Cross Culture Analysis Report

Andy (Shao-An Chien, sc4426@columbia.edu)

Introduction

Cross-culture analysis has been an active research among multiple fields of study, and we continue to focus mainly on analyzing the cultural differences based on news reports on the Internet. In this term of research, we applied a new idea and approach to find word outliers of news reports between different cultures. Moreover, in addition to the news videos, we broadened our work such as looking into the worldwide full-text news report on the Internet. The reason why we think it is important is because that text news reports contain more concrete news content than text descriptions of YouTube news videos.
In this term of research, we have some new results and thoughts that we did not expected such as finding outliers in between two different cultures, synonyms across different cultures. Our approaches are based on previous works and making some modification to automate the pipeline and use it to produce more results.

Previous Works

In last semester, we developed and set up a video scraper using python. The descriptions of videos and videos themselves can be downloaded by official YouTube API and open source library pytube (Figure 1). The pipeline included text preprocessing techniques such as tokenization, part-of-speech tagging, lemmatization, stopwords removal and name entities recognizer (Figure 2).

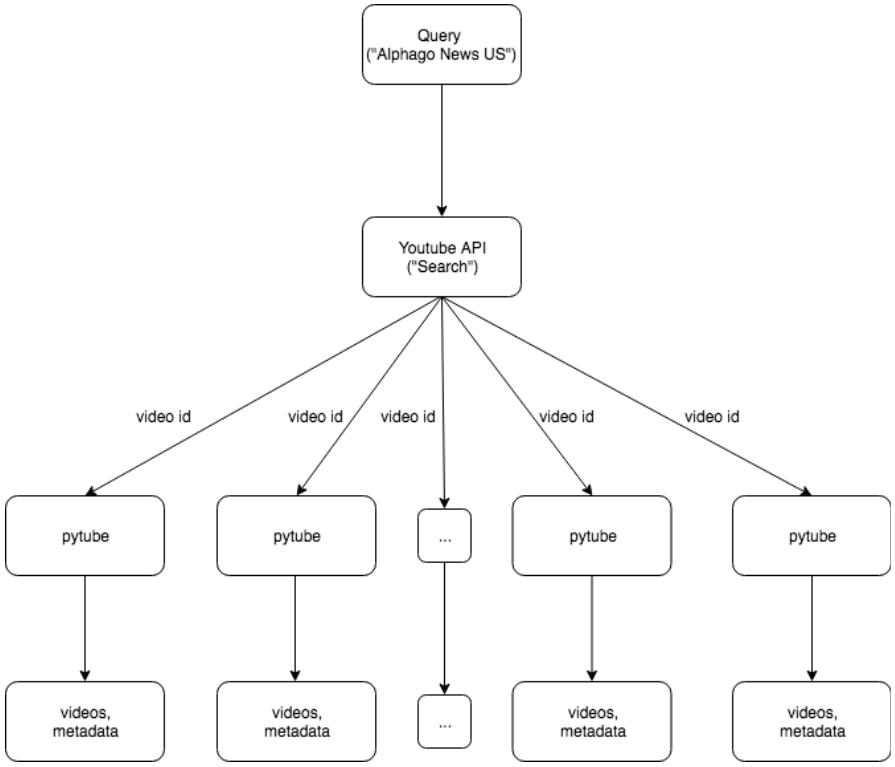

Figure 1

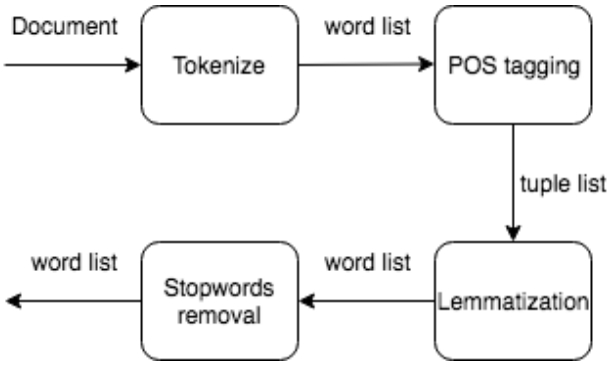

Figure 2

We used NLTK for text preprocessing and SpaCy for Named Entities Recognition, which are both being widely used in python for Natural Language Processing. One of our goals before was to find the differences between Chinese and American cultures by looking into the word frequency distribution. The words we considered "important" are those appears in the middle range since high-ranked words are

usually like stop words and low-ranked words are exactly unimportant just like their frequencies. We use the approach of log-log plot to have a better visualization so that we can easily capture the words in middle range. The other main goal was to look into the Named Entities to see if there exist some obvious differences of usage (which were all based on English but from different culture). We plotted a 2D graph to visualize the distribution from two different cultures. The points on the diagonal line were those words had similar frequency with respect to each document of Chinese and American news reports, so the points we were interested in were the outliers, which were not on the diagonal line but showed up in the upper left or bottom right corners.

YouTube Video Descriptions

Approach

For the analysis on YouTube Video Descriptions, the pipeline are basically the same as previous works. Here we introduce more news topics to see if we can find more outliers between two different cultures.

New Topics

Once we decided to use the approach of finding outliers, looking into variance of topics is an option to observe more interesting differences between Chinese and American cultures. Hence, we have come up with several new topics other than AlphaGo.

Thailand Cave Rescue

In June and July in 2018, there was a rescue team who rescued a junior soccer team from a cave (Tham Luang Cave) in Chiang Rai Province, Thailand. The reason we choose this topic is because it was a widely reported global news.

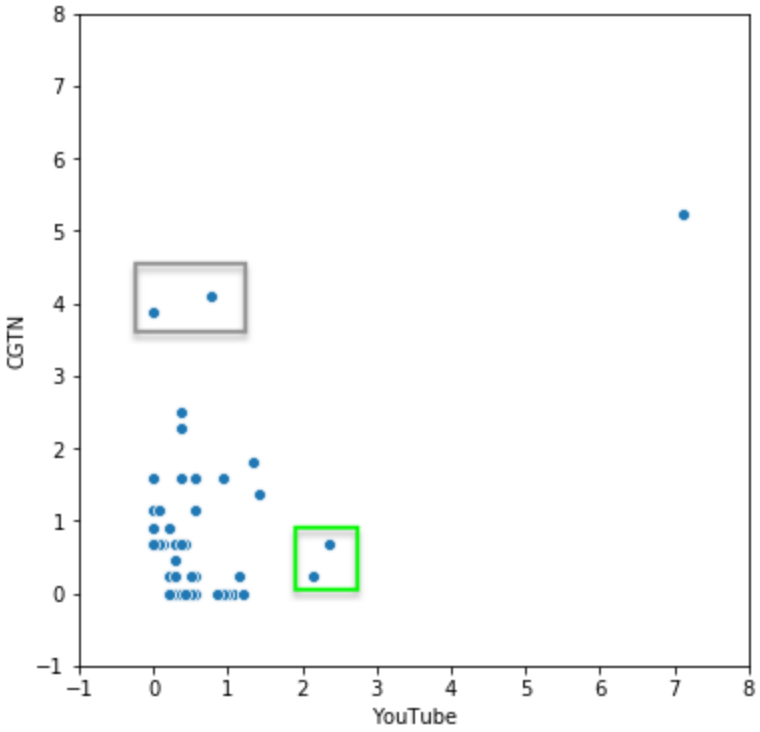

Figure 3

The outliers in the top-left (gray box) apparently account for more proportion in CGTN than in YouTube. These two words are ”two” and “Phuket”. The outliers in bottom-right (green box) are “Chiang Rai” and “Tham Luang”. The way we understand why there is a number “two” used significantly more often in CGTN (Chinese) than in YouTube (American) is that there might be some countries (or cultures) where prefer not to count the objects as the number of them goes up, while this is just a guess from our discussion and still to be verified. “Chiang Rai” is a big and well known city and province in Thailand which make sense that it showed up frequently in news reports, but, however, it is not that popular in CGTN. In the other hand, “Phuket”, which is a smaller and unpopular province is significantly popular in CGTN but not in YouTube. Base on intuition and these informations, we conclude that the countries near to the place where the incident happened are more familiar to it since it make sense that there might be more contact among them.

Lunar Rover

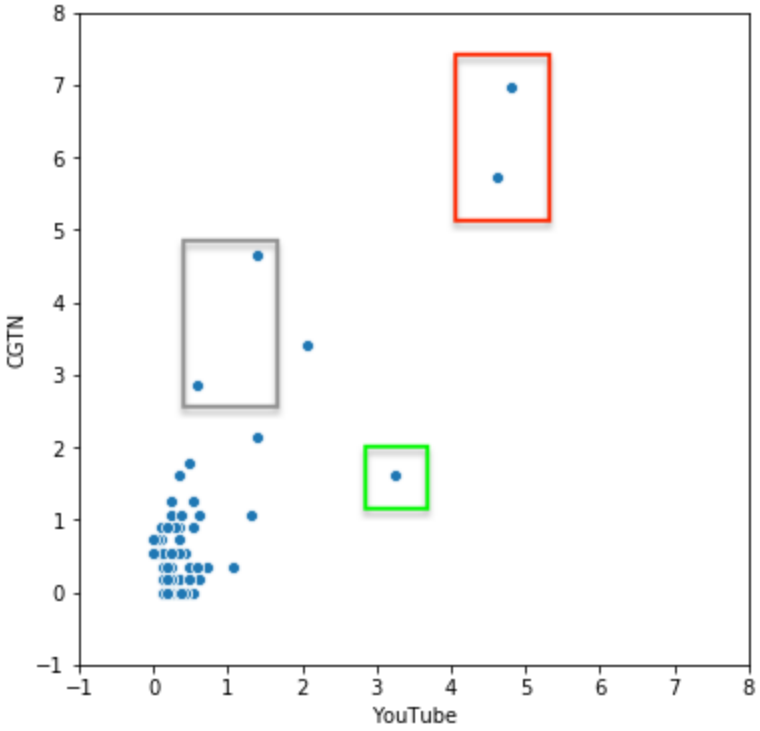

Figure 4

The outliers in the top-left corner (gray) are “Yutu” and “Change’s-4”, which are interesting because Yutu is the missions name and Change’s-4 is the name of the spacecraft built by China. And the outlier in the bottom-right corner (green) is “NASA”, on which we conclude that both cultures are reporting different content based on same topic. The outliers in the top-right corner are “first” and “moon”, and we can tell that there is a common point that there is some mission and made some achievement firstly related to the moon. Lunar Rover is the most interesting topic which exactly reflects the content of current events.

Florida Shooting

This is a shooting incident happened in Stoneman Douglas High School in Florida. 17 people including students were killed. We chose this news topic since there are

some Chinese-American involved (killed) in the incident and we wanted to see that if there were any biased reports among the two cultures.

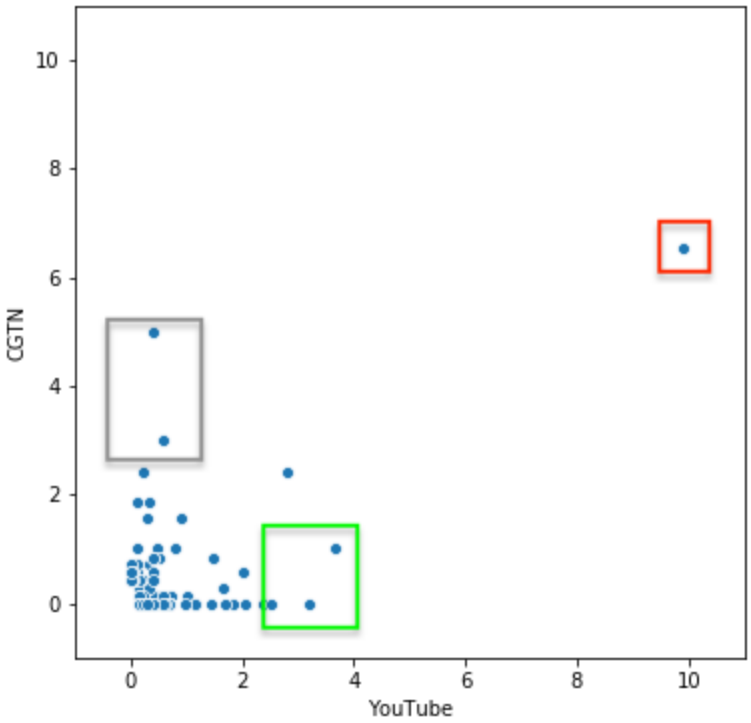

Figure 5

The outliers in the top-left (gray) are “US” and “two”. The strange number “two” showed up again here and we still wonder that why “two‘ is so popular in Chinese culture but not in American. We think the term “US” is more popular in CGTN is because that people in United States will not mention “United States” when stating an domestic incident, they will only mention the name of states or cities. The outliers in the bottom-right (green) are ”today”, “the next day” and “Parkland”. As we mentioned, Parkland is the name of the city where the high school located in. People and media in or near the country will state about the location more precisely. The high frequency word in the top-right corner (on the diagonal line) is “Florida”, which is not surprising at all. But we can see that the point is a little lower than the actual diagonal line, which means it is used more frequently in American culture and is consistent to the previous assumption.

Full-Text News Reports

As we are analyzing the text part of the news report, the data generated from full-text news reports are intuitively more powerful than from the descriptions of news videos on YouTube. There are two main problem from videos descriptions. First, the content is usually too short and brief since they want people to focus on the video but not the descriptions. Second, there are too many advertisements in the video descriptions such as “Download our App on App Store”, “Download our App on Google Play”, …, etc. Hence, we modify some part of the pipeline in this term of research, detail we be discussed below.

Approach

Rather than focusing on YouTube, we target on the news websites themselves such as Washington Post, CNN, BBC for American media. There are also news media based in China but the content are all written in professional English such as People’s Daily, China Daily.

Bing News Search (did not work out)

At first, we were trying to get news using general news apis. Bing News Search is a strong api for news searching, however, we were not available to get enough data for “News written in English by China media”. Moreover, Bing News Search started charging money after one week trial.

News API (did not work out)

After Bing News Search started charging, we turned out use News API. It is good at getting trending news and topics, however, we were trying to get the news in the past which were once trending then and we were also unable to get “News written in English by China media”.

Scrape Directly From News Websites

To avoid above problems, we finally turned into the web scraping techniques. The websites we scraped for this research are Washington Post and China Daily. We

looked into how they call the api to there backend every I click search on their websites. To achieve this, we use debug tools by Chrome:

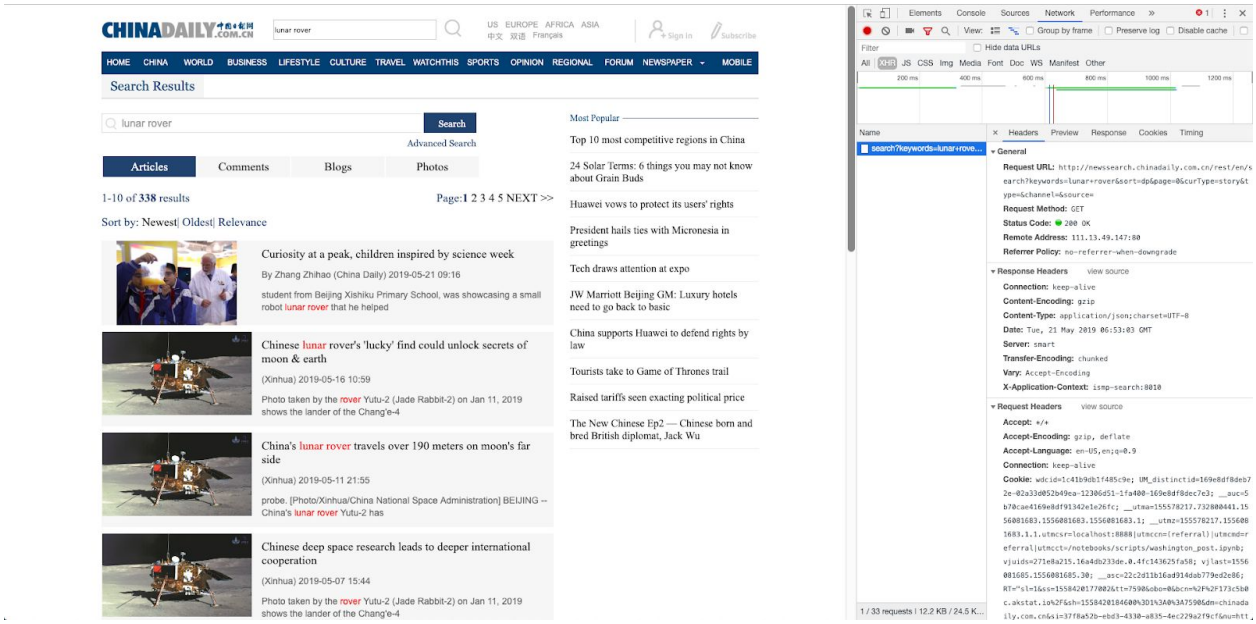

Figure 6

Usually the response of api will contain list of news “previews”. Each preview contains thumbnail, title, url, …, ets. Then we store the url and use it as an input of another python package “newspaper”.

newspaper

“newspaper” is a python open-source library, and is widely used to do the full-text, and article metadata extraction. We input the urls obtained in previous progress (scraping), “newspaper” will return the metadata of such article including the whole content of the news report. Then we can finally store the results and use them in the further outliers analysis (Figure 7).

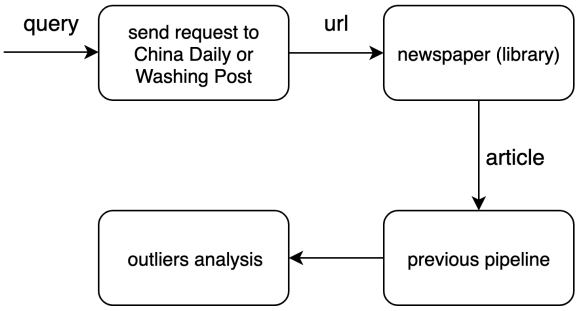

Figure 7

Topics

Thailand Cave Rescue

In June and July in 2018, there was a rescue team who rescued a junior soccer team from a cave (Tham Luang Cave) in Chiang Rai Province, Thailand. The reason we choose this topic is because it was a widely reported global news.

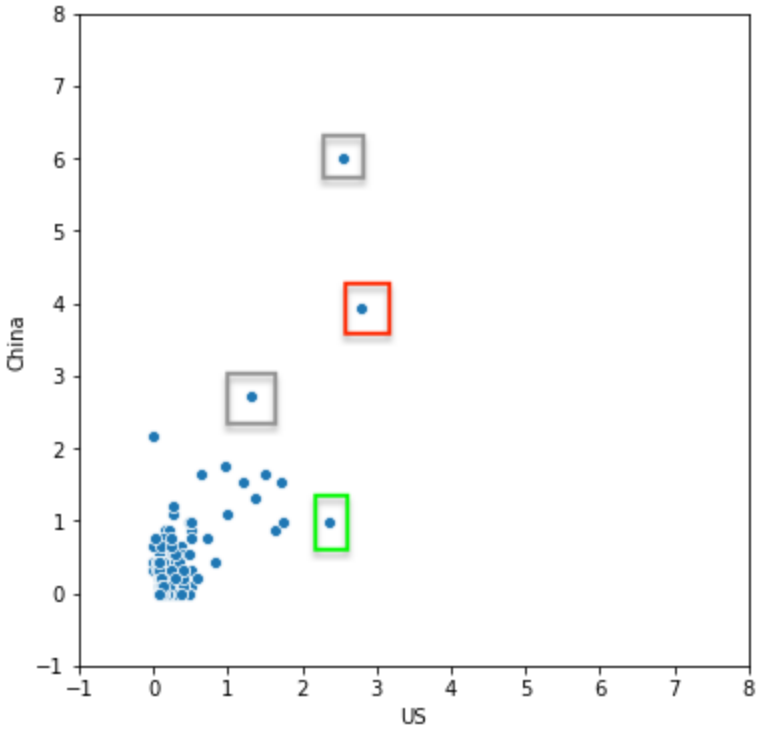

Figure 8

The outliers in the top-left corner (grays) are “Monday” and “Chiang Rai”. Our understanding for the reason why there is day of the week showing up is there is time difference between two cultures. The outlier in the bottom-right corner (green) is “first”. This is also interesting since we did not know why people tend to use the term “first” in American culture. When we looked into the document how they use “first” in sentences, the distribution looked like random and normal such that it was used to describe variety of instances. The word which is popular in both culture in the red box is “12”, which is the number of the days that the rescue process took.

Trump Trade War

Since this is an political sensitive issue, especially in between the two cultures we are discussing, we wanted to give it a try and see what is in common and difference.

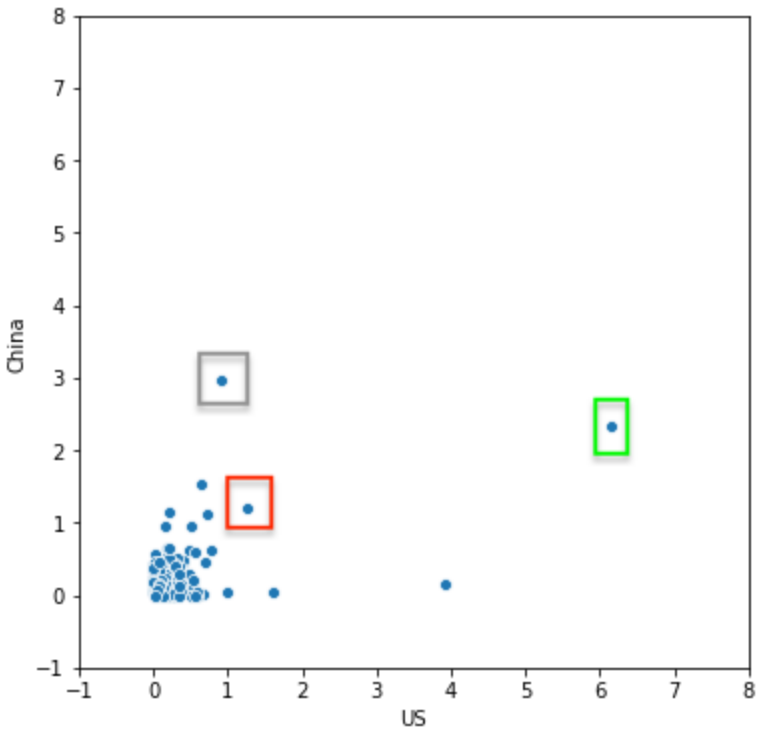

Figure 9

The outlier in the top-left corner (gray) is “two”. Again, the word “two” in Chinese culture. However, rather than leaving the truth why they like the word “two” in air, we looked into the documents to see how they is this word. “two leaders”, “two sides”, “two countries” dominate the majority of the usage of “two”. So, the media in Chinese culture tends to combine two sides and give opinion or report together, while the media in American culture tends to separate them. This is an interesting conclusion that we did not know before, however, it still has to be further and carefully verified.
The outlier in the bottom-right corner (green) is “Trump”. Apparently people in United States mentioned there president more often.

Lunar Rover

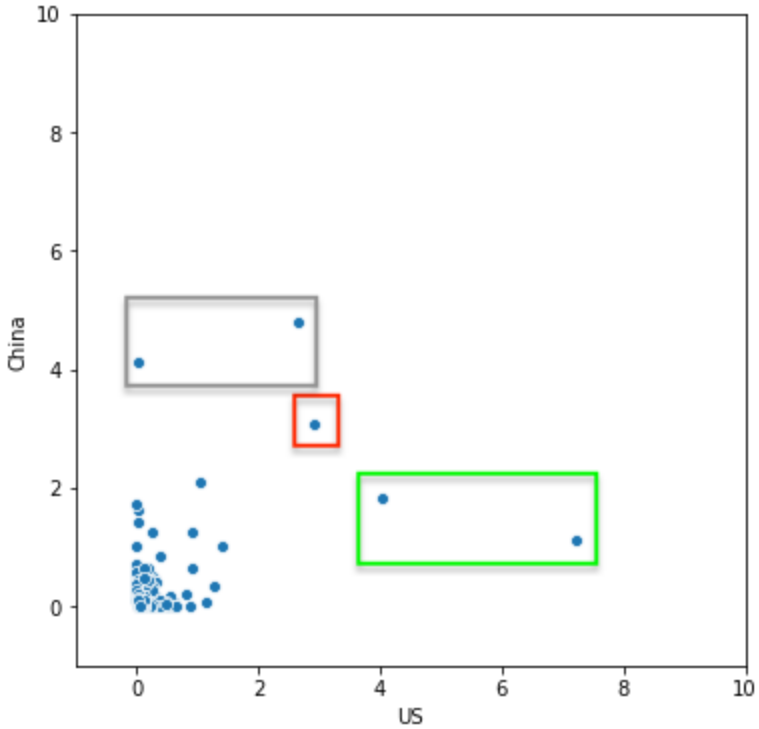

Figure 10

The outliers in the top-left corner (gray) are “two” and “us”. Once again, “two” showed up in one of the outliers. However, unlike the previous that using “two” to combine two things or sides together and comparing to each other, “two” was

randomly and simply used to say two any kind of things such as “two components”, “two years”, “two pixels”, “two systems”, …, etc. This observation is like the more previous ones which China media tend to write “two” things. Our conclusion (guess) is that people might not count the number intuitively in some cultures once the number goes large although people in other cultures do not share the same perspective.

The outliers in the bottom-right (green) are “Mars” and “NASA”. We looked back into the documents and observed that “Mars” usually appeared together with “NASA”, saying that there is some mission in progress or is being planned to rove around the Mars. It is not surprising when reporting an achievement in other country, the media from United States would relate something about the issues in their own country.

Notre Date Fire

On 15 April 2019, a tragical fire which severely harmed the structure of Notre Dame Cathedral in Paris.

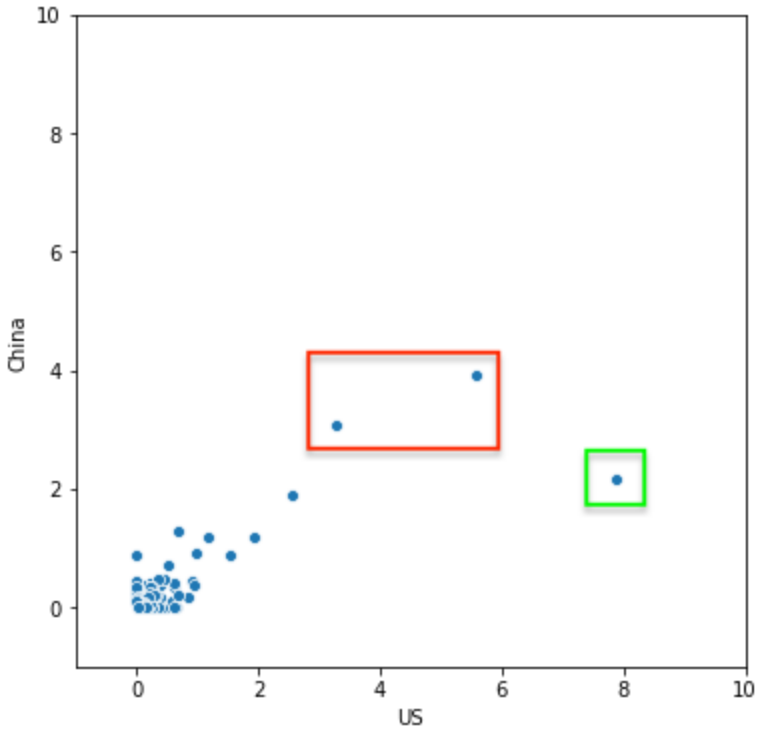

Figure 11

There is no outlier in the top-left corner, which means there is no word popular in Chinese culture but not in American culture. The outlier in the bottom-right corner (green) is “Notre Dame”, and we think this can be ignored since there are multiple ways to represent “Notre Dame” such as “Notre-Dame”, “Notre-Dame Cathedral”, …, etc. Our Named Entities Recognizer was unable to map them to the same entity. The conclusion is there is no outlier in this topic, which means both cultures were using the same ways and terms to report.

Chinese Scientist Gene-Edited Babies

A Chinese professor He Jiankui announced he successfully edited the gene of a baby, which cause extreme condemnation on what he did.

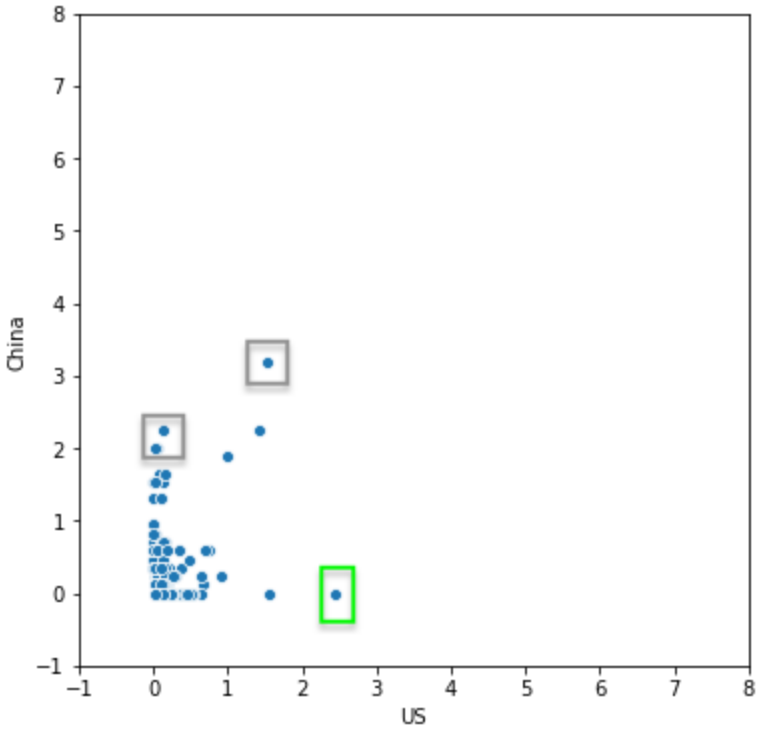

Figure 12

The outliers in the top-left corner (grays) are “first”, “Shenzhen”, “Monday”, “Guangdong”. “Shenzhen” and “Guangdong” are both names of cities in China. It is not surprising that media in China will report the location more detail than in

United States. And here we see “Monday” again, which is a day of the week again in the outliers in Chinese culture. It started raising our attention that why Chinese tend to mention “Monday” since this is not the first topic “Monday” had showed up. However, we think we still need more data to make the conclusion on this outlier. The one in the bottom-right (green) is “Trump”, and it is also not surprising to showed up in the United States culture.

Synonyms Across Cultures

When looking into the documents of the topic Trump Trade War, we found an interesting fact that Chinese media use “DPRK” (Democratic People's Republic of Korea) rather than North Korea. The difference might due to difference of political background. Then we started a research branch to find if there there are other synonyms, and some of them are used in one culture but not used in another.
Our approach to find such synonyms are searching on the Internet if there are different usage of words across British and United States, then query the words in the Chinese news website such as China Daily to see which one has more results. Below are some results we have found:

Chinese	**United States**
railway	railroad
registration plate	licence plate
share	stock
carriage (train)	car (train)
wardrobe	closet

Conclusion

Future works

“two”

The word ”two” consistently showed in the outliers in the top-left corner. There are two observation in this term of research. First, when reporting an international issue between Chinese and United States, Chinese media tends to report in the way of combining two sides and mention what they did. Second, they use the number of “two” frequently and randomly to report the amount, while we did not see other numbers more than two in the outliers.
Hence, we suggest to do further research on whether there is a cultural difference such that people in Chinese tend to use “two” more frequently than people in United States.

More news topics

As we have developed the pipeline, we can easily get and visualize the data and analyze than like the works above. Once we came up with more and more news topics, the automation of this process will take an important role in analyzing the culture differences.

Disclaimer

There are multiple topics we discussed above might involve in political and sensitive issues. The conclusions we made are all based on the perspective of science, rather than having personal perspective of politics.