**DATA301 Proposal**

11969443 Zhenyuan He

**1. Summary**:

a.subset of GDELT data:

I am going to use the data from GDELT 2.0 Event Database with a subset selected:

1. EVENTID AND DATE ATTRIBUTES: GlobalEventID, FractionDate

2. ACTOR ATTRIBUTES: Actor1Code, Actor1CountryCode, (or more)

3. EVENT ACTION ATTRIBUTES: EventCode, GoldsteinScale, NumMentions, NumSources, NumArticles, AvgTone (mainly using these six or more parameters)

4. MENTIONS TABLE: MentionType, Confidence (or more).

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

c.algorithm: A-Priori Algorithm and Naïve Algorithm

d.There may be China, Russia, the USA, Japan, etc.

e.The significance of predicting which countries are likely to emerge can be predicting economic trends, analyzing the global situation, etc.

**2. Motivation**: As an international student, I am very concerned about the global situation and curious about the world landscape. This research can be used to predict future international situations.

**3. Background：**

The relationship between countries is studied under the epidemic by the severity of the information released.

GlobalEventID in GDELT 2.0 Event Database is unique and is the main attribute of an event. actor ATTRIBUTES has two Actor names, country, and ethics, and we will mainly apply the country attribute to divide it. event ACTION ATTRIBUTES in NumMentions, NumSources, NumArticles, AvgTone can be used to count the social impact of an event. Confidence in MENTIONS TABLE: can predict the credibility of an event.

**4. Research Question or Hypothesis：**

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

change: What is the theme of the article in 2021 in China, the United States, and Russia respectively?

Hypothesis: The result includes the United States, China, and Russia.

Implement a shopping basket analysis similar to the one in LAB3, using the A-Priori algorithm to iteratively find which countries are most likely to have conflicts between them (construct one with roughly K=5 or 6).

**5. Design and Methods**:

1. Items that appear ≥ 100 times are the frequent items(just guess 100 first, change during the project, or min support >= 30%). and filter out events with GoldsteinScale <= 0 and social impact (NumMentions, NumSources, NumArticles above average).

2. Read baskets again and count in main memory only those pairs where both elements are frequent (from step 1).

3. Repeat the first and second stages until you find a set containing 5 or 6 or items. (summary1)

4. Starting from the second step of the analysis

1. Calculate the support of individual countries.

2. calculate the confidence between two countries

3. calculate the INTEREST values for each country in the set generated in the third step. (summary 2)

5. Summarizing the results of summary1 and summary2, summary1 indicates the countries that are more likely to be in conflict, and summary2 indicates which other countries these countries are more likely to be in conflict with.

**6. References**

just source on learn.