**Individual Progress Report**

**11969443**

What have you done so far on your project? Please be as detailed as possible. List any code you wrote or tested, papers/documentation you read, data you attempted something with, etc. Note any changes to project direction that resulted from your work.

The original plan was：

1. week1: complete extraction of all data and basic processing, filtration of the required data.
2. week2: Start summarising the data using the A-Priori Algorithm.
3. week3: Complete all debug work and wrap up the project.
4. Week 4: Completion of the report

Now that I have completed step two, I am starting step three.

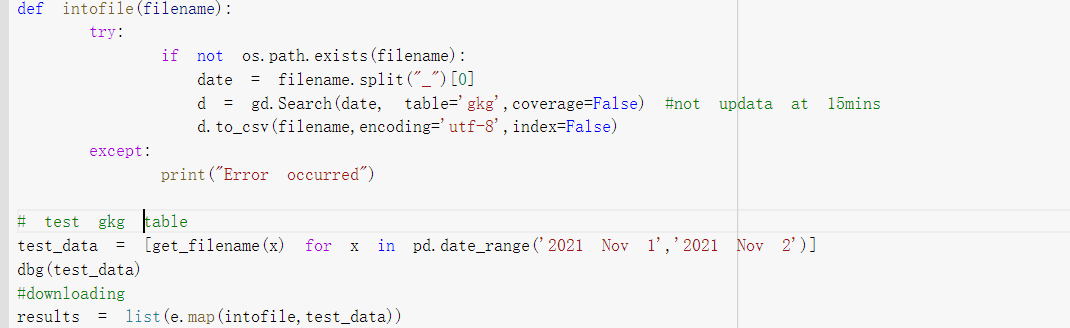
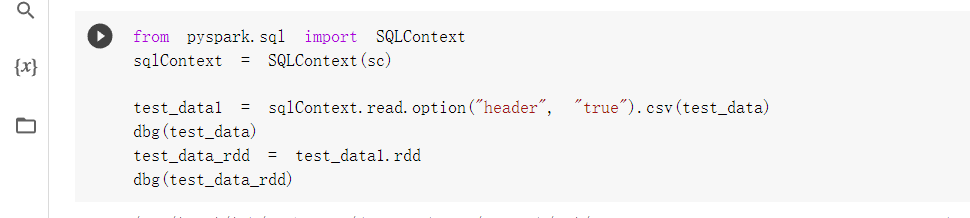
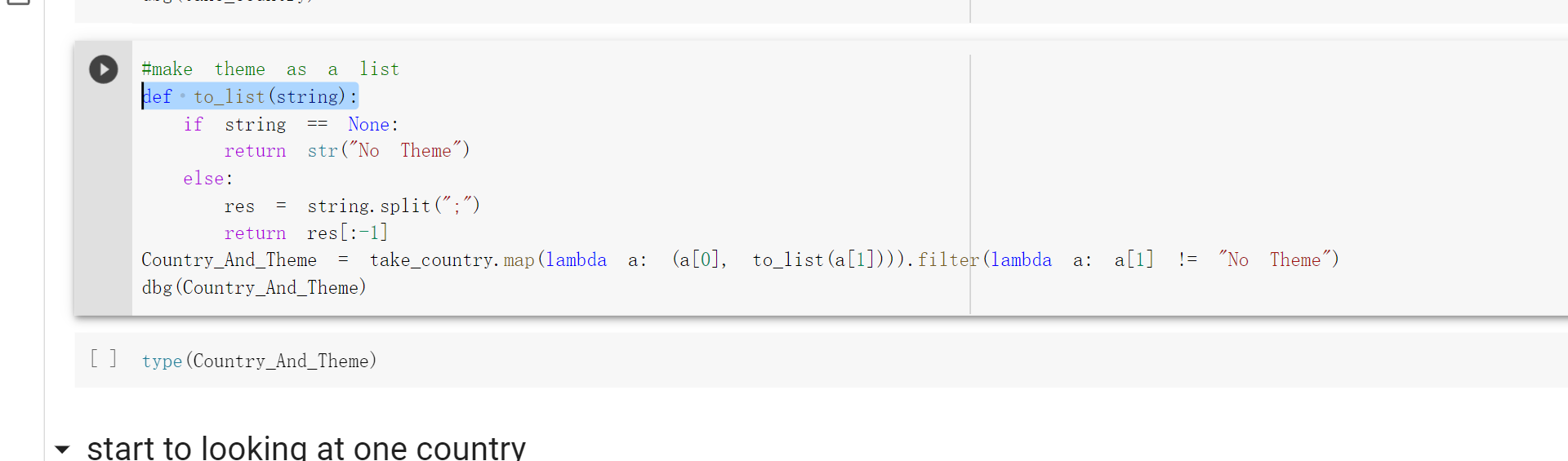
At the beginning of the second week, I used the Gdelta event table to analyse the country relationships to solve the problem I had at the beginning of the survey. Since my goal was to do a Market basket analysis of them, a Market basket can only contain a maximum of two countries. This made it much faster to summarise ((country1, country2), times\_mentions) directly using the Map and reduce methods than to analyse them by first calculating the support values of individual items. the A-Priori Algorithm was not powerful in this case. So I changed the original problem.

***Previous Question: Which countries are most likely to generate sub-conflict during the Coronavirus period (2021)?***

***Change Question: What is the theme of the article in 2021 in China, the United States, and Russia respectively? Use Market Basket Analysis (MBA) on them to identify the themes of the article that often appear in pairs.***

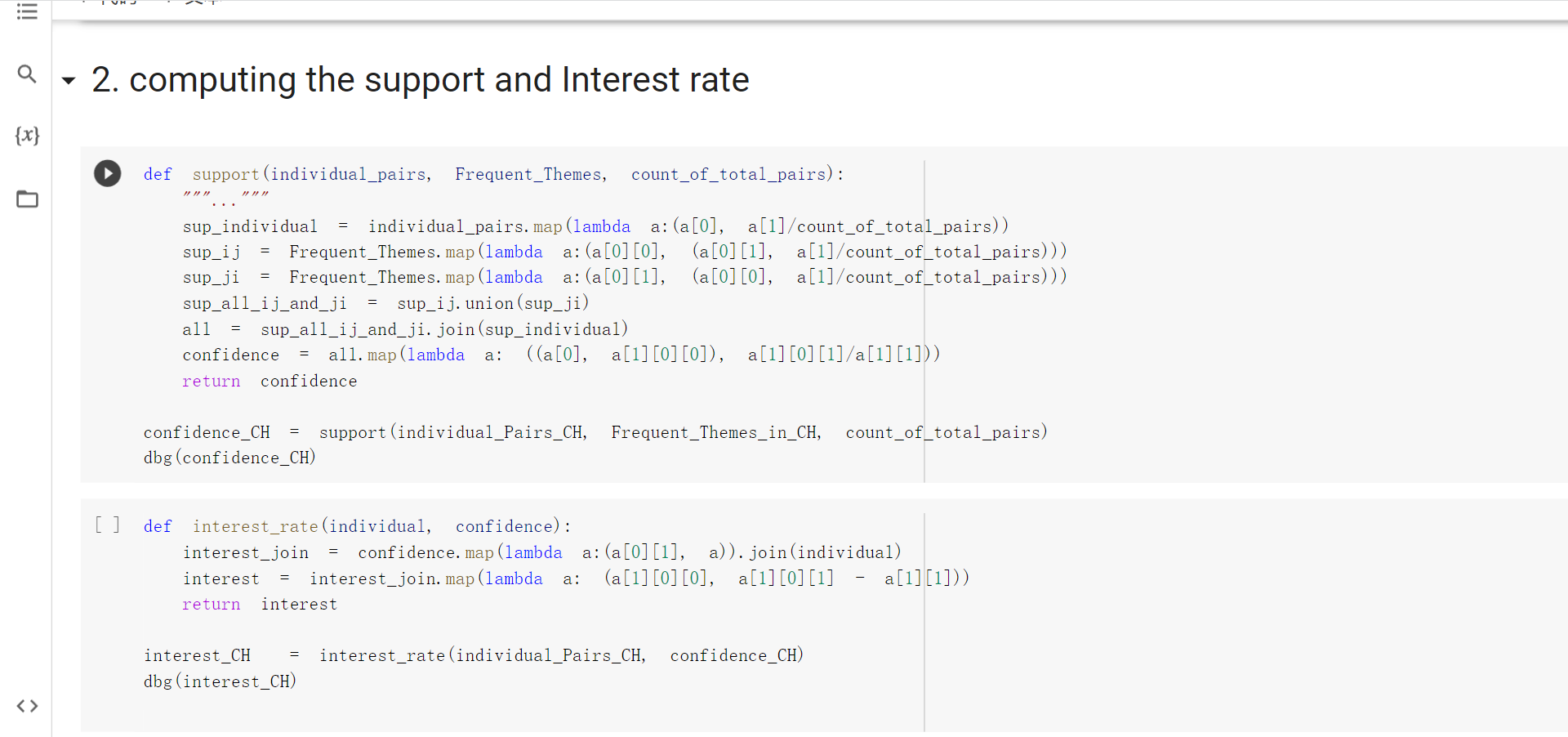
After changing the problem I used GKG Table to analyse the topic of the article for a particular country. First I filtered the articles from a particular country out of the Table using the filter function. Afterwards, I carried out a Market Basket Analysis on the topics in V1Theme. Finally, a summary was made.

What have you done so far on your project?

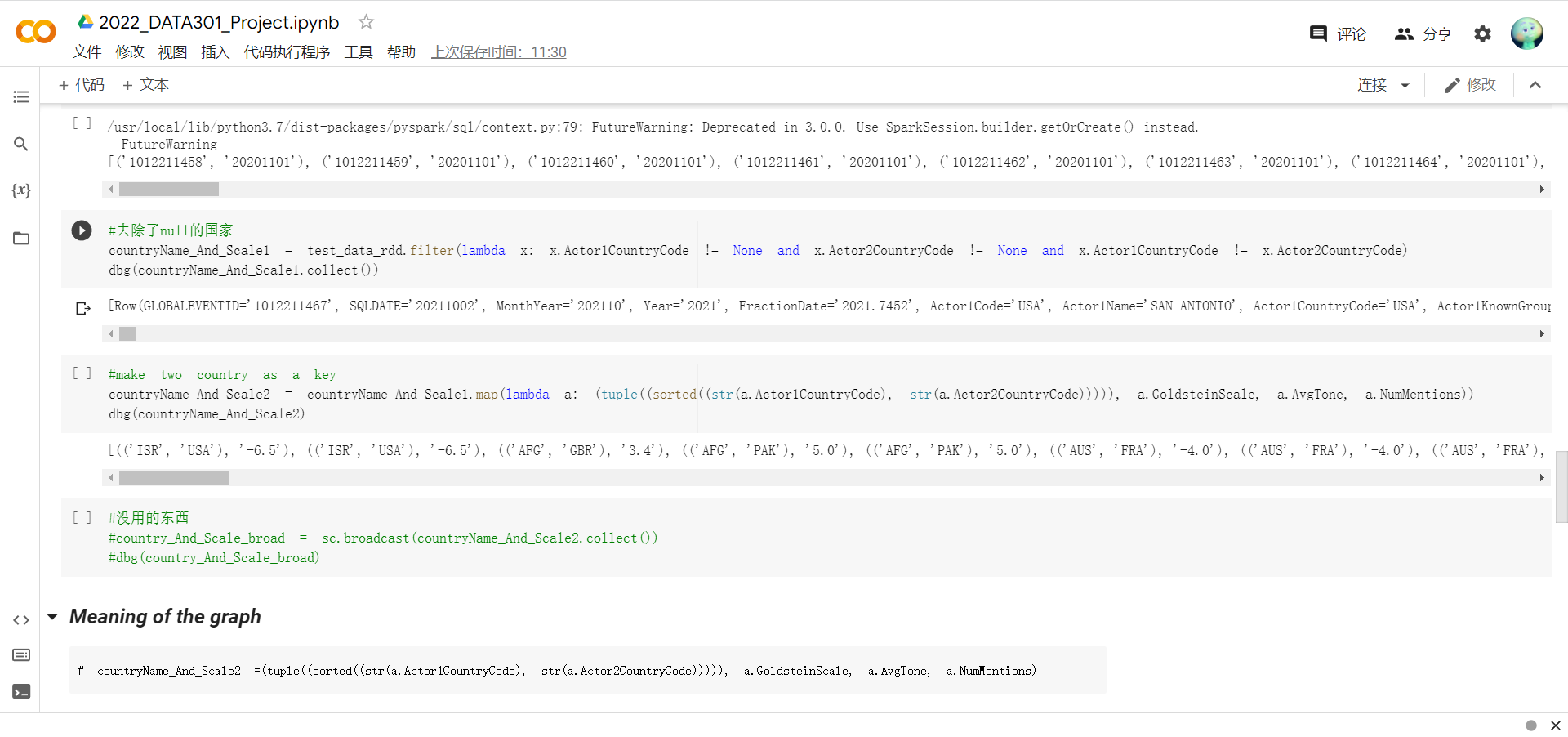
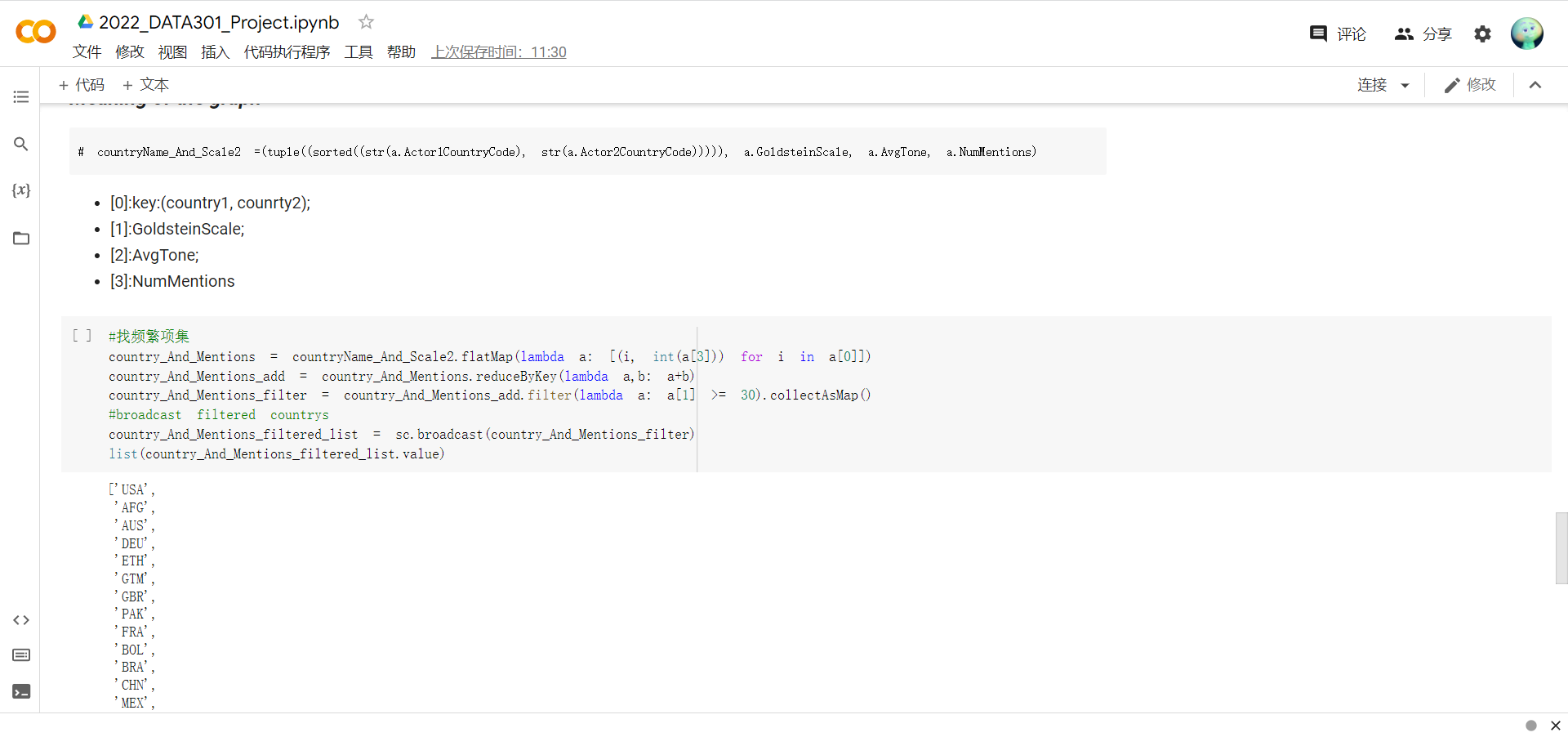
* 1.I have now managed to download two days of data (due to technical problems I am not yet able to manipulate one month or one year of data).
* 
* 2. I then read the resulting data into the RDD to easily manipulate it later
* 
* 3. I used the Row feature of the RDD to extract the fields I needed to process using a.Locations and a.Themes. I then created (def find\_country(string)) and (def to\_list(string)) to extract the country name and the subject of the article from a.Locations and a.Themes respectively.
* 
* 
* 4. def Frequent\_Itemsets(countryCode) Enter the countryCode and the function automatically finds the relevant countries in the RDD for each article's topic. and fits these lists into a large List, thus performing a Market basket analysis. This function will output

1. the number of times each item appears individually in the Market basket 2. the number of times frequent Pairs appear

3. the size of that Market basket.

* 5. Start performing the first step of the A-Priori Algorithm. I wrapped the first step into a function for later use in the analysis of more countries. I find the frequency set according to the first step in Lab3 and Lecture7 PPT and then look for Pair in each row of the Market basket so that both items of Pair are in the frequency set.
* 
* 6. The second step is to calculate the confidence\_rate and interest\_rate separately as mentioned in Lecture 7.
* 

Screenshot of the test code (above is the completed code)

* 
* 

What are you doing next? What is your plan to complete the required project components on time?

* I will be running my code with a larger data set. Any problems that arise there I'll work out slowly.
* As soon as I can run more data I will run the Market basket algorithm once for China, once for the US and once for Russia to find out which themes appear at the same time and analyse them.
* Finally， start writing a report summarising the project.

Do you have any roadblocks? Please describe things that are preventing you from making progress such as computing resources or difficulties using software/algorithms. Note that time itself is not a road block.

* 1. When converting Context to RDD, the conversion process fails for some missing date data, and this is the problem I have concentrated on so far.
* 

Issue 1 Solution: This has now been resolved. Solution: In the intofile function, return the filename if the download is successful. Then it iterate through the returned filenames, removing the "None" variable, to get the actual downloaded files, and then read them into the SQLContext.

* 
* 2. For the analysis of the data obtained. What information can I get from mining for relevant topics？ Is this information useful?

The main problem is that it is not clear how to apply the interest rate to the data analysis and what meaning to give to the data.