Data 301 Project Proposal

Shivin Gaba (sga113)

Summary

I propose to research how similar was the trend of the public sentiment towards the economy over the period of Covid 19 lockdown in New Zealand when compared to this time last year.

This will use the GDELT 2.0 data set that records the URLs of the newspaper articles for each day. Those URLs further with the help of the GDELT summary API would be used to extract the interesting set of keywords from the title of the news articles. To compile the data I will filter the news articles by the country name and range of dates that I am interested in. The main algorithm that I will use to investigate the proposed question would be Apriori which has applications in domains such as market-based analysis. I will also be using filtering and map-reduce operations.

The intended result should reflect the rise in the negative sentiments related to the economy and the housing market with keywords such as unemployment, rate of interests, job-losses, AVRs etc when compared to the past year.

Background

I will use the GDELT 2.0 summary API to generate a table that will include columns such as URLs, dates, and titles. I will be studying the change in the trend of the keywords that appear in the title of the article. Each of those articles to the title will then be used as a separate basket. With the help of features available on the GDELT summary API, I can then query the articles for the specific country for a specified date, which would then allow me to compare the result for various times and compute the change in the general trend. I am planning to use the Apriori algorithm to investigate the above research question. This algorithm proceeds by identifying the frequent individual items and then extending them to a larger and larger item sets. The frequent items set determined by the Apriori can be used to determine associations rules which highlight general trends in the database.

Hypothesis

Due to the COVID-19 and NZ imposing the lockdown I the hypothesis that there has been an increase in the public sentiment towards the economy when compared to this time last year.

I am planning to research the similarities in the trend by going through the news article provided in the GDELT dataset and look for specific keywords.

I will specify up-to 20 interesting keywords related to the economy such as job market, unemployment, benefit, job-loss, interest rates, property, foreign investment, housing, AVR, etc. I will use the change in the trends of these keywords to measure the effect of this pandemic on

New Zealand’s economy and housing market. Specifying the above keywords in the GDELT summary API in return would filter my data set and generate a table with only relevant articles in relation to my research questions. I can then use the association rule in order to find the interest and use the title column to retrieve the keywords from each article.

Design and Method

After creating the relevant GDELT dataset tables, I will be generating a CSV file. Specifying interesting keywords should already filter my data to some extent but I might end up dealing with the huge amount of dataset. Therefore, I will be just researching the trends in New Zealand’s economy. If the processing of data is fast enough and does not cost lots in terms of computing cost then I might also end up comparing the trend with another country. Initially, I will start computing the data for 3 consecutive days and test the storage and computing costs. Based on those results, I might end up adding more days to increase the quantity of data that will be used to measure the trend.

After generating the CSV file I will map it into an RDD and then will apply the two-step Apriori algorithm to get the item pairs and their counts and arrange them in descending order. Then I will calculate the confidence and the interest score of each pair and select the top 20 pairs. I will then follow the same process on the set of dates for the previous year and find the top 20 pairs and compare the similarities between them. Confidence and interest would be two measures that I will use to compare my frequent keyword pairs which provide me the information to see the similarities or changes in the trend over time.

References

https://en.wikipedia.org/wiki/Apriori_algorithm

https://api.gdeltproject.org/api/v2/summary/summary