Data301 Project Final Report

Li Jiaqi 75766881

Abstract / Summary

The project aims to discover the changes in attitudes towards China in various types of events before and after the coronavirus pandemic. The Global Database of Events, Language, and Tone data from 1st January 2019 to 31st December 2020 and related to the United States, United Kingdom, Australia, Russia, North Korea, Pakistan against China. How similar was the attitude in various types of events from these countries to China in 2019 and 2020? The Cosine Similarity Algorithm will be used to Analyse the similarity of the differences between the attitudes in various types of events from these countries to China in 2019 and 2020. Because of the differences of international camps, the negative response to China of western developed countries headed by the United States should have an obvious increase in 2020 compared with developing countries which have good relationships with China such as Russia, North Korea and Pakistan. Therefore, the similarity of the attitudes of the United States, the United Kingdom, and Australia towards China in 2019 and 2020 should be significantly lower than Russia, North Korea and Pakistan. The results will show whether there will be major changes in relations between countries from the same or different camps during a special period.

Introduction

In the GDELT Event file, every event record has a Actor1CountryCode and Actor2CountryCode which are 3-character CAMEO codes for countries belonging to two actors of the events. This can filter the events which are relevant to this topic. There is also an EventCode in the GDELT file. It is a raw CAMEO action code describing the action that Actor1 performed upon Actor2. This can be used to distinguish the attitude of the actions between these countries and China is neutral, positive or negative. Both actors in an event have 3 ActorXTypeXCode CAMEO code attributes, these describe the type of the actor is. If an event has several different ActorXTypeXCode, it can be regarded as several different types of events.

The GDELT file covers events that occur every day of the year. The numbers of events in a year in different attitudes in various types of events can be counted to vectors and apply the Cosine similarity algorithm to analysing.

How similar was the attitude in various types of events from the United States, United Kingdom, Australia, Russia, North Korea and Pakistan to China in 2019 and 2020? The events in the data set which are participated by 2 countries in various types of events with an event description will be used to figure out this research question. To apply the cosine similarity algorithm on the relevant data will produce a number result to show the accurate similarity.

Experimental Design and Methods

The cosine similarity algorithm is used to calculate the similarity between 2 vectors. The result interval of this algorithm is between 0 and 1. The 2 vectors are totally the same when the result is 1. When the result is closer to 1, the two vectors are more similar. The attitude of these countries to China in various types were categorised into positive, neutral, negative and unknown.

On the first step of this project, the pyspark, java virtual machine and gdelt are setted up on the google colab. Then fetch the gdelt events table of 2019 and 2020 to the environment. Next, the downloaded csv event table files are translated to RDD by the pyspark function. The different CAMEO event codes are categorised into 4 different attitude sets based on the CAMEO event Codes description document. Then, these 6 countries and Types of actor CAMEO codes are stored into two lists. In order to reorganize the RDD into a structure that can be intuitively operated, a main function and few help functions are created. The first help function will be entered with the CAMEO code of an event and return a string of attitude level. The second help function will be input to a list containing ACTOR role types. If there is only one type in the list, it will return a string of this type. If there are multiple types, it will return a list containing multiple types. In the rest of the cases, it will return as an Unknown type. In the main RDD processing function. The first step is to use the map function to turn each row in the rdd structure into a tuple. The left side of the tuple is the CAMEO code of 2 countries and these country's attitude between each other in various fields in the event, and the type of event. The right side is the integer 1. Then use the filter function to filter out the events of the analyzed countries and China. In the next step, Divide rdd into two, one of which contains only one event type, and the other contains multiple event types. In the rdd that contains multiple event types, use the flatmap function to divide an event that contains multiple event types into multiple events. Then use the union function to combine these two rdds. Furthermore, Use the reducebykey function to get the rdd that contains the counts of various countries' attitudes toward China in different event types. In the next step, An rdd containing the attitudes of various countries towards China in various events with a count of 0 is created. Finally in this function, combine these two rdds and return it. This function is used to transfer the data of 2019 and 2020 into a structure that can be intuitively operated. After this, use a series filter function to get the data by each country. Then use map function to transfer these data to 6 vectors and apply the cosine similarity to get the result. In order to make the results look clearer, pandas library and matplot library are used to output the graph of the results at the end.

```
isNone(array)
```

Determine how many event types are in the input list, if there is only one event type, return this event type, if there are more than one, return a list containing these event types, otherwise return “*unknown*”.

```
get_need_rdd(data)
```

Convert the rdd that stores the data to the rdd that is convenient and intuitive to operate.

```
rdd_2_vector(rdd)
```

Get a vector of counts of various countries' attitudes towards China in various events from rdd

```
total_events(rdd)
```

Get the counts of events of various countries towards China in a year.

```
cosine_similarity(list_a,list_b)
```

Calculate the cosine similarity from two vectors.

```
from math import sqrt
```

Import this module to calculate the square root of a number.

```
import pandas
```

A data analysis library.

```
import matplotlib.pyplot as plt
```

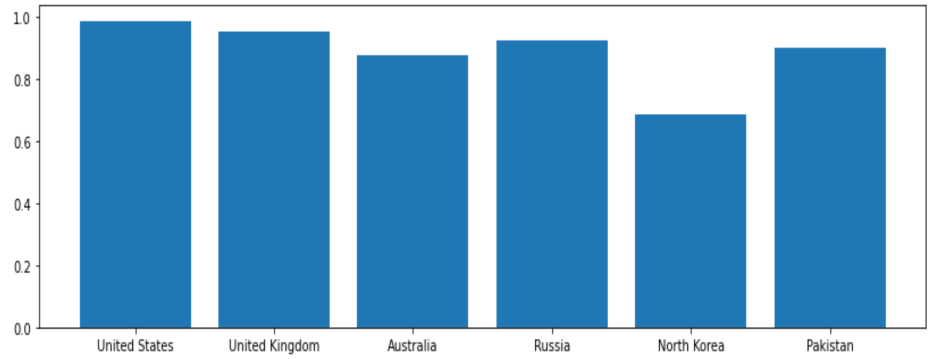
Draw charts.

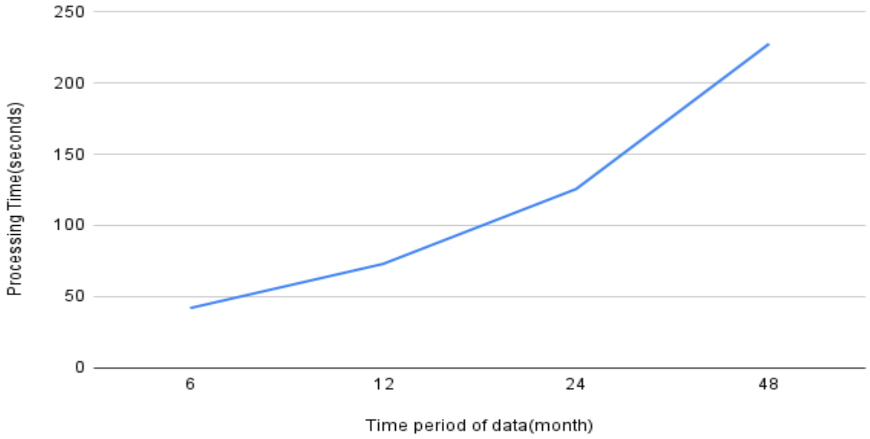
```
import time
```

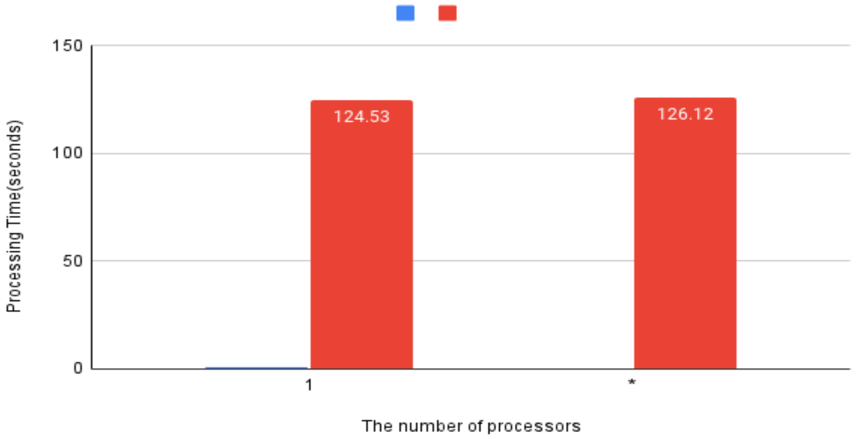
Time libarary

Results

```
               cosine similarity  2019total  2020total
United States           0.986352       1184       1214
United Kingdom          0.952339        166        209
Australia               0.877515        110        289
Russia                  0.922113        171        137
North Korea             0.686819         57         44
Pakistan                0.899390         77         64
```







Unexpectedly, the cosine similarity of the United States's attitudes towards China in various types of events in 2019 and 2020 is the highest at 0.986. Followed by the United Kingdom, Russia, Pakistan, Australia and North Korea with 0.952, 0.922, 0.899, 0.878 and 0.687 respectively. With the exception of North Korea, China's closest ally, other countries' attitudes towards China in all aspects of events in 2019 and 2020 are highly similar.

Conclusion

Yes, able to answer. The cosine similarity of six countries' attitudes towards China in various types of events in 2019 and 2020 clearly shows their similarity. The closer the cosine similarity is to 1, the more similar the two sets of vectors, and the more similar the two sets of vectors, the more similar the two sets of data.

The results of the project show that before and after a huge event, the change in the country's attitude towards the country under various types of events is not directly related to the country's international camp. For example, in this project analysis, Britain, Australia are in one camp, and Russia and Pakistan are in another camp. However, their attitudes towards China in all aspects in 2019 and 2020 are very similar.

Because the size of each country is different, in subsequent analysis, international alliances such as EU member states and Arab League member states can be regarded as a country compared with a superpower such as the United States. This can increase the equivalence of data scale and Increase the persuasiveness of data analysis results.

Critique of Design and Project

Finding the frequency of keywords in the event article to determine the type of event seems to be more accurate than directly using the role type of the event member. But if this method is used, it seems that the program will run very slowly on the two-year data.

In the initial design, it was originally intended to compare the overall attitudes of various countries towards China in 2019 and 2020. But this seems too simple and cannot reflect the changes in different types of events. However, it would seem better to remove all events that did not appear in the events between the analyzed country and China in the comparison.

Reflection

Concepts: Parallelism, Distributed Data, Map Reduce, Similarity.
Tools: Google Colab, Pyspark, pandas, matplot GDELT DATABASE.

In this project, I learned how to apply the algorithms learned in this course to such a larger project. And intuitively feel the efficiency of Parallelism. I also learned to use the GDELT database. More importantly, this is the first time I have asked a research question myself and analyzed it using the knowledge I have learned in the course. I believe this experience will provide me with invaluable help in my future work.

References

https://linwoodc3.github.io/gdeltPyR/
http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf
https://www.gdeltproject.org/data/lookups/CAMEO.eventcodes.txt
https://www.jianshu.com/p/e912987e1e64