DATA301 Final Report

Adam Davidson
43408024

Abstract

For this project, my research question is as follows: What is the worldwide sentiment of the upcoming 2022 FIFA world cup in comparison to both the host country Qatar's sentiment of the tournament, and the worldwide sentiment of the 2018 competition? The subset of data needed for answering this question is in the GDELT Event Database. More specifically, the data will be filtered down using the GDELT DOC API. The algorithm used for the similarity analysis will be cosine similarity, which will be used three times for various tone comparisons of the data. The intended result of this analysis will be to determine if there is negative worldwide sentiment of the 2022 competition in comparison to the 2018 competition, or if negative sentiment is only coming from a few specific countries.

Introduction

Background

To better understand the GDELT data I am using, I need to understand the workings of the GDELT Event database, in which all of the news articles that I wish to examine will be. More importantly, it is vital to understand how the GDELT DOC API works, as this is where I will be refining the dataset to try and get the specific data that will help answer my research question. This includes understanding and making sure I have all of the correct parameters, such as the correct date ranges, the correct keywords, the correct countries, and the correct data mode. I will also need to understand how to exactly format the query so that the data that is fetched is the data that I want.

I will need to understand similarity algorithms, and the way that cosine similarity works to compare sets of items. There are many different similarity algorithms such as Jaccard, which measures distance between vectors as the size of their intersection divided by the size of their union. This is used for binary representations of data however, and in cases of data duplication cosine similarity is a better choice as it measures the cosine of the angle between two vectors to determine whether the vectors are pointing in a similar direction. Since my data will include multiple tone counts for each vector position, cosine similarity is a more suitable choice for comparisons.

Motivation

Football has always been a sport that has interested me, as I have played it up until a few years ago and have watched it all my life. The world cup is something that elevates the spectacle of the sport, which is why it interests me to do research on the topic. The upcoming tournament in Qatar later this year is especially interesting, as it has been controversial for a number of reasons. These include worker conditions and deaths while building the stadiums, corruption allegations when making the bid to host, and LGBT rights concerns for spectators as homosexuality is illegal in the country. There seems to have been a lot more negativity surrounding this tournament than previous competitions, and I wish to find out whether this negativity is largely shared worldwide, or mainly just coming from western english-speaking sources of which I have been reading. The result of this research can help me share my interest in football with others, especially when talking about the upcoming world cup.

Research Question / Hypothesis

My research question: What is the worldwide sentiment of the upcoming 2022 FIFA world cup in comparison to both the host country Qatar's sentiment of the tournament, and the worldwide sentiment of the 2018 competition? This question is relevant for the GDELT Event dataset and DOC API. This is due to the Event dataset looking at news articles and the DOC API being able to refine the news article data down to search for specific parameters that are relevant to news about the World Cups and setting the mode for looking at tone values for sentiment.

The cosine similarity algorithm is very useful at comparing different vectors of items. The first two cosine similarity analyses will be for tone comparisons of worldwide countries to the host country in the respective World Cups. The third cosine similarity analysis will be for tone comparisons between the same countries leading up to the different world cups. These sentiment results will then be able to answer the research question as it includes both sentiment comparisons to the host nation and comparisons to the 2018 tournament.

Experimental Design and Methods

The first step taken in my project code was setting up the spark and GDELT libraries, and also starting the spark context. This was done in the same way as the sample projects. Next, I have listed 22 different queries which fetch data from the GDELT DOC API, from the 11 different countries to be analyzed over the two tournaments. Along with the source country, a number of parameters were placed such as the keywords of "world cup", a year long date range, and the word "qatar" being mentioned at least twice. I also selected the output mode of the API as a tone chart, which will help me determine the distribution of emotion from each source country. The next step I took was to make a csv file for each query and then pull the links from the DOC API and write them to their respective files. Next, for each of the csv files a dataframe was created, along with the name of the country. Two empty RDDs were then created to represent the two world cups, and each dataframe was mapped into their respective one, based on a tuple of each tone, and count value, and also the name of the country the frame came from.

The next steps taken in my project was to prepare the data using pyspark methods for performing cosine similarity on the 2022 World Cup RDD. This was done through a number of steps such as grouping the RDD by tone values, then adding in countries with a tone count of 0 if they were not already included in the respective tone value. The RDDs were then grouped by country and mapped into vectors. These vectors could then be analyzed using cosine similarity between the respective country and the host country Qatar. The results were then sorted into decreasing order to make comparisons easier. These steps of preparing the data and performing cosine similarity were then repeated but for the 2018 World Cup RDD, and comparing the cosine similarity between the host country Russia and each of the other respective countries. Finally, the two RDDs were reformatted and joined together based on the (country, tone) key tuple. This joined RDD was then grouped by each country, and mapped into two vectors for each country representing the two different tournaments. Finally, these vectors were analyzed using cosine similarity between each country for the two different tournaments.

To help implement the pyspark methods listed above, I created a number of helper functions which are listed as follows:

- add_country: Function to help add countries with a tone count of 0 that is currently not being represented by a given tone value.
- country_as_key: Function to reformat all of the countries and counts in a tone value as (country, (tone, count)) tuples.
- vector_country: Function to create a numpy array of tones for each country

- multiple_vectors: Function to create two numpy arrays (for the two tournaments) of tones for each country
- cos_similarity: Function to compute the cosine similarity between a country and the host country
- cos_similarity_2: Function to compute the cosine similarities between a country in the 2018 world cup and the same country in the 2020 world cup

Along with these functions, I also imported the numpy library for arrays and to help compute the cosine similarities.

Results

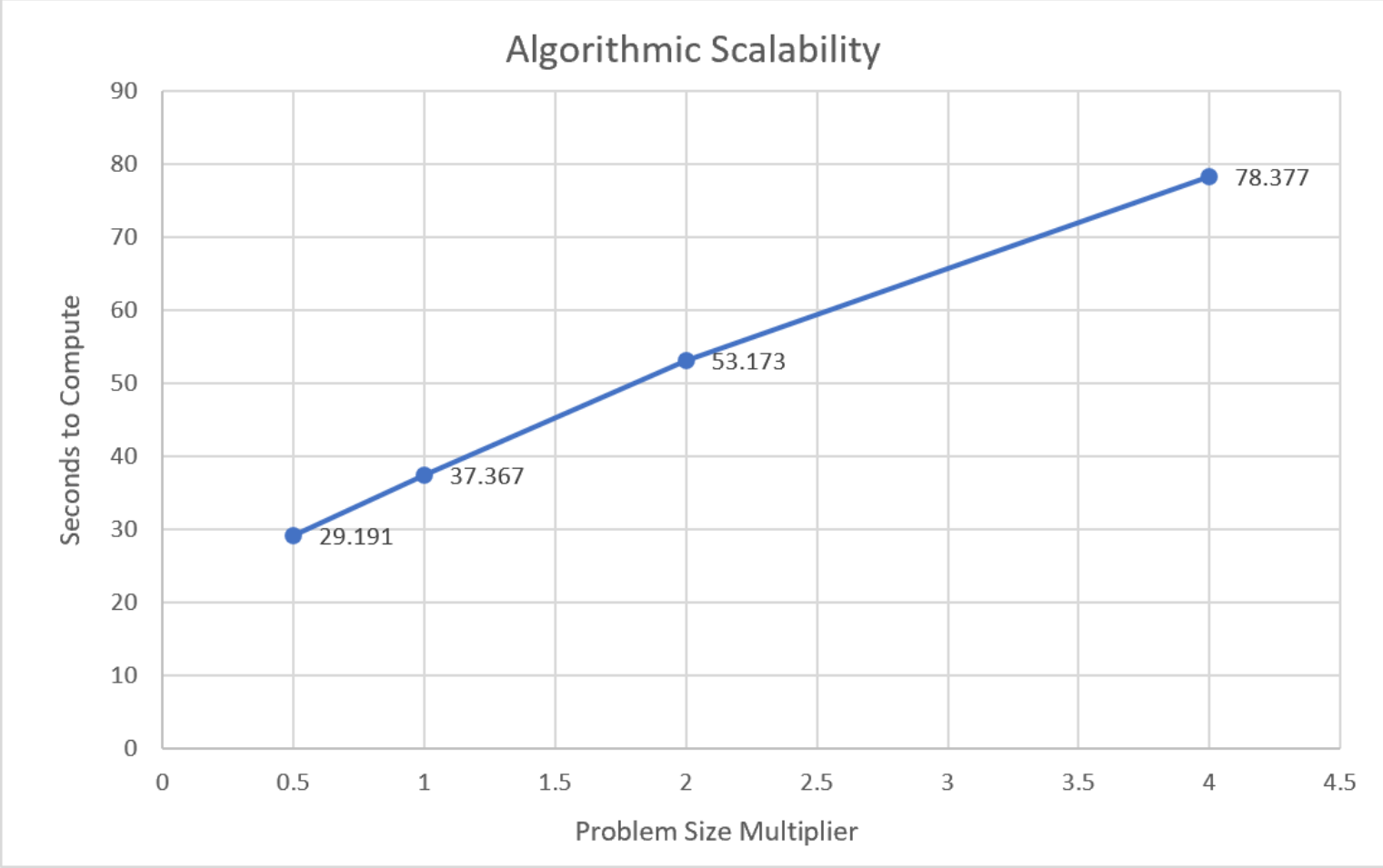

Figure 1: Graph showing the scalability of my solution as the number of processors remains constant and the size of the problem changes. The original problem size has 22 queries, which is represented by the 1 problem size multiplier. The remaining x axis multipliers represent 11 queries for the 0.5 multiplier, 44 queries on the 2 multiplier and finally 88 queries on the 4 multiplier.

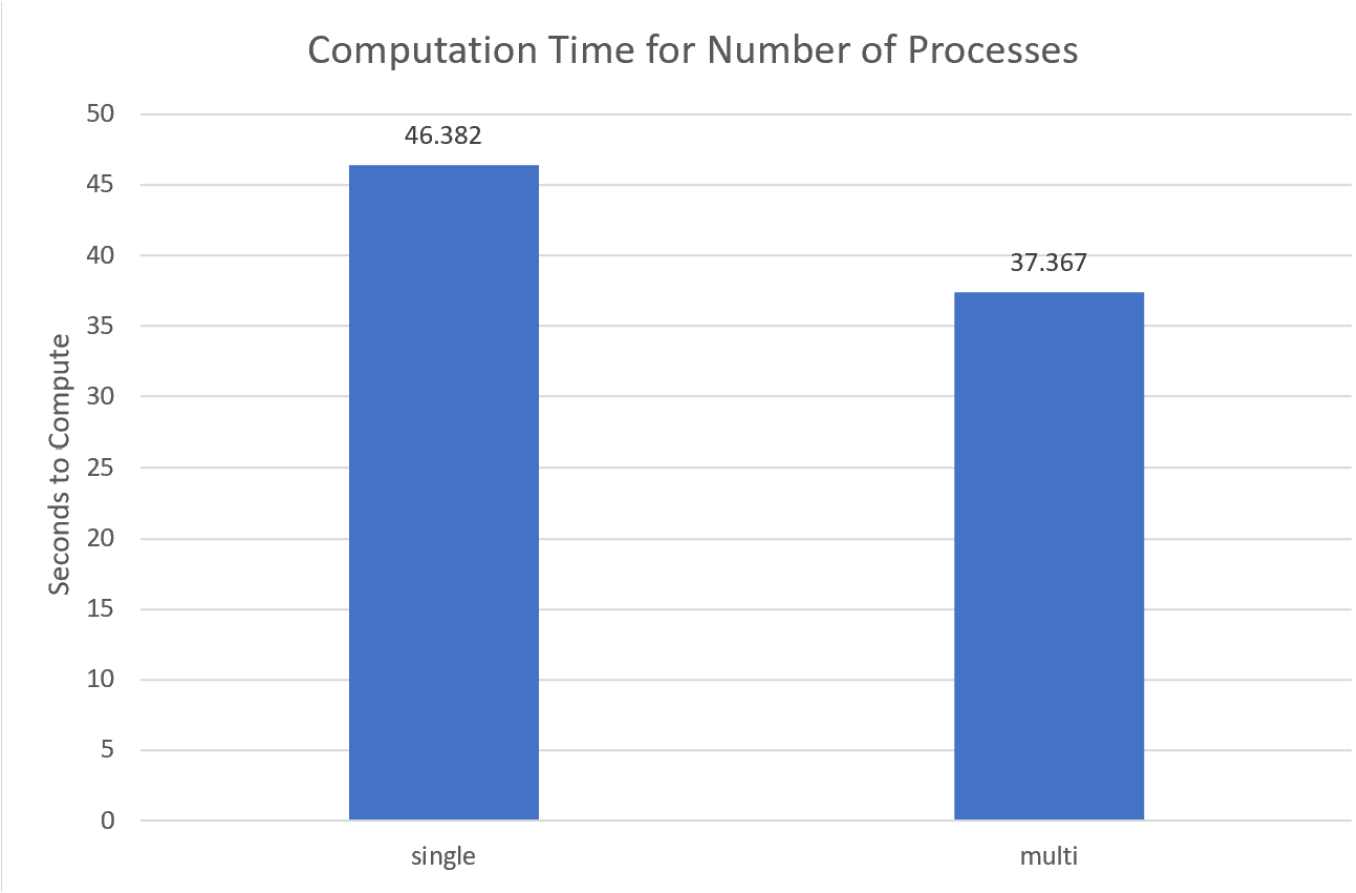

Figure 2: Graph showing the scalability of my solution as the problem size remains fixed and the number of processors changes. The single process represents the solution run with just one worker thread, while the multi process represents the solution running with as many worker threads as there are available.

The first cosine similarity computation compared the tone sentiment of host nation Qatar to that of 10 other countries worldwide for the 2022 World Cup. The results showed Ghana to be the country most similar with a cosine similarity value of 0.890 (3dp). This was followed by Iran (0.79), Brazil (0.746), Argentina (0.671), South Korea (0.616), Senegal (0.539), New Zealand (0.502), United Kingdom (0.48), United States (0.43), and finally Germany (0.396). The results indicate that Qatar does view the 2022 World Cup differently to countries worldwide, but that the difference is much more extreme in certain nations.

The second cosine similarity computation compared the tone sentiment of host nation Russia to that of 10 other countries worldwide for the 2018 World Cup. The results showed the United States to be the country most similar with a cosine similarity value of 0.965 (3dp). This was followed by South Korea (0.96), Germany (0.958), Brazil (0.948), Ghana (0.926), Senegal (0.925), Argentina (0.888), New Zealand (0.878), United Kingdom (0.871), and finally Iran (0.76). These results indicate that Russia viewed the 2018 World Cup largely similarly to other countries worldwide.

The final cosine similarity computation compared the tone sentiments of each country against each other for the different World Cups. The results showed Argentina to be the country with the most similar tone values between world cups with a cosine similarity value of 0.974 (3dp). This was followed by South Korea (0.967), Senegal (0.965), United States (0.949), Brazil (0.942), Germany (0.93), United Kingdom (0.869), Iran (0.77), Ghana (0.76), New Zealand (0.753), and finally the host countries of each tournament (0.624). These results indicate that there largely isn't too much of a difference in worldwide sentiment between the 2018 and 2022 World Cups, but there is a fairly large difference in how the host nations viewed their respective tournaments.

Conclusion

My research question was able to be answered well in some areas and not so well in others. I was able to find out that the sentiment that Qatar has of its upcoming tournament is different to other countries, and that this difference is in fact more extreme in mainly western nations. This supports the notion that countries worldwide generally view the Qatar World Cup more negatively than Qatar does. This negativity however, doesn't seem to have much of a difference between World Cups, as most countries had a high cosine similarity when comparing their sentiments between tournaments. This, along with the fact that the sentiment Russia had of its 2018 World Cup being generally similar to the same countries, could mean that Qatar just has an overly positive outlook on their tournament instead of other countries having an overly negative outlook on it. This point is inconclusive however, as cosine similarity just tells how similar two sets are, and it could also be that both Russia and other countries viewed the Russia World Cup in a negative light as well.

These results have implications for both the sport and society in general. The fact that there is a sizable difference in how the rest of the world views the upcoming Qatar World Cup gives evidence that there are legitimate concerns about the tournament. The fact that the sentiments are shared similarly for the Russia World Cup is not a massive surprise, as there was also a lot of controversy in them holding the tournament, but it does dispel the notion that the Qatar World Cup is unique in its negative view to the world. A more interesting observation from my results was the fact that Russia viewed their own tournament in a similar way to the rest of the world. This could possibly be the result of tensions between Russia and other Governments resulting in more negativity overall on all sides.

A future question for my project would be to extend the comparison between world cups to tournaments before the 2018 one so that a more accurate view could be formed on whether or not the upcoming tournament is actually seen in a negative light. This unfortunately was not possible with the data for the DOC API only available going back to 2017. Another implementation I could have made to my project would be to add another algorithm so that supporting claims could be made around the sentiments and my conclusions. Finally, In the future it would be interesting to analyze the worldwide view of the tournament after it has been played, so as to get comparisons of before and after each respective tournament. This would help me answer further questions such as: Do the controversies of a World Cup stick around after it has been played?

Critique of Design and Project

A part of my design that could have worked better was the repetition of preparation methods for finding the Cosine Similarity between Russia and worldwide countries and that of Qatar and worldwide countries. Instead of initially creating two RDDs, it would be better to map all of the data into one RDD. Instead of repeating the methods to refine and reformat the RDDs separately, it could all be done to the single RDD and having that RDD being split up just before the actual cosine similarity part is implemented. Approaching the implementation from this way would have made it more efficient when processing and executing the first two cosine similarities.

An addition to my design that would have made the project more successful would be another algorithm such as computing a TF.IDF score. This could be implemented by going through the actual articles in my queries for the Qatar world cup, and collecting the words and comparing them with words from articles mentioning World Cups in general using the TF.IDF algorithm. This would further help conclude sentiment surrounding the Qatar World Cup, as it would highlight words that are used a lot when mentioning the tournament in comparison to other World Cups.

Reflection

I found many concepts learnt throughout the course useful for my work in the project, along with other resources I found. The course lectures and notes on similarity were the main source of concepts I used for my project, and also online sources for a more deeper understanding of cosine similarity. The GDELT documentation and blogs were also extremely helpful for learning exactly how to fetch the material I wanted for my project. While completing this project, I have learnt a great deal more about similarity algorithms, scaled data processing using PySpark, and about GDELT and all the analysis that is possible with the data they provide which was previously unknown to me.

References

- GDELT DOC API blog/documentation: https://blog.gdeltproject.org/gdelt-doc-2-0-api-debuts/

- More GDELT DOC information I used to help set up my queries: https://drkblake.com/gdeltintro/

- Lab 5 notebook in where I got help with the cosine similarity algorithm: https://colab.research.google.com/drive/1JW7wEz1PoYLQ37Tp_PyLE1cSYeu8wfDy?usp=sharing

- Lab 2 notebook in which I got the idea for the potential of using a TF.IDF algorithm: https://colab.research.google.com/drive/1l9r-rlP0g_aR4tNesxin8ImNvL-907xs?usp=sharing

I have not worked directly with any other student on the project.