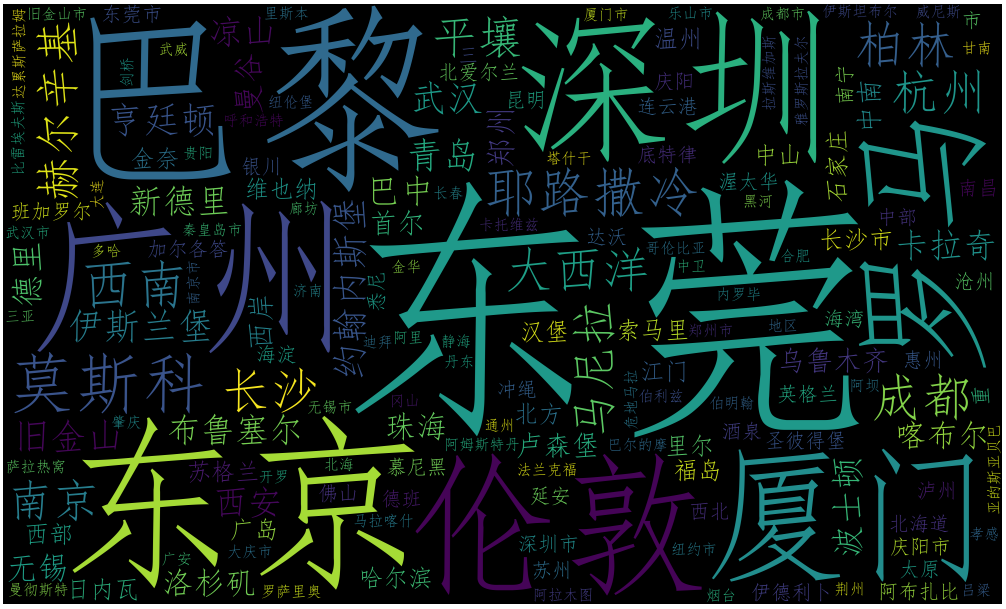
Mentions of organization are centered around international organizations like UN, EU, WTO, with the exception of the Communist Party of China (CPC), which is much more frequently mentioned in the English version.



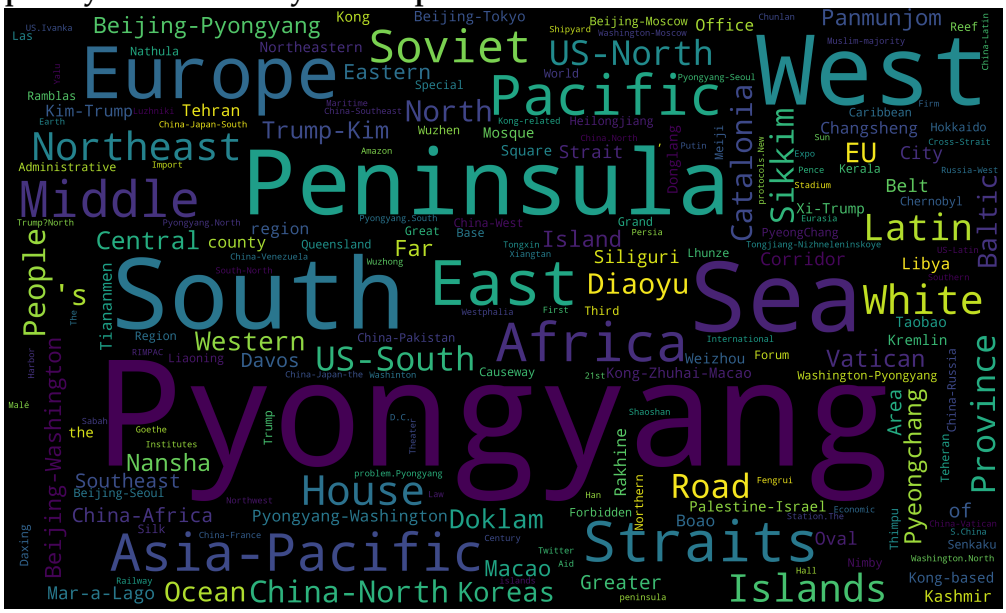


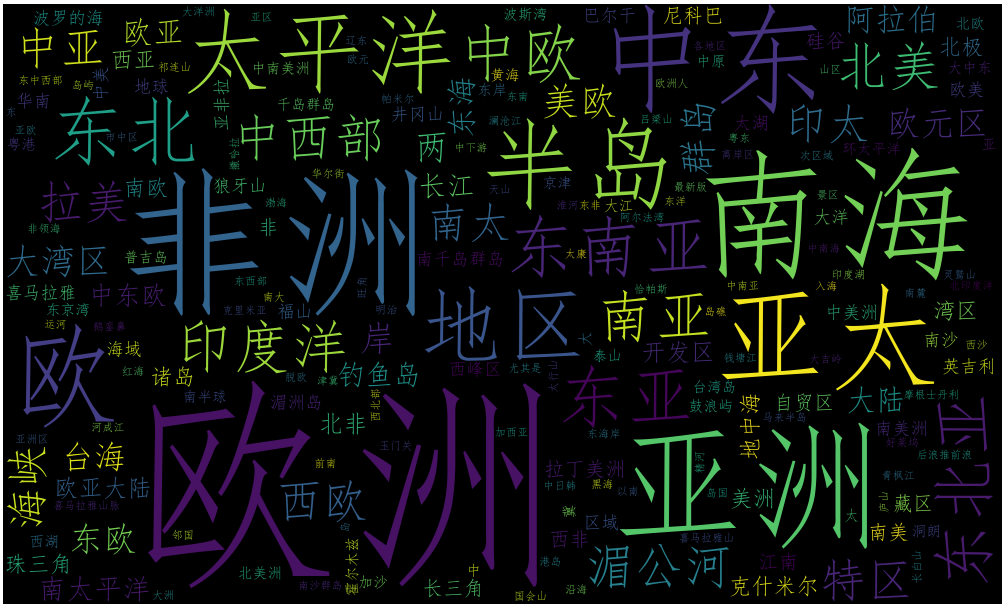
The most mentioned title in the English version is president, just the same as the Chinese version. However, most frequently mentioned titles in the English version are centered arond officials from other countries, while in the Chinese version, titles like farmers, workers and teachers are also frequently mentioned.





For the city category, cities attached with political importance like Beijing, Moscow and Delhi are very frequently mentioned in the English version. Washington DC is not very frequently mentioned, because it's recognized as a state or province. Domestic cities like Shenzhen, Guangzhou and Dongguan made the Chinese list. TThe frequent appearances of Dongguan, instead of Beijing, come as a surprise to me. There was a big "eradication" movement on the adult pornography industry in Dongguan, but the fact that it beats all other cities and became the one most frequently mentioned city still surprised me.

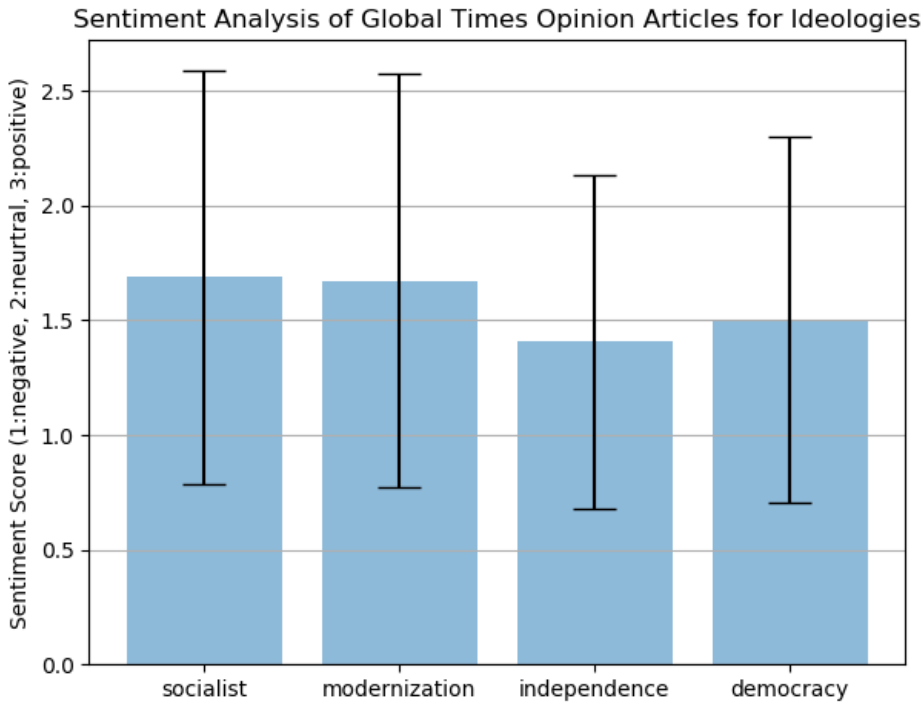




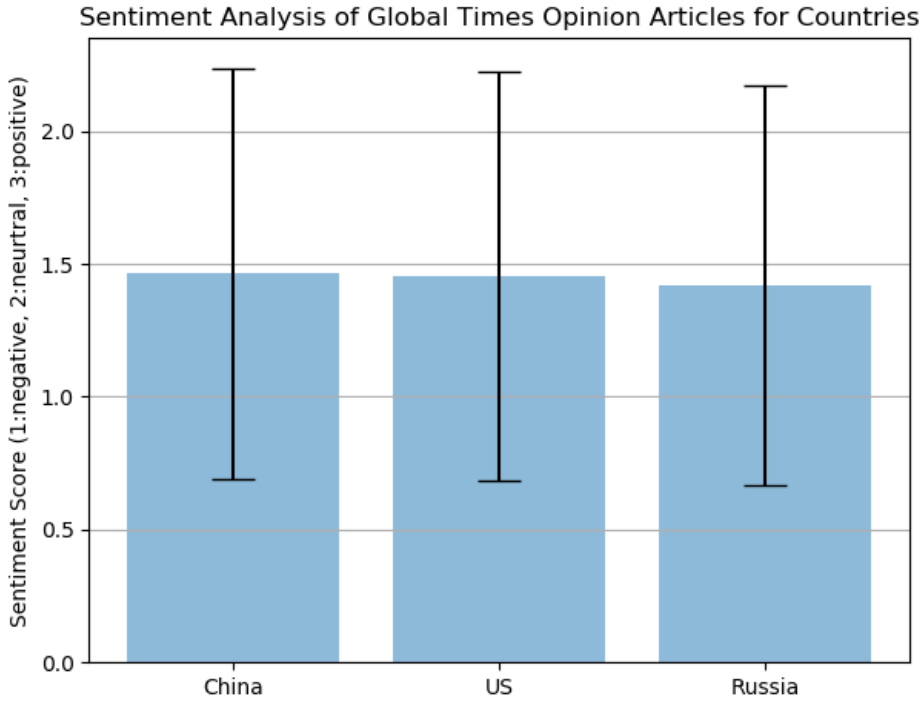
Location-wise, North Korea (Peninsula, Pyongang) is the center of attention for the English version, while EU, Africa, and South China Sea are the focus of the Chinese version.

Overall, the English and Chinese versions of Global Times are consistent in their focus on issues. Naturally, the Chinese version is more domestic-oriented, while the English version has more of an international focus. The China-US relationship, with a very high mention of Donald Trump, is of particular interest to Global Times.

Sentiment Analysis With these results, I wanted to go an extra step and perform sentiment analyses on certain key words within different categories. Since CoreNLP doesn't seem to have good sentiment analysis tools for Chinese, I am applying it on the English version only. Due to time contraint I am applying it on selective keywords within selective categories. See the below results for sentiment analysis on ideologies:



All ideologies, including socialism have their mean sentiment scores below neutral. However, upon investigation of standard deviation, socialism and modernization have better chances of being on the positive side than democracy and independence.



The sentiments of articles mentioning the three countries seem to have very similar narratives, countrary to my initial hypothesis that the narrative might be more hostile towards US. One potential explanation could be the co-existence of multiple country names within the same article, which could give

them similar scores even though within sentences they can have differential sentiment values.

0.0.3 Future Directions

First, due to time constraints I am not performing the sentiment analysis on the 10 most popular keywords within all categories (it takes me an hour to run on the ideologies alone, which have very few mentions overall in all of its keywords). Categories like titles, organizations and locations are all very interesting to further invetigate. Also, as mentioned above, using article as the unit for sentiment analysis might not be the best idea. One potential next step can be to use sentence as unit and calculate associated setiment values.

In addition, this project is largely descriptive. Another potential next step is to build a predictive/regression model on how the frequency of certain key words in an article can predict the article's overall sentiment.