**What is the impact of the epidemic on Chinese society**

11969443 Zhenyuan He

***Abstract***

The research question: **What is the impact of the epidemic on Chinese society.** The research question for this project was a market basket analysis of themes in China at different times, based on themes from Gdelt's Global Knowledge Graph 2.0. The expected result was that Chinese society would focus more on health and epidemic themes after the onset of the epidemic. The results of this study help to understand the changing themes of concern in Chinese society before and after the beginning of the epidemic and after it has lasted for some time. It can also help to study the impact of the epidemic on Chinese society and to study the changing trends of the epidemic.

***Introduction***

***Background:***

Research data background: The research data is from the Global Knowledge Graph 2.0 in the Gdelt database, which contains world news and articles in 65 languages, with each row having variables such as the link to the event, the location where it occurred, the article theme, and the article sentiment. This project extracts the Location and V2Themes fields from the GKG Table, which represent the geographic location of the event and the article's theme, respectively. After downloading the data for the three periods, the data were filtered according to CountryCode, and the rows with CountryCode of CH (China) were filtered out, and then the three data sets were analyzed separately.

***Algorithm background***

1. Market basket analysis is an analysis method to correlate different commodities and explore the connection between them by studying user consumption data, called Market basket analysis. Talking about Market basket analysis, we have to mention the classic case: beer and diapers: in fact, this is a very classic case in marketing. After investigation, it was found that when men came to buy diapers for their children, they would buy a few more bottles of beer for themselves. The manager later placed the beer and diapers together and significantly increased the sales of both beer and diapers. The purpose of Market basket analysis is to find out some potential patterns of users' buying habits by mining and analyzing their consumption record data so that they can provide users with the matching packages they want. The increase in package sales will generally lead to an increase in customer unit price, thus increasing the company's revenue.

2. Association rule mining allows us to discover item-to-item relationships from a dataset. The form of an association rule is I → j, where I is a set of items and j is an item. The association rule studies the effect of the set I on J. Analyzing the support, confidence, and interest of an association rule, one can learn more about a particular association rule and thus discover meaningful association rules.

3. Apriori algorithm: Apriori algorithm is the process of finding frequent itemset. The frequent itemset is an itemset whose support(number of occurrences) is greater than or equal to the Min Support threshold. And the itemset whose support is less than the minimum is called the infrequent itemset. It is a layer-by-layer iterative approach, first finding the frequent1 itemset L1, then using L1 to find the frequent2 itemset, and so on ...... Due to the consideration of the running time, this project only ran the Apriori algorithm once.

4. On the way to applying the algorithms, this project makes significant use of the knowledge taught in the Data301 course. The project will use Spark, an open-source cluster computing framework, to store data using resilient distributed data storage (RDD) and to compute data using Spark's operators such as map/reduce. After the basic framework is built, the project will upload the program and data to the Google Could Platform for multi-core operation.

***Motivation***

From the beginning of the COVID-19 started, people's lifestyles were greatly affected and restricted. Since I am from China, researching this issue is a great way for me to learn more about the themes discussed passionately in Chinese society at different times. I came to New Zealand to study in 2019 and have not been home once due to the COVID-19, which prompted me to learn about some of the challenges and social themes that China faces through this project. Since I lived in the age of the epidemic and could personally feel the impact of the epidemic on my life, the data was relevant to me. In Lab3 of DATA301, I was asked to complete a market basket analysis of user purchase data, and that lab allowed me to understand the steps of market basket analysis, and this analysis method was applied to this project. By studying this theme, through my research, I was able to introduce people around me to the changes that Chinese society has seen over the years on a theme that they are passionate about. This is meant to make my friends or international students around me more aware of the changes in Chinese society.

***Question***: **What is the impact of the epidemic on Chinese society.**

The solution to the problem was for the project to study the themes covered by Chinese social media through three time periods.What are the differences in the themes of Chinese social discussion, before the COVID-19 starts, just after the COVID-19 starts, and after the COVID-19 lasts for some time? Hypothesis: Social themes in China were more related to taxation and politics before the start of the epidemic, and more related to the epidemic, health and population after the start of COVID-19. The data used in this project are from the Global Knowledge Graph 2.0(GKG) of the Gdelt project, which extracts the Location field and the V2Themes field from the Graph, extracts the CountryCode from the Location field, and then extracts the Country with CH (China) as the study set. The location field is extracted from the CountryCode and the country with CH (China) is extracted as the research data set, so the data set is selected to be relevant to the research question.

The project applies market basket analysis to the dataset. Market basket analysis is performed on the set of Baskets, while each row of the GKG Table is an integration of information from the articles. The V2Themes field contains the themes of the articles (more than one and much less than the total number of baskets), so the program use the themes of each article as a Basket. Filter the CountryCode for China baskets and perform market basket analysis on them. The steps to apply the market basket analysis are: The first step is to find the set of frequent items using the Apriori algorithm. The project has the function Frequent\_Itemsets(Country\_And\_Theme, countryCode), which inputs the data set and the country code that needs to be found, and outputs a set of frequent items pairs containing two themes and the number of occurrences, a set containing only a single theme and the number of occurrences of the theme (individual) and the length of the data set.

The second step of the project uses the function support(Frequent\_pairs, n) to calculate the support of the Frequent\_pairs data set. confidence(individual\_pairs, Frequent\_Themes, count\_of\_total\_pairs) to calculate the confidence of the Frequent\_Themes(same as Frequent\_pairs) data set. interest\_rate(individual, confidence) calculates the interest of the Frequent\_pairs dataset.

After the above calculations, the program produces the results and analyzes them. By ranking support, confidence, and interest, the themes with significant values are identified for analysis.

Support is the probability of both A and B themes appearing together, which represents the reliability of the combination. Confidence is the conditional probability of theme A appearing followed by theme B, which is the percentage of articles with A and B themes among all articles containing theme A. A high absolute value of interest and a positive number indicates that both themes are interesting and have value for research.

***Experimental Design and Methods***

***detail of the algorithm***

Step 0: Set up the library to use Spark and set up the dbg() function for debugging.

Step 1: Firstly, to prevent redundant files in the downloads section, the project first deletes all CSV files in the downloads section. Then download the Global Knowledge Graph 2.0（GKG table） from the Gdelt database for February 2019, 2020 and May 2021 (one CSV file for each day of data). The program first constructs a list of dates to be downloaded and then iterates through the list of dates to try to download. If the download succeeds, the date is returned. Else, None is returned as a marker. Dates or None are stored in Before\_Epidemics, Start\_Epidemics, and Still\_Epidemics.

Step 2: Extract the dates of successful downloads, in preparation for reading the downloaded data into the RDD. The specific operation is to remove None from Before\_Epidemics, Start\_Epidemics and Still\_Epidemics to get all the successful download dates, read them into sqlContext and then convert them into RDDs for storage.

Step 3: In the RDD stored in the GKG table, extract the CountryCode field from the Location and convert the V2Themes field into a List.

Step 4: Construct the Frequent\_Itemsets(Country\_And\_Theme, countryCode) function to implement the Apriori algorithm to find frequent itemsets. First set 50 to Min Support, then filter out the rows with a CounrtyCode of CH (for China) as the object of study. Then the flatMap operator is used to find the frequent itemsets containing a single theme and broadcast them to each Working Node, which can speed up the program. Afterwards, each basket is iterated through to find pairs, where both elements of the pair must be in the frequent itemset. The result is integrated using reduceByKey to obtain the set of pairs with occurrences where both elements are in the frequent itemset.

Step 5: After using the Apriori algorithm to find the set of frequent items, the project can begin to calculate the support, confidence and interest of the association rule. Support measures the probability of two frequent themes occurring together, and the program operates the support function to calculate the support value. support( The frequent\_pairs, n) function calculates the ratio of the number of occurrences of each pair to the size of the total data set by entering the frequent pairs and n, which represents the size of the corresponding data set and using the map operator to obtain support.

Step 6: In the association rule A->B, confidence is the conditional probability that theme A occurs when theme B occurs and is the percentage of articles in which theme A and B occur at the same time out of all articles that include theme A. The confidence function is used to calculate the confidence value. confidence(individual\_pairs, Frequent\_Themes,count\_of\_total\_pairs) function by entering individual\_pairs, which represents the number of occurrences of a single theme. Frequent\_Themes represents the number of times two frequent themes occur together, which can be interpreted as the number of times themes A and B occur together. The project takes the two themes of Frequent\_Themes as Key and constructs them as A->B and B->A. They both represent the probability of A and B occurring at the same time but in different forms. Afterwards, the probabilities of A->B, B->A and individual themes are combined into a large RDD using the union method, and then the confidence of each set of A->B and B->A is calculated using the Map operator.

Step 7: The interest of an association rule A->B is defined as the difference between the confidence of A->B and the probability of occurrence of B in the whole dataset, which reflects whether the association rule is interesting or not. The project uses interest\_rate(individual, confidence, n) to calculate the interest. The individual represents the number of occurrences of a single theme, the probability of a single theme occurring is first calculated, and then the map operator is used to calculate the interest of all association rules. This is done by merging the RDD containing the probabilities of individual themes with the confidence dataset from step 6 using the join function. Afterwards, the frequency of occurrence of the corresponding theme is subtracted from the confidence of the association rule to obtain the interest.

Step 8: Analyse the results of the data

***Functions used***

1. The functions intofile(filename), get\_filename(x) are from the Sample GDELT project available on the learn page. They are functions for downloading files and the project has modified the intofile function a little so that it returns the date of successful download and the download of Global Knowledge Graph 2.0.
2. find\_country(string), and to\_list(string) are used to process the dataset that has been read into the RDD and taken out of the Locations field with the V2Themes field. The find\_country function raises the CountryCode from the Locations field and the to\_list function converts the V2Themes field into a list.

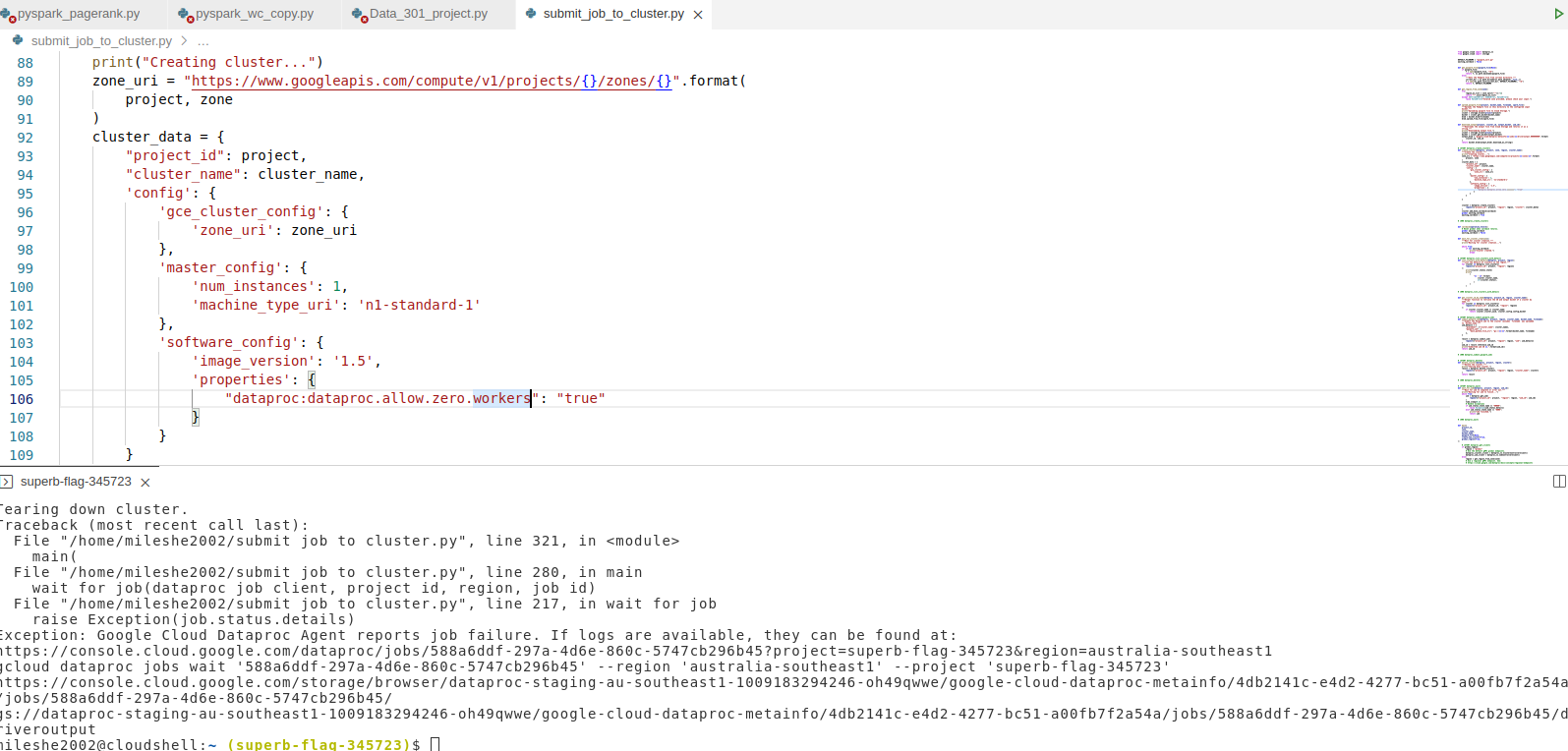
## The Frequent\_Itemsets(Country\_And\_Theme, countryCode) function inputs the name of the country and the data set to be searched for, runs the Apriori algorithm to find the frequent itemsets and returns them.

1. support(Frequent\_pairs , n)， confidence(individual\_pairs, Frequent\_Themes, count\_of\_total\_pairs)， interest\_rate(individual, confidence) These three functions find the support, confidence and interest of all association rules respectively.
2. indi\_support(individual\_pairs, count\_of\_total\_pairs） Function is used to calculate the probability of occurrence of a single theme in the dataset.

***Results***

***The difference in speed between different numbers of processors***

sequential takes 20 min+ and goes error



2core 313.77718925476074s



4core 154.28740191459656s



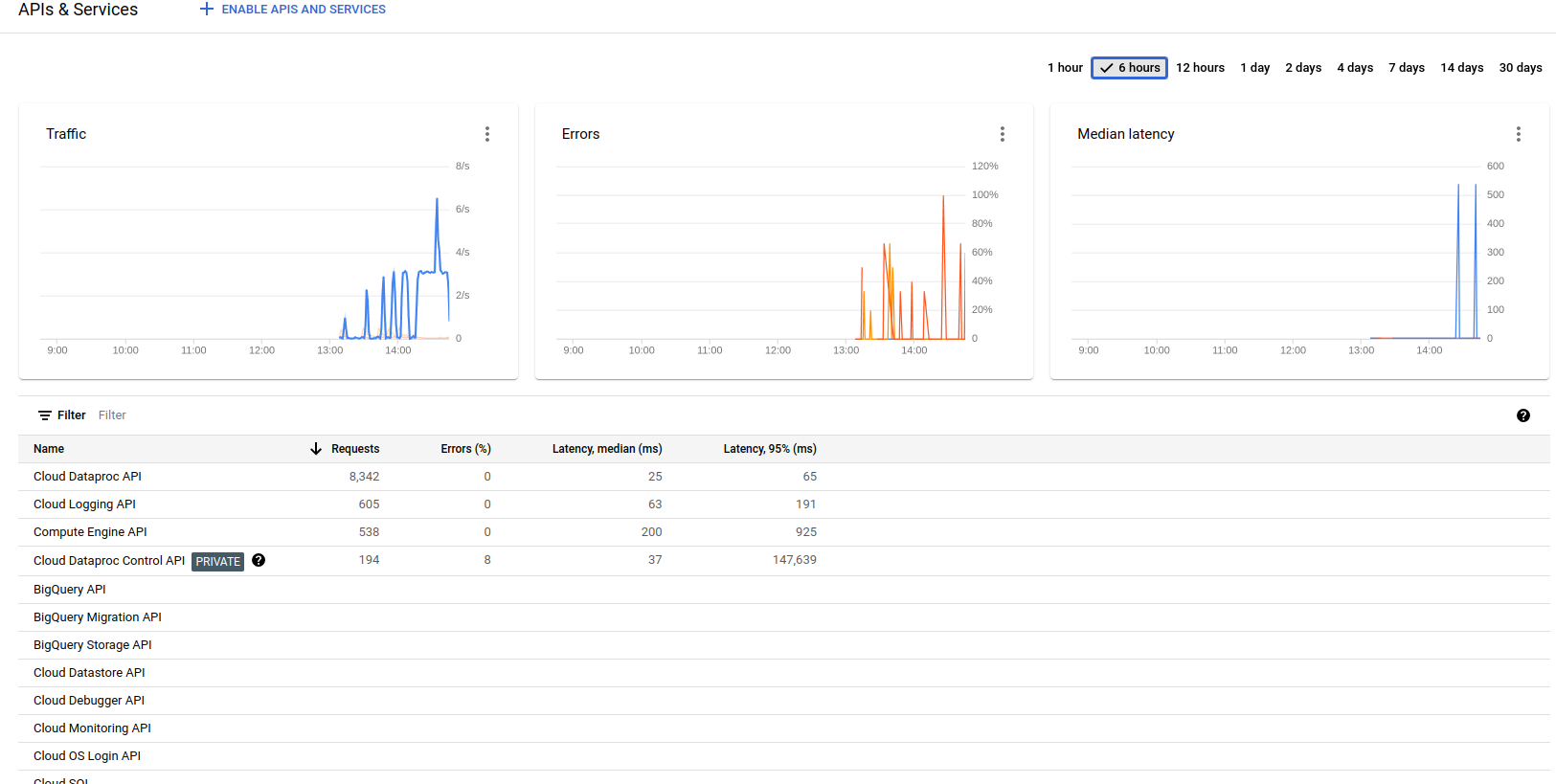
8 core 90.13733100891113s

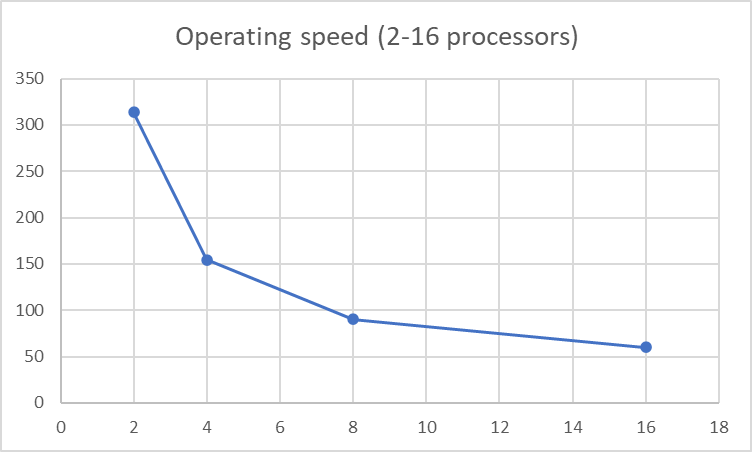


16 core 60.23396849632263s



graph





As can be seen from the graph above, the speed of running a program from two processors to 16 processors increases all the time. This process is roughly exponentially distributed, with the cost of communication causing the speed to increase slowly once the number of processors is greater than 16.

***result***

| individual\_Pairs | | |
| --- | --- | --- |
| Before | Start | Still |
| ('EPU\_ECONOMY\_HISTORIC', 1902),0.4725465838509317 | ('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 5018)，0.7160388127853882 | ('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 1394)，0.4671581769436997 |
| ('TAX\_ETHNICITY\_CHINESE', 1634),0.40596273291925467 | ('TAX\_DISEASE\_CORONAVIRUS', 4831)，0.6893550228310502 | ('GENERAL\_HEALTH', 1382)，0.4631367292225201 |
| ('TAX\_WORLDLANGUAGES\_CHINESE', 1634),0.40596273291925467 | ('GENERAL\_HEALTH', 4789)，0.6833618721461188 | ('USPEC\_POLITICS\_GENERAL1', 1260)，0.4222520107238606 |
| ('USPEC\_POLITICS\_GENERAL1', 1554),0.38608695652173913 | ('WB\_635\_PUBLIC\_HEALTH', 4383)，0.6254280821917808 | ('CRISISLEX\_CRISISLEXREC', 1243)，0.4165549597855228 |
| ('LEADER', 1494),0.37118012422360247 | ('WB\_2165\_HEALTH\_EMERGENCIES', 4100)，0.5850456621004566 | ('TAX\_ETHNICITY\_CHINESE', 1224)，0.4101876675603217 |

These are the top 5 occurrences of individual themes for three different periods. Due to the bias in the data set the size of the dataset is 4025 for 2019, 7008 for 2020 and 2984 for 2021, so the second year will seem to have more themes. Looking at the 2019 data, the themes that appear most often are more related to the economy, taxation, and international politics. The most occurring theme in 2020 was new coronavirus and public health events, with a theme rate of 70 per cent. The other top 5 health-related themes are all greater than 50 per cent. 2021 will have the most health-related themes at 40 per cent, so we can deduce that the impact of the epidemic is diminishing.

| support | | |
| --- | --- | --- |
| Before | Start | Still |
| (('TAX\_ETHNICITY\_CHINESE', 'TAX\_WORLDLANGUAGES\_CHINESE'), 0.4057142857142857) | (('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 'WB\_635\_PUBLIC\_HEALTH'), 0.6254280821917808) | (('GENERAL\_HEALTH', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.4195710455764075) |
| (('LEADER', 'USPEC\_POLITICS\_GENERAL1'), 0.311055900621118) | (('GENERAL\_HEALTH', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.6242865296803652) | (('TAX\_ETHNICITY\_CHINESE', 'TAX\_WORLDLANGUAGES\_CHINESE'), 0.4101876675603217) |
| (('EPU\_POLICY\_GOVERNMENT', 'GENERAL\_GOVERNMENT'), 0.3003726708074534) | (('TAX\_DISEASE\_CORONAVIRUS', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.6090182648401826) | (('UNGP\_HEALTHCARE', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.3930965147453083) |
| (('LEADER', 'TAX\_FNCACT\_PRESIDENT'), 0.2732919254658385) | (('WB\_2165\_HEALTH\_EMERGENCIES', 'WB\_635\_PUBLIC\_HEALTH'), 0.5850456621004566) | (('GENERAL\_HEALTH', 'UNGP\_HEALTHCARE'), 0.3900804289544236) |
| (('EPU\_ECONOMY\_HISTORIC', 'USPEC\_POLICY1'), 0.26881987577639754) | (('WB\_2165\_HEALTH\_EMERGENCIES', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.5850456621004566) | (('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 'WB\_635\_PUBLIC\_HEALTH'), 0.36896782841823056) |

These are the top 5 paired themes that appear in three different periods.

Prior to the start of the epidemic, in 2019, articles in Chinese society frequently discuss themes such as taxation, international politics, and the United States, with a greater degree of similarity between the two themes, either the same taxation or the same government. The theme that appears in pairs in 2020 and 2021 are the epidemic, nutrition, interpopulation combinations, and to lesser extent taxation issues. This indicates to some degree that the coronavirus is having an impact on Chinese society.

| confidence | | |
| --- | --- | --- |
| Before | Start | Still |
| (('ECON\_FREETRADE', 'WB\_698\_TRADE'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'USPEC\_POLICY1'), 1.0) | (('WB\_1427\_NON\_COMMUNICABLE\_DISEASE\_AND\_INJURY', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_FNCACT\_OFFICER'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'WB\_1920\_FINANCIAL\_SECTOR\_DEVELOPMENT'), 1.0) | (('WB\_1427\_NON\_COMMUNICABLE\_DISEASE\_AND\_INJURY', 'WB\_1406\_DISEASES'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_MILITARY\_TITLE\_OFFICER'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'EPU\_CATS\_MONETARY\_POLICY'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'USPEC\_POLITICS\_GENERAL1'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_FNCACT\_CHIEF'), 1.0) | (('WB\_2934\_COPPER', 'WB\_507\_ENERGY\_AND\_EXTRACTIVES'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'TAX\_FNCACT\_PRESIDENT'), 1.0) |
| (('EPU\_CATS\_TAXES', 'USPEC\_POLICY1'), 1.0) | (('WB\_2934\_COPPER', 'WB\_895\_MINING\_SYSTEMS'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'LEADER'), 1.0) |

This is the confidence of the top 5 pairs of themes for the three different periods, and the table shows that the confidence is 1.0, indicating that whenever the first theme appears in the pairs the second theme also appears. In 2019, 'TAX\_FNCACT\_OFFICER', 'TAX\_MILITARY\_TITLE\_OFFICER' and 'TAX\_FNCACT\_CHIEF' appear whenever TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER appears. 2020, 'TAX\_FNCACT\_OFFICER' appears when EPU\_ POLICY\_CENTRAL\_BANK appears 'WB\_1920\_FINANCIAL\_SECTOR\_DEVELOPMENT, 'EPU\_CATS\_MONETARY\_POLICY and WB\_507\_ENERGY\_AND\_EXTRACTIVE both appear. 2021 The theme of population and disease is mentioned whenever non-communicable hazards are mentioned. Also when 'TAX\_FNCACT\_VICE\_PRESIDENT is mentioned, USPEC\_POLITICS\_GENERAL1, 'TAX\_FNCACT\_PRESIDENT and LEADER will also be mentioned.

| interest | | |
| --- | --- | --- |
| Before | Start | Still |
| (('WB\_707\_POVERTY\_MEASUREMENT\_AND\_ANALYSIS', 'WB\_1152\_ECONOMIC\_AND\_SOCIAL\_MOBILITY'), 0.9873291925465838) | (('TAX\_MILITARY\_TITLE\_SOLDIERS', 'TAX\_FNCACT\_SOLDIERS'), 0.992579908675799) | (('TAX\_WORLDLANGUAGES\_ENGLISH', 'TAX\_ETHNICITY\_ENGLISH'), 0.9829088471849866) |
| (('WB\_1152\_ECONOMIC\_AND\_SOCIAL\_MOBILITY', 'WB\_707\_POVERTY\_MEASUREMENT\_AND\_ANALYSIS'), 0.9873291925465838) | (('TAX\_MILITARY\_TITLE\_COMMANDER', 'TAX\_FNCACT\_COMMANDER'), 0.9924372146118722) | (('TAX\_ETHNICITY\_ENGLISH', 'TAX\_WORLDLANGUAGES\_ENGLISH'), 0.9829088471849866) |
| (('TAX\_FNCACT\_CAPTAIN', 'TAX\_MILITARY\_TITLE\_CAPTAIN'), 0.986832298136646) | (('TAX\_WORLDLANGUAGES\_GERMAN', 'TAX\_ETHNICITY\_GERMAN'), 0.9921518264840182) | (('WB\_2601\_TRADE\_LINKAGES\_SPILLOVERS\_AND\_CONNECTIVITY', 'WB\_772\_TRADE\_FACILITATION\_AND\_LOGISTICS'), 0.9819034852546917) |
| (('TAX\_MILITARY\_TITLE\_CAPTAIN', 'TAX\_FNCACT\_CAPTAIN'), 0.986832298136646) | (('TAX\_ETHNICITY\_GERMAN', 'TAX\_WORLDLANGUAGES\_GERMAN'), 0.9921518264840182) | (('TAX\_ETHNICITY\_KOREAN', 'TAX\_WORLDLANGUAGES\_KOREAN'), 0.9815683646112601) |
| (('TAX\_RELIGION\_MUSLIM', 'TAX\_ETHNICITY\_MUSLIM'), 0.986583850931677) | (('WB\_2120\_SATELLITES', 'WB\_2329\_ACCESS\_AND\_CONNECTIVITY'), 0.9920091324200914) | (('TAX\_WORLDLANGUAGES\_KOREAN', 'TAX\_ETHNICITY\_KOREAN'), 0.9815683646112601) |

These are the top 5 interests for pairs of themes over three different periods. Interests in the table are generally greater than 0.98 and less than 1, which means that these themes are interesting. For example, in 2019, the two themes of interest are property management and economics. The most interesting themes in the three years above are generally tax themes, while no keywords related to the epidemic appear.

***Conclusion***

The project believes that the current study produces data that demonstrates a shift in the themes of greatest interest to Chinese society from economic and political themes to health and epidemic themes after the start of the epidemic in 2020. It is interesting to note that the 2020 dataset has a larger volume of data than the other two datasets and a very high proportion of epidemic themes (the highest being 70%), while the 2021 dataset is smaller than the 2020 dataset and the proportion of epidemic themes is not as high as the 2020 dataset (the highest being 40%), suggesting to some extent that the epidemic situation has improved. Looking at the top 5 themes in the SUPPORT value, it is clear that the combination of the three themes of Taxation, International Politics, and the US often appear two by two in 2019. In 2020 and 2021, the themes that appear in pairs are a combination of epidemic, nutrition and population. In particular, the top 5 theme combinations in 2020 all have a support value greater than 0.5, which demonstrates that people often discuss epidemics, nutrition and population together. This is an indication of the impact of the epidemic on Chinese society.

By looking at the themes in Chinese society, this report is able to infer that the emergence of the epidemic had a notable impact on Chinese society, with the impact diminishing over time. This project can infer whether the epidemic has had a serious impact and whether its trend is in a good or bad direction by looking at the level of attention given to the theme of the epidemic by societies. As well as inferring whether there is a possibility of a new coronavirus variant emerging. Countries with severe epidemics can learn from countries with low epidemic impact about how to deal with epidemic viruses.

The project also calculates confidence and interest, looking at the results of interest and finding that there are often links between different tax themes. Confidence indicates that the emergence of one theme often leads to another, e.g. EPU policy leads to the mutual implication of economic issues and tax issues. This allows future projects to correlate these high-confidence themes and to further explore the relationships between them.

***Critique of Design and Project***

1. The project's implementation of the Apriori algorithm only iterated once, producing a relevant set consisting of only two elements. One of the priorities of the improvement project is to implement an iteration of the Apriori algorithm that produces 5 or 6 frequent item sets. By adding more themes to the frequent itemset, correlations between four or five themes can be studied simultaneously. Or the possibility that adding more themes can be strongly correlated, and the feedback on the results will be more accurate. This project only used the Apriori algorithm once, lacking the simultaneous exploration of more than two themes, and although the current approach of the project runs more efficiently, it is lacking in accuracy in terms of results.
2. The project did not perform a correlation analysis of the themes with high values of interest and confidence from the send after calculating interest and confidence. In the association rule, the lift can represent the correlation between themes and can prove whether there is a correlation between themes. This would be useful for further research to investigate and explain, for example, what events are highly correlated with epidemics and what events are correlated with a particular policy. This project has calculated interest and confidence, but the lack of correlation investigation on the way to analysis has not produced impressive or meaningful conclusions from these two data.
3. The project does not have a scientific method for testing the similarity of the data generated over three periods. It is possible to compare the similarity between themes in a similar way to Cosine Similarity after the data has been calculated. Doing so would allow an assessment of how much the epidemic has affected changes in social themes, as well as preparing for future research into trends in epidemic change. This idea was not implemented in the project due to my inability to develop a good implementation method.

***Reflection***

1.Market basket analysis. The project uses the idea of market basket analysis as a theme of analysis, and the market basket analysis methodology is integrated throughout the project. (lab3)

2.Association rules. Association rules are a way of finding interesting relationships between variables in large databases. (lab3)

3.The Apriori algorithm is the process of finding frequent itemsets.

4. Throughout the DATA301 course the spark and Google Could Platform are used as the analytics platform. spark-specific arithmetic such as Map and Reduce are used.

Through this project, I was able to gain more insight into the idea of market basket analysis and its implementation. In the process, I was introduced to the Gdelt database, a global social database. I learned to find and use the data I wanted from the Gdelt database, using the gdelt library in Python. This was the first project that I worked on alone, from designing the problem, applying the algorithm and completing the report, which improved my skills in many areas.

***References***

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of Massive Datasets* (3rd ed.). Cambridge University Press.

Wikipedia contributors. (2022, June 2). *COVID-19*. Wikipedia. https://en.wikipedia.org/wiki/COVID-19