

DATA301 Final Project

Isaac Dann - 47452629

Abstract / Summary:

My project is on comparing the type of events when Actor1's CountryCode is the United States of America. I will be using the GDELT 2.0 Events Database for my analysis and I will use a "Market Basket" set up and the two step A-priori algorithm to gather relevant data and use support, confidence and interest equations to analyze my data. The intended result is to end up with a list or set of top pairings that have high confidence and interest, which in theory I could use to analyze what world events are specific to an individual or a small group of countries in the set time frame I want to analyse. My results will also give information on how different event types are commonly paired together, possibly giving insight into how the people, government or organizations from the US conduct themselves internationally.

Introduction:

Background:

I have looked into how the GDELT 2.0 database works, how it chooses its country codes for its Actors, classifies events using language and what event type the number codes correspond to. I have extensively played around with pyspark and gdeltPyr to gather data from dates in the events table and I am confident I am able to do this.

Market Basket analysis is commonly used in retailers, but can be used to analyse my dataset, as the whole dataset where USA is the Actor1CountryCode is the market, the baskets are defined by the Actor2CountryCodes, where each country is its own basket and the Unique event Codes are the Items. The resulting interesting pairs will be event type pairings which have a high confidence score but aren't common in the dataset

Motivation:

As a follower of world politics and a USA born New Zealander, I am particularly interested in the USA's overbroad activity. The US is seen as a global power and I want to get a set of interesting event pairings to see if it gives any insight into the way they engage with other countries. Especially event type pairs with high interest scores, as these will be in theory more relevant to the time period and rarer or specific to certain countries, conflicts or situations. I am looking forward to gathering and analysing my results so I can see what event pairings are of interest according to data and the algorithms and sharing my results in this report

Research Question:

What are the most interesting pairings of types of events in the month of January in the past 3 years where an organisation, government, person/s etc from the United states is involved as the main actor in the event?.

This question will use the ‘ActorCountryCodes’ and ‘EventCode’ attributes from the Gdelt event database to gather relevant data about the USA and types of events. It also uses the CAMEO manuel to define ‘Event Types’

Experimental Design and Methods:

The dataflow begins with importing the necessary module, namely gdeltpyr, pyspark among others. Gdeltpyr is a python module that can pull GDELT 2.0 data into dataframes. We use gdeltpyr to import data using the gdelt Search built in, multi-process the query using ‘ProcessPollExecuter’ and write it to a csv file to be manipulated later.

Using the SQLContext module, the data is read from the csv and I can begin processing it. I then mapped the data to an rdd and filtered it by Actor1CountryCode == ‘USA’. Then I mapped it to an rdd where the format is tuples of form ((Actor1CountryCode, Actor2CountryCode) , Event Code).

The Data is then grouped by key to give tuples of form ((Actor1CountryCode, Actor2CountryCode) , list(set(Event Codes))), where we have country codes and all the unique events associated with them. This gives us events by country.

Now we can start the first step of the two step A-priori algorithm. We go through the baskets and flatMap them by the event codes, then map and reduce them into tuples with (eventcode, Num of countries where the event type occurred with USA as actor1), and filtered by events with greater than 5 occurrences. This is done to speed up the algorithm and take out one off events / event types, as these are less likely to give insight into how the USA operates abroad. This is then broadcasted using spark as a collected map, which acts as a dictionary.

We then begin step two of the A-priori algorithm. I wrote three functions, srtTup, Pairs and Filtered_Pairs. SrtTup sorts a given tuple and both Pairs functions take a list of events (items) from a country (basket). Then the function takes for every event type in the list twice, if the event types are both in the the broadcasted map, they get sorted into an ordered pair which is output by the function. In Filtered_Pairs, the same happens but an extra check occurs to see if one or the other event types are in another broadcasted map. This is so we can check all event types against a subset of event types, such as ‘PROVIDE AID’ or ‘PROTEST’. This can be useful for seeing specific subsets of our data, and investigating what event types cause the USA to provide aid to themselves or other countries, or what event types cause Protest, or how the USA responds to protest. Obviously this is just an assumption based on our data, but could provide insight into how the USA can react based on real world data.

Using our Functions, the Event Types are grouped into pairs of event types and the frequency of the occurrence of both event types in countries with the USA. We then perform our Support, Confidence and Interest equations. These are :

- Support : $\text{support}(i, j) = \text{freq}(i \text{ and } j) / \text{number of countries}$
- Confidence : $\text{conf}(I \rightarrow j) = \text{support}(I \cup j) / \text{support}(I)$
- Interest : $\text{interest}(I \rightarrow j) = \text{conf}(I \rightarrow j) - \text{Pr}[j]$

After these equations have been calculated we are left with a list of confidence value and event type pairs and a list of interest values and event type pairs. I then convert the event types into their text description using a dictionary I made to give the top 10 event pairs of interest.

Results:

In the following Tables, the order of the format is: The first Event Type, The Associated Event Type and the value for the Interest of the association rule, which is the difference between its confidence and the fraction of the baskets. In all cases I have chosen a top 10, based on both their Interest scores as well as how interesting I personally found them to be.

Results

January 2020 - No Specific filtered events / top event type pairs of interest

| Event Type | Associated Event Type | Interest |
|--|--|----------|
| Demand policy support | Use tactics of violent repression | 0.957 |
| Demand policy support | Increase military alert status | 0.957 |
| Demand policy support | Appeal for military protection or peacekeeping | 0.957 |
| Demand policy support | Reject plan, agreement to settle dispute | 0.952 |
| Engage in mass expulsion | Accuse of crime, corruption | 0.948 |
| Threaten political dissent, protest | Coerce, not specified below | 0.943 |
| Forgive | Reject request or demand for material aid, not specified below | 0.943 |
| Demand policy support | Reject request or demand for material aid, not specified below | 0.943 |
| Forgive | De-escalate military engagement | 0.938 |
| Express intent to cooperate militarily | Ease economic sanctions, boycott, embargo | 0.938 |

From our interesting values, we can see in January 2020 the US was involved in events where there was a demand for policy support. Since our data is filtered to only use event types with greater than or equal to 5 occurrences, we can see that in cases where there was a “Demand policy support” event type, from our data, we can say it is likely that there was a military conflict or violent conflict resulting from government policy, either of the USA or another country, where a policy request or request for peace was denied and military protection was requested and military alert status was increased. The use tactics of Violent repression is an interesting read from the data, and would be interesting to see if this is a common event type pairing for the USA if we looked at past years from the GDELT 2.0 events table.

Another interesting pairing is “Engage in Mass Expulsion” and “Accuse of Crime or Corruption”. As this data point also has a confidence of 1.0, 100% of the time where the USA engaged in a mass expulsion, there was an accusation of crime. This could mean the USA don’t act without what they see as reasonable justification on their behalf, but the “Accuse of Crime or Corruption” event type, might be referring to a much more minor crime in some cases than the severity of an “Engage in Mass Expulsion” event type. This could in theory suggest the USA punish certain groups harshly when given what they see as a justifiable opportunity. Cross referencing the data to find the specific events would be necessary for further investigation into this topic.

The final one I’ll look at for January 2020 is the pairing of “Threaten political dissent, protest” and “Coerce, not specified below”. This implies in the rare cases the US Threatened political dissent, they acted upon it in some manner. The “Coerce” event type means “Repression, violence against civilians, or their rights or properties not otherwise specified.”. I believe this is evidence that the US are very decisive, or at least were in January, that they followed through on their threats in some manner, 100% of the time they were made.

January 2021 - No Specific filtered events / top event type pairs of interest

| Event Type | Associated Event Type | Interest |
|---|---|-----------------|
| Allow international involvement not specified below | Increase police alert status | 0.965 |
| Impose state of emergency or martial law | Conduct suicide, car, or other non-military bombing, not spec below | 0.946 |
| Allow international involvement not specified below | De-escalate military engagement | 0.936 |
| Express intent to mediate | Express intent to settle dispute | 0.936 |

| | | |
|---|---|-------|
| Allow international involvement not specified below | Impose restrictions on political freedoms | 0.931 |
| Express intent to mediate | Apologize | 0.926 |
| Allow international involvement not specified below | Express intent to de-escalate military engagement | 0.926 |
| Demand de-escalation of military engagement | Fight with artillery and tanks | 0.921 |
| Express intent to provide military aid | Fight with artillery and tanks | 0.921 |
| Engage in mass killings | Fight with artillery and tanks | 0.921 |

From our interesting Values for this January 2021, our most interesting event type correlations seem to imply that there was specific regional conflict the US was involved with. It appears that there was martial law and police presence increased, then eventually an attempt to mediate was made.

Our 8th, 9th and 10th most interesting event type correlations show that where the US “provided military aid” and “demanded de-escalation”, there were Mass Killings and fighting with artillery and tanks. From this we could infer where the USA sees global injustices such as mass killings, they will provide military aid and give demands. On the other hand, it could have been the US engaging in mass killings and other countries demanding for de-escalation. This is a limit of our data that we have found and further investigation into specific events could show us more about what the data shows as interesting. On that note, the table does give insight into what our algorithms and equations deem as “interesting”, so they do provide a good idea of what an area of further research would be if you wanted to look into these event types more rigorously.

January 2022 - No Specific filtered events / top event type pairs of interest

| Event Type | Associated Event Type | Interest |
|------------------------------------|--|-----------------|
| Express intent to accept mediation | Appeal for material cooperation, not specified below | 0.967 |
| Express intent to accept mediation | Veto | 0.957 |
| Express intent to accept mediation | Reject plan, agreement to settle dispute | 0.952 |

| | | |
|--|---|-------|
| Express intent to accept mediation | Expel or withdraw, not specified below | 0.947 |
| Express intent to accept mediation | Threaten non-force, not specified below | 0.947 |
| Appeal for military protection or peacekeeping | Conduct strike or boycott, not specified below | 0.943 |
| Express intent to accept mediation | Threaten conventional attack | 0.938 |
| Express intent to accept mediation | Demand rights | 0.938 |
| Appeal for military protection or peacekeeping | Express intent to de-escalate military engagement | 0.928 |
| Appeal for military protection or peacekeeping | Appeal for release of persons or property' | 0.923 |

As we can see, all top event type correlations of interest this month had “Express intent to accept mediation” and “Appeal for military protection or peacekeeping” as the event 1 type. From these we can see there was an interesting correlation between these two event types and event types such as “conduct strike”, “boycott”, “demand rights” and “reject plan”. This suggests that the event correlation possibly was one where a marginalized group felt unfairly treated in their situation. The “Appeal for release of persons or property” and “Express intent to de-escalate military engagement” event types also suggest the military was involved. The small range of event types and close correlation to one another suggests it could be related to one event, but not much more assumptions could be made.

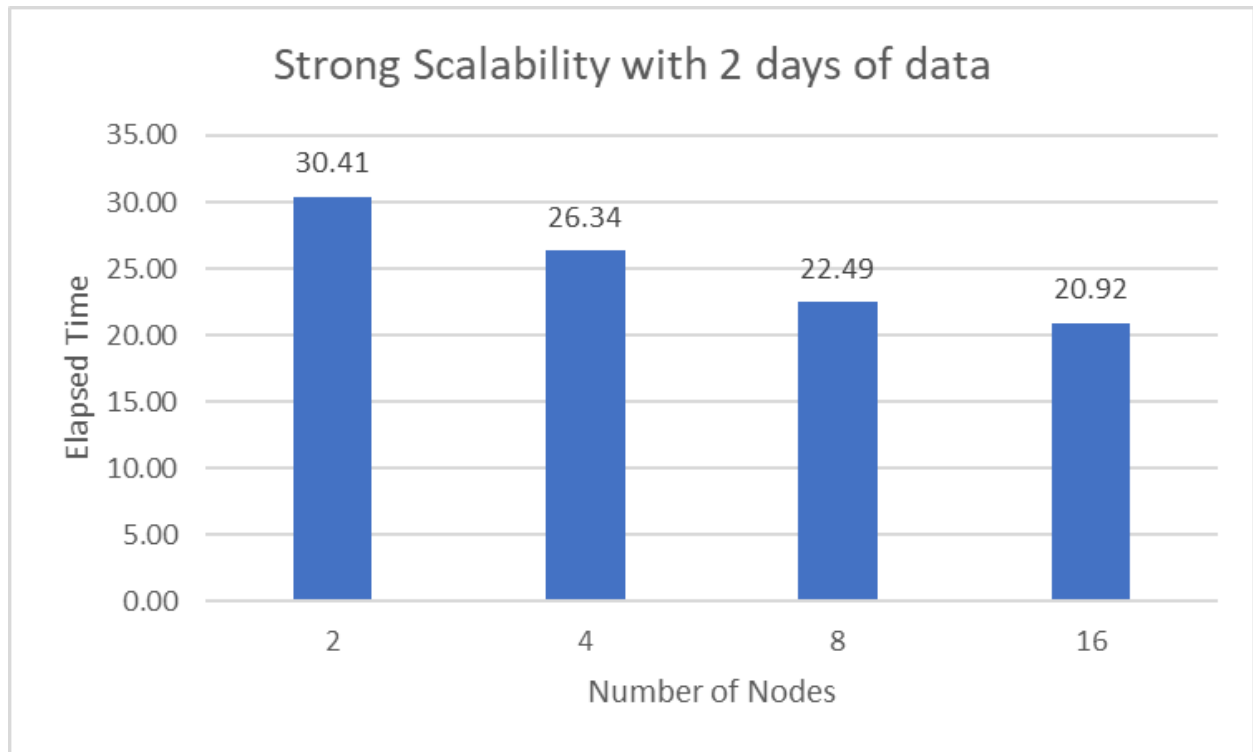
Conclusion of Results:

From my question, “What are the most interesting pairings of types of events in the month of January in the past 3 years where an organisation, government, person/s etc from the United states is involved as the main actor in the event?” I have gathered the top 10 interesting pairings for each month of January in the past 3 years and made inferences from the data. However, any assumptions we made cannot be proven with just the data alone. So although we can see interesting event type correlations for individual months, using more results and algorithms to compare results in months to find trends of interesting event types correlations would be needed to better analyse how the United States conducts themselves internationally and possibly answer a more complex question.

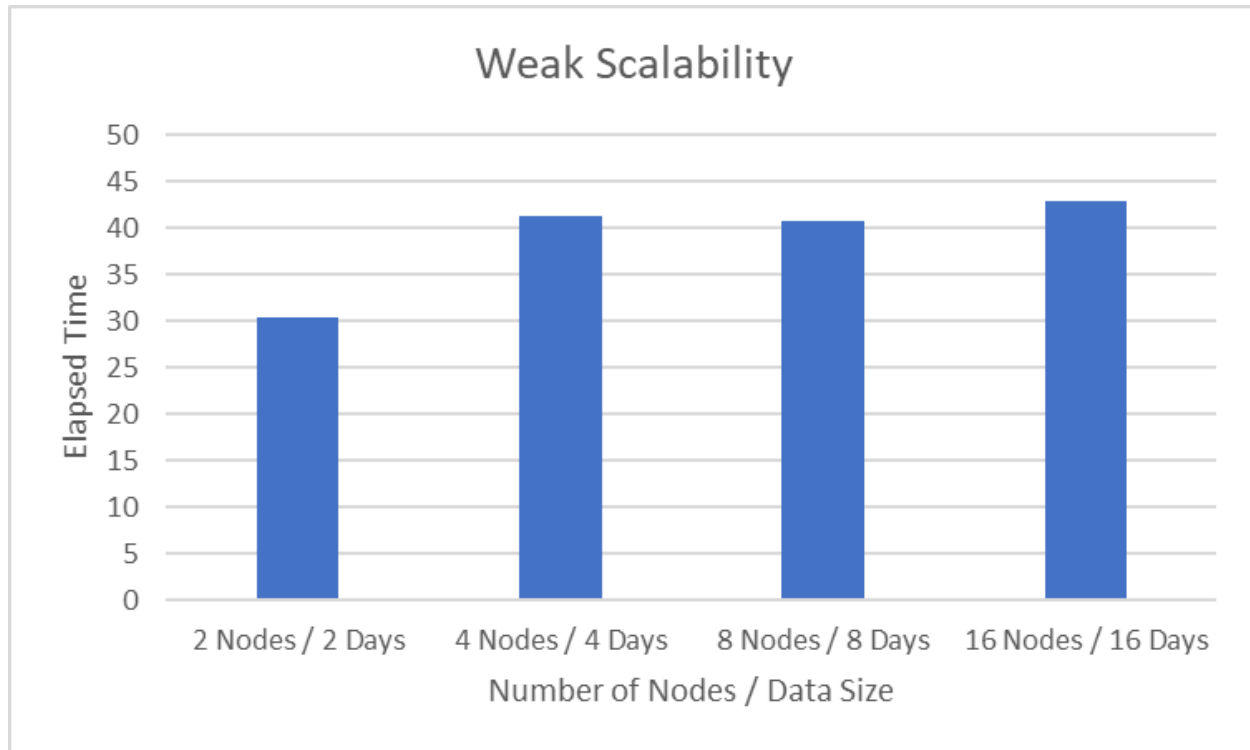
Scalability Graphs:

These Graphs were created using results gathered from Google Cloud and plotted using Excel

Strong Scalability (fixed problem size with an increasing number of processors)



Weak Scalability (problem size increases at the same rate as number of processors)



Conclusion:

From the data gathered and the way it was processed, I was able to give an answer for what were the most event type pairings where the US was involved for each January in the past 3 years. The implication of the results is that we can make quite broad inferences from the event correlations. Our algorithm managed to find interesting event pairings where correlation was at 100%, potentially implying causation for all events which could be taken into account when making inferences. It also found event pairings which were rarer, but not too rare, using >5 “baskets” as a baseline. This means our most interesting event pairings were where only a few countries besides the US were involved, or it is a rarer world event type and separate events resulted in the event types. All this meant we were able to make good inferences about our data, and even drew some small but interesting ideas about how the USA operates abroad. However, it doesn’t fully answer any bigger questions about how the US operates abroad. This could be an area of further research and using more results and algorithms to compare results from different months to find trends of interesting event types correlations would be needed to better analyse how the United States conducts themselves internationally and possibly answer a more complex question like “What event type pairings have been the most common where the USA were involved with over the past 5 years?”. This Question if answered would give very good results and create a better picture of how the USA has conducted themselves internationally in the last few years.

Critique of Design and Project:

A part of my design that could have been better was the way I filtered event types. I used every single event type from the CAMEO database and thus could have greatly reduced my processing time for the later stages where sorts were used if I took the time to filter what event types were worth looking into. I did however build the code to do so, but I did not utilize it in my final project as I felt looking into everything was more valuable than not. In theory I could have gathered only violent conflicts or where aid was given or where demands were made. This although would have given less results, might have given more interesting results where deeper conclusions could have been made about a smaller subset of data.

Reflection:

I found Market Basket Analysis extremely useful to answer my question, as well as the definition of what an interesting item pairing is and how to calculate it. Also using spark and parallelization to work with code so much faster was very useful and the amount of data I worked with would have not been possible without it. Finally the GDELT dataset made everything possible. The main takeaway for me on this project was how many times I had to restart and what it taught me about persistence when doing a project. My initial project pitch did not utilize the algorithm (PageRank) I originally wanted to do very well so I had to start again. My second restart was due to not being about to use the GDELT GKG database due to an error where complete coverage would not work. And the python version required to update gdeltpyr to its latest version so that it would fix the error, (python 3.6) was not compatible with pyspark, which threw my second idea of doing market basket analysis of themes of different countries related to a world event. Besides that, I also learnt how to deal with a large dataset to answer a question, and optimizing my functions, such as removing unnecessary sorts and filtering, to better my algorithm times and make my testing and data gathering more streamlined.

References:

Cameo event type codes & Translation Document

<http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf>

https://drive.google.com/file/d/1AlyqTzz8HR1qoeFkaxTUJ_IUU5UUMNVH/view?usp=sharing


GDELTPyR

<https://github.com/linwoodc3/gdeltPyR>

GDELT 2.0 Events Database

<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>

DATA301 Lab 3 - Association rules

 **DATA301 - Lab 3 2022**

DATA301 Lab 4 - Cloud computing, scalability

 **DATA301 - Lab 4**