Department of Computer Science and Software Engineering

DATA301 Project Final Report

What is the similarity in average sentiment of health-related events over the last few years between New Zealand and other countries?

Lily Williams
42415299
lfw25

June 1, 2022

Abstract

Using data taken from the GDELT 2.0 Event Database, the similarity in average sentiment between New Zealand and other countries was determined. This was done by pulling 222 average tone CSV files, spanning 6 countries and 37 months from March 2019 to April 2022, from the GDELT database, filtering the data using Map-Reduce functions, determining each country's tone in a given month and putting that into a vector, and calculating the cosine similarity of the tone vectors. It was found that New Zealand had a very similar tone progression to the 5 other countries (US, UK, China, Russia, South Africa), with similarities ranging from $S \approx 0.916$ for the US to $S \approx 0.983$ for China. It was also found that between the 6 countries, the lowest similarity was $S \approx 0.881$ between South Africa and the Unites States, and the highest similarity was $S \approx 0.984$ between Russia and China. These results suggest that the global media response to COVID-19 has largely been consistent between countries and cultures. Additionally, this analysis was also run on parallel computing cores to determine its speedup response to parallelism and thereby investigate the efficacy of the code construction. The code responded poorly to being parallelised, taking longer to execute on 2, 8, and 16 nodes than on a single node. This result was verified by running an asymptotic algorithmic scalability test where the amount of data was varied while keeping the number of nodes the same. Overall, this suggests that there is significant forced sequentialism in this method that is not overcome by splitting the task onto more nodes.

Contents

1 — Introduction

The COVID-19 pandemic continues to overwhelm countries, economies, healthcare systems, and the general psyche of human culture. In New Zealand, the pandemic response was initially received as bold but wise restrictions on travel, work, and general life, and many people felt like the general feeling was one of unity and cooperation. However, now, two years later, it feels as if most people view it as a thing of the past, and continued restrictions feel like a cage preventing people from moving on. It seems like the belief in the advice of governments and healthcare professionals has gone from the obvious choice to feeling like an unnecessary burden. Investigating the perception of health events and health-related news like COVID-19 over the last few years, both in New Zealand and overseas, could reveal unique insights into the similarities and differences in how these events are being received amongst people of different countries. This task would require access to vast amounts of data, and methods that allow the similarity between different sentiments to be analysed. Fortunately, combining similarity and sentiment analysis methods with the data available in the GDELT database will allow the research question to be answered.

GDELT is the world's largest high-resolution open source database containing over a 250 billion event records, covering from 1979 to the present day. It was created by Kalev H. Leetaru from a dream to better understand society and the events that shape it. Today, it is sourced from hundreds of thousands of news sources in different formats from all over the world. The GDELT database features a summary API which allows users to search full-text articles with a human-friendly interface. Using this GDELT event database, the impact of the COVID-19 pandemic can be analysed. By using the summary function, it is possible to search news articles for mentions of the pandemic, and to then analyse how news agencies around the world have been portraying the crisis from its beginning to the present day.

Grouping the events by months, the general portrayal on a numerical scale for each country can be averaged. This is a concept known as sentiment analysis, and, over time, can show the progression of how a particular issue is perceived or portrayed. This sentiment can be compared between a list of countries using an algorithm known as cosine similarity. Cosine similarity is a mathematical similarity measurement between two vectors, where each vector represents the sentiment progression of a country, obeying the formula

$$S = \frac{u \cdot v}{||u|| * ||v||} \tag{1.1}$$

Where S is the similarity between vectors, u and v are sentiments of countries represented as vectors, and $||u||$ and $||v||$ are the magnitudes of each sentiment vector.

The research question to answer in this project is this: what is the similarity in average sentiment of health-related events from March 2019 to April 2022 between New Zealand and other countries? The countries investigated will be New Zealand, USA, UK, China, Russia, and South Africa. By formatting the average sentiment of health-related events each month of different countries as a vector, the cosine similarity algorithm can be used to find how

similar each country's sentiment is to New Zealand's. This similarity can show how each country is interpreting health-related events, particularly over the course of the pandemic, and how different that progression is to that of New Zealand.

Additionally, this project will use the concepts of parallelism to investigate the speedup response of adding more cores to the analysis process. Parallelism is the concept of strategically splitting a task up into smaller parts and having those tasks be solved simultaneously. These smaller solutions can be pieced back together afterwards, resulting in a dramatic reduction in the time taken to solve the overall problem. The proportion of time saved using parallelism is calculated using Eq. 1.2, where Z is the speedup ratio, T_s is the time taken for the analysis to be completed sequentially (with only one core), and T_p is the time takes for the analysis to be completed in parallel (with more than one core). The less dependence on the order the sub-tasks need to be completed in, the more speedup can be attained.

$$Z = \frac{T_s}{T_p} \tag{1.2}$$

2 — Methods

Data Flow and Analysis

The project made use of the GDELT 2.0 Summary feature, which allows the user to search for specific keywords appearing in news reports, enabling filtering by country and sorting data into monthly chunks; the summary has a built-in tone function which buckets the approximate tone of the articles in the specified period of time, producing an easily readable and, most-importantly, a downloadable CSV which contains these buckets. Figure 2.1 displays an example of this tone chart. The keywords for this example, and the project, were "health -mental", the country was New Zealand, and the date was 01 March 2019, the first day of my proposed sample range. Table 2.1 shows a small section of the data contained within the associated CSV file.

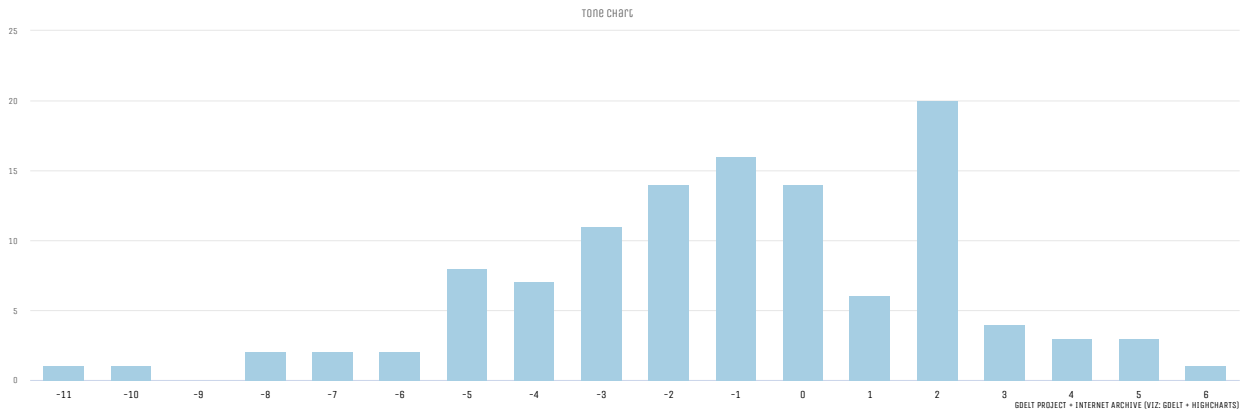


Figure 2.1: This is the tone chart of health-related events on 01/03/2019 in New Zealand.

Table 2.1: A small section of the data contained within the CSV for the example tone data shown in Fig. 2.1.

Label	**Count**	**TopArtURL1**	...
-11	1	https://www.nzherald.co.nz/...	...
-10	1	https://thedailyblog.co.nz/2019/03/01/...	...
-9	0		...
-8	2	https://livenews.co.nz/2019/03/01/...	...
-7	2	https://www.stuff.co.nz/national/crime/110967765/...	...
-6	2	https://i.stuff.co.nz/national/education/110954487/...	...
-5	8	https://www.nzherald.co.nz/nz/news/...	...
-4	7	https://www.nzherald.co.nz/northern-advocate/...	...
-3	11	https://www.stuff.co.nz/national/110958855/...	...
...	...	...	...

Using GDELT's Summary feature, individual CSV files were acquired for each country and month within the desired date range. This was be done by editing the URL to contain the

appropriate country code, keywords, and date range, and then saving the resulting tone file as a CSV. The country codes were "NZ", "US", "UK", "CH", "RS", and "SF", the keywords were "health -mental", and the dates ranged in sections of one month from "20190301000000" to "20220401235959" in the YYYYMMDDHHMMSS format. Each country and month had its own RDD containing the CSV files of the tone buckets. For 6 countries and 37 months each, there were a total of 222 RDDs created. Once these files were pulled from GDELT's database, each was be parallelised and mapped to only retain the first two columns of data, "Label" and "Count". Another map found the product of these two columns, summed these up over every column in the CSV, and divided by the total number of articles, giving an average tone for each month. Finally, this average tone was rounded to an integer on a scale from 1 to 5 aligning with the general tone of the articles in that month, with 1 being "Very Negative" and 5 being "Very Positive". From there, the tones for each country were sorted into a vector, then the similarity, S, between New Zealand and the five other countries' tone vectors was calculated using the cosine similarity formula (Eq. 1.1). Another vector was made of all the tones for each country without aligning them to the aforementioned integer scale in order to plot the progression of each country in more detail, and to more visually represent the similarities in sentiment progression between the 6 countries.

Required Functions

The following libraries were imported for use in this analysis:

- PySpark: SparkConf, SparkContext, SQLContext
- OS, sys
- concurrent.futures, PocessPoolExecutor
- datetime, date, time, timedelta
- pandas, numpy, matplotlib
- urllib.request
- operator

The following functions were defined and used in this analysis:

- getFilename
  - Formats the date and returns a filename to save the target CSV as.
- intoFile
  - Pulls down the tone data and saves it in the given filename
- getTone

  - Takes a dataframe containing the CSVs of each country, uses Map-Reduce algorithms to find the weighted average tone for each CSV, returns a list of the average tones for each CSV.

- toneOnScale
  - Takes a float of the tone, matches it to an integer tone on the scale, returns the integer tone.

- cosineSimilarity
  - Finds the similarities between two vectors, returns a float representing the similarity.

Parallelism and Speedup

In order to analyse the effect of parallelism on this problem, it was run on the Google Cloud distributed computing Platform. Excluding the downloading of the GDELT files, the Map-Reduce and similarity analysis were timed for 1, 2, 4, 8, and 16 nodes. This speedup was calculated using Eq. 1.2, where T_s was the time with 1 core and T_p was the time taken with 2, 4, 8, and 16 cores. These ratios were graphed to determine the effectiveness of parallelism on this data set. In order to verify this data, a more "traditional" asymptotic algorithmic complexity analysis was also completed, in which the size of the data set was changed while keeping the number of nodes the same. For this, the number of months of data for each country was varied through 2, 6, 12, 18, 24, and 37 months while keeping the number of nodes the same at 1, and the time taken for the analysis was recorded. Additionally, using the full dataset, the number of local nodes within Google Colab was changed from 1 to "*" and the times compared. Having these four metrics allowed the full scalability of this program to be determined.

3 — Results and Discussion

Dataset Analysis

After running the previously mentioned analyses on the data set, it can be shown that the similarity in sentiment progression between New Zealand and the other six countries is quite alike. This is displayed in Fig. 3.1.

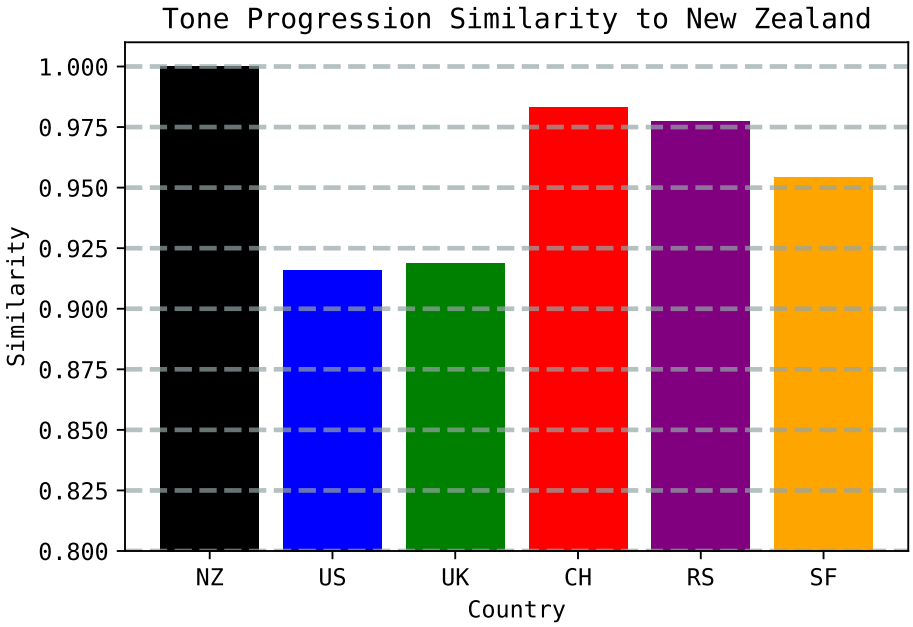


Figure 3.1: It can be seen in this bar chart that every other country had notably similar sentiment progressions to New Zealand.

It is clear that all of the other six countries have sentiment progressions very similar to that of New Zealand. The most similar progression was China, with $S \approx 0.983$ and the least similar was the United States with $S \approx 0.916$. These are still very similar progressions, which suggests that the way that health-related events have been communicated has changed in a similar way in each of the six countries. This is even clearer in Fig. 3.2, where the shapes of these tones very closely match.

It can be seen that the peaks and troughs occur at similar positions; most notably, the significant drop 11 months after March 2019, February 2020, is present in every country's sentiment progression. Clearly, the similarity in sentiment progression between New Zealand and the five other countries is quite alike. Out of interest, the similarities between each country and all the other countries was also computed. These progression similarities are shown in Figs. 3.3.

The two countries with the most similar progressions were Russia and China, with progression similarity of $S \approx 0.984$, and the two countries with the least similar progressions were South Africa and the United States, with progression similarity of $S \approx 0.881$. These are still fairly similar though, which makes this result surprising, as the six countries were

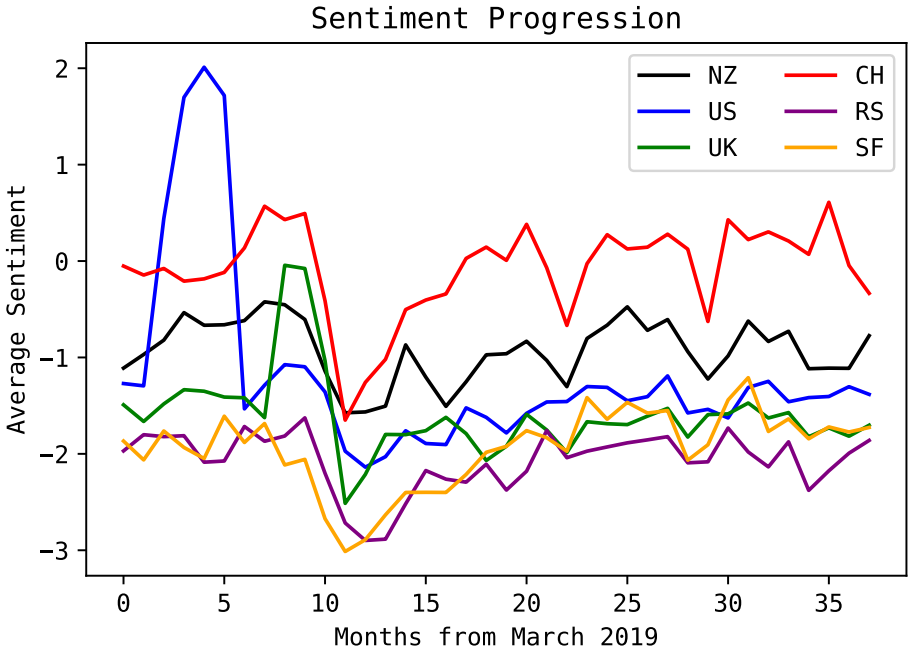


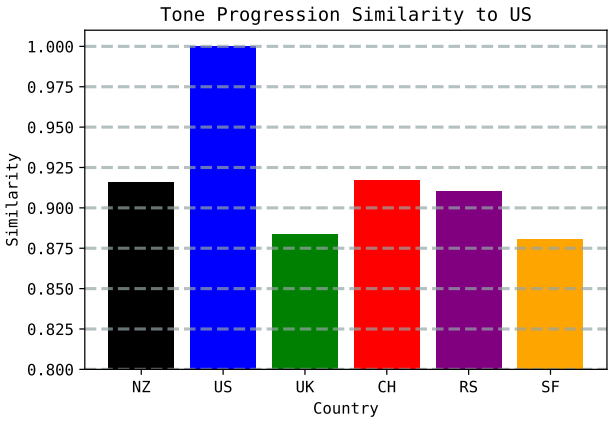
Figure 3.2: It can be seen in this bar chart that every other country had notably similar sentiment progressions to New Zealand.

chosen for their range in political and cultural beliefs. The sentiment progressions being so similar could suggest that, since COVID-19 had a similar devastating effect on every country, the way health-related events are perceived over this same course of time is inherently linked to these effects. It could also imply that different cultures find similar events interesting, so media are more likely to report on similar events in a the same manner to retain engagement. This second explanation seems unlikely though, given the cultural differences between the chosen sample countries. In order to more thoroughly investigate this result, the experiment should be repeated with more sample countries, to increase the cultural diversity, and with a larger date range, to track the progression outside of the COVID era.

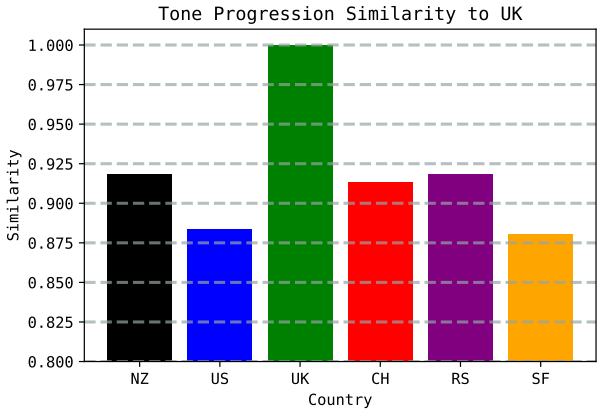
Parallelism and Speedup

Running this code (minus the downloading of the data from the GDELT database) through the Google Cloud Platform, changing the number of nodes, and recording the time taken for the analysis allowed the speedup times to be analysed. Figures 3.4 and 3.5 show that the code did not respond at all well to parallelism; it actually took longer for the analysis to be completed on 2, 8, and 16 nodes than on 1 node.

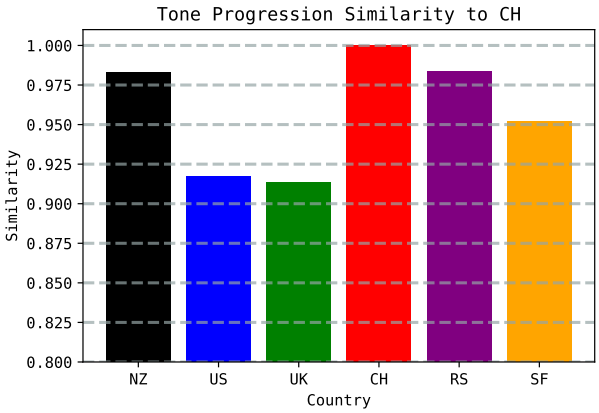
This indicates that there is a significant amount of compulsory sequentialism in the code. This is code that cannot be parallelised and thus will not reduce in time when more nodes are added. This is very likely in the creation of the RDDs, which is necessary for the parallel computing to be complete but also cannot be parallelised. Creating the 222 RDDs using iterative (and therefore sequential) for-loops created a dependancy that will not be overcome by adding more nodes; while the actual processing of the CSV files using Map-Reduce functions is reduced in time, this was already a much more inexpensive computation



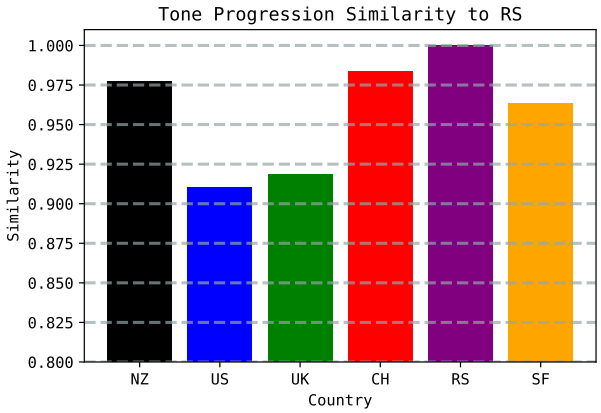
(a) Similarities to the United States



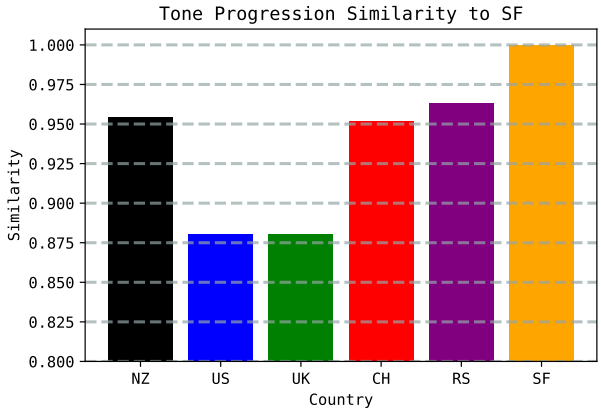
(b) Similarities to the United Kingdom



(c) Similarities to China



(d) Similarities to Russia



(e) Similarities to South Africa

Figure 3.3: Tone Progression Similarities between each country.

than forming the RDDs from the CSV data frames. In fact, it is possible that forcing the code to be run on more cores than is necessary is the reason that the time taken actually increased beyond 4 cores. The overhead cost of splitting the tasks up between so many nodes seems to be much more costly than any time saved by running the calculations in parallel. Overall, this indicates that the construction of the code to answer this specific research question with this dataset is not suited to being parallelised. It is possible that

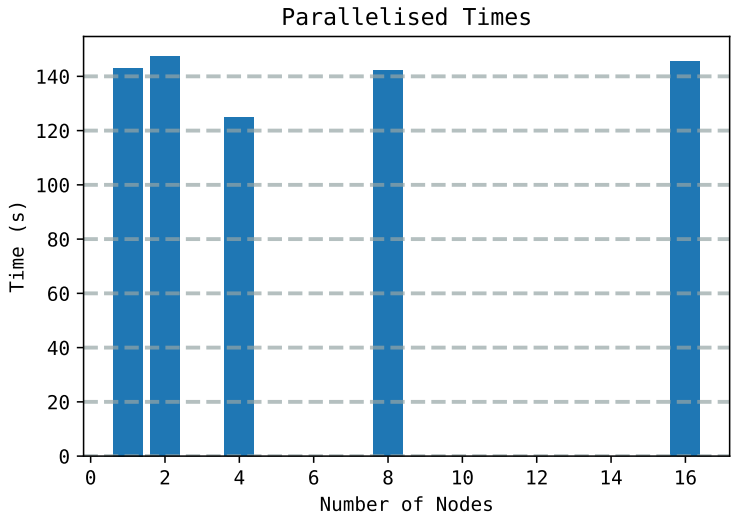


Figure 3.4: It can be seen in this bar chart that the time taken with more nodes does not steadily decrease.

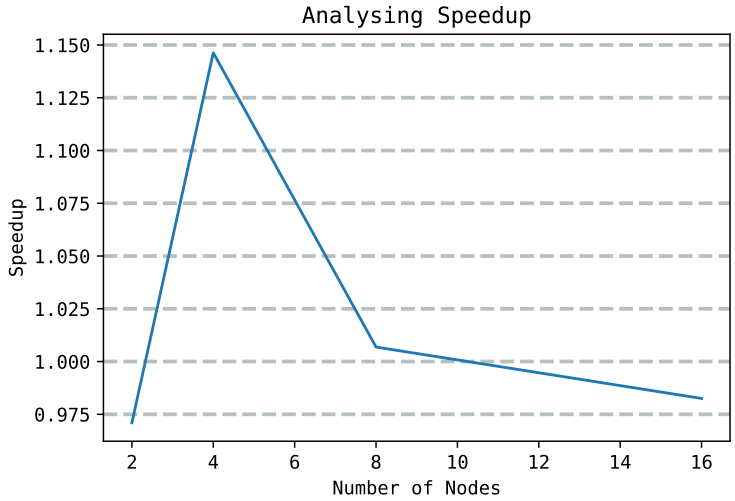


Figure 3.5: This graph shows that the code did not respond well to parallelism.

using a different dataset, perhaps with larger CSV files or a smaller date range, would result in the speedup being more significant with more cores, so it may in fact result that this method would pay off when scaled up significantly.

This particular result is quite unusual though, so more tests were run using Google Colab rather than the Google Cloud Platform in order to determine validity. Figure 3.6 shows that when the number of processes remains the same, there is a relatively linear increase in processing time in relation to the amount of data. However, Fig. 3.7 shows that the time taken to process all the data (222 RDDs) is actually faster with 1 node than with all the nodes (* Nodes). This backs up the idea that this code contains too much sequential code that isn't able to be parallelised, and further suggests that the code construction could be significantly optimised in order to reduce the time taken for analysis.

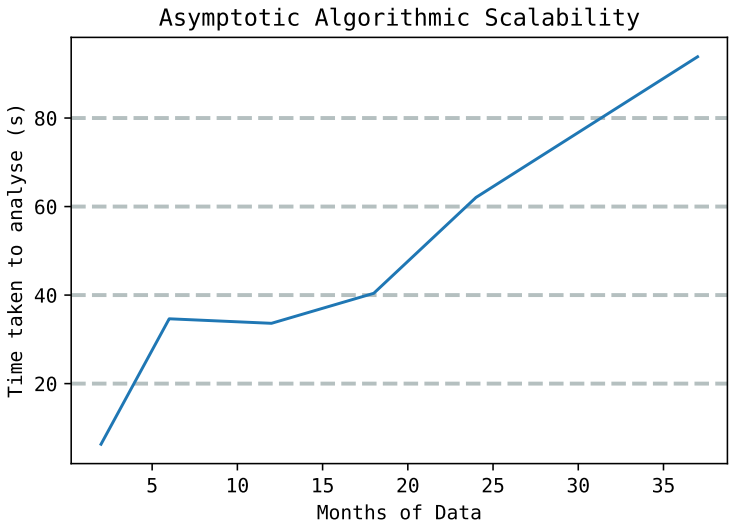


Figure 3.6: It can be seen in this graph that there is a relatively linear relationship between time taken to process and amount of data.

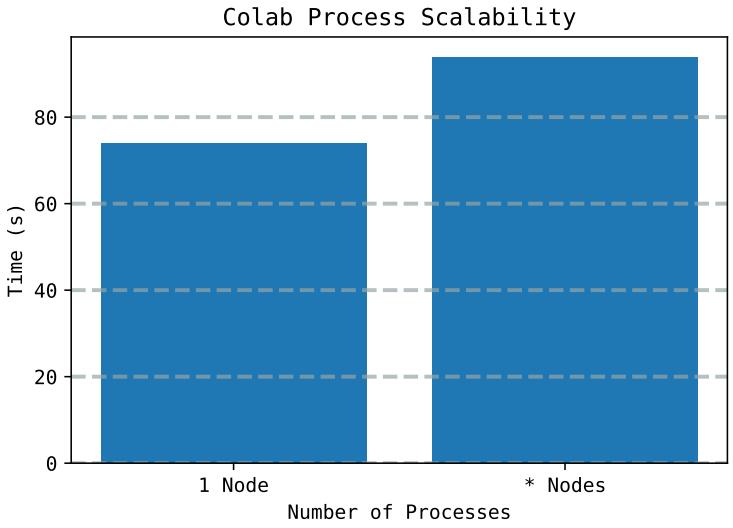


Figure 3.7: This bar graph shows that the code did not respond well to parallelism, and in fact it was faster to process all the data with 1 node than with all the nodes.

4 — Conclusion

The aim of this project was to ascertain how similar New Zealand's sentiment progression to health-related events was to the progression of five other countries: the United States, the United Kingdom, China, Russia, and South Africa. This was done using parallelised data methods, fine-grained sentiment analysis concepts, and cosine similarity. It was found that the sentiment progressions were indeed very similar; the most similar was China with a similarity of $S \approx 0.983$ and the least similar was the United States with a similarity of $S \approx 0.916$. This was reinforced by plotting the raw[1] sentiments, where it was observed that, for the most part, the progressions were aligned: sharing peaks and troughs in the same places throughout the time scale. The research question was therefore answered using the methods described above.

It was also found that this method responded poorly to parallelism, with very limited speedup. The analysis actually took longer to complete on 2, 8, and 16 nodes than it did on 1 node. This was likely due to the significant amount of compulsory loops in the code, which is forced sequentialism that cannot be parallelised. Unfortunately, this is because this method required each CSV to be put into an RDD, which in itself is a very time-consuming process compared to the Map-Reduce analysis that is sped-up by parallelism. There is a lot of room for improvement in this method to reduce the necessity of so much sequential code.

In future, however, it would be interesting to extend this analysis to include more countries, and to perhaps consider a wider time range, in order to see if this sentiment similarity is a cause of the COVID-19 pandemic and it's wide-ranging effects, or if it is an inherent similarity in the way the media portrays health-events. It would also be worth investigating if using larger CSV files results in parallelism being more worth the overhead cost, or if this method just has inherently bad scalability as well.

[1] That is, the value of the tone before it had been aligned to the integer scale.

5 — Critique of Design and Project

One major issue with the design of this project was the $\mathcal{O}(n^2)$ analysis of the CSV files. The parallelism was severely limited by having to run the process over each country and each CSV file, and resulted in significant sequentiality that was not reduced by splitting the tasks between more nodes. A related issue was the creation of 222 RDDs, which is incredibly inefficient and time consuming. Having one RDD per CSV per country resulted in the single longest section of code also being the most integral. The asymptotic complexity of each section of code is listed in Table 5.1 below.

The only part of the code which has the time reduced is a subsection of the *Analyze Tone Files* section. Specifically, it is within the **getTone** function that the Map-Reduce processes, which are the only processes which benefit from the parallelism, are. This specific subsection is the only part whose effective complexity is reduced, but clearly the complexity of the overall program is still quite slow. Particularly, the parallelism doesn't prevent the *Analyze Tone Files* section from being $\mathcal{O}(n^2)$, as the loops which iterate through the CSV files are an element of forced sequentialism and thus are not affected by the parallelism. In future, it would be much more useful to find a method that did not require so much iteration. Since the *dataframe.rdd* function, which converts each CSV into an RDD, is so essential yet does not respond at all to parallelismis used so often in the method used here, perhaps using more Map-Reduce functions to filter and sort the CSVs as one large data frame rather than as dozens of smaller ones would respond better to parallelism and thereby take less time for execution. This change would also greatly improve the scalability of the process, which would allow the proposed future extensions to this research question, mentioned in the section above, to be more efficiently completed.

Table 5.1: A table describing the sequential asymptotic complexity of each section of the code.

Section	**Function**	**Complexity**
Setup Libraries	Imports all necessary libraries and installs pyspark	$\mathcal{O}(1)$
Pull Files	Pulls down 222 tone CSVs and saves as a specified filename	$\mathcal{O}(n^2)$
Analyze Tone Files	Get a tone vector using Map-Reduce algorithms	$\mathcal{O}(n^2)$
Find Similarity	Find the similarity between the tone vectors for each country	$\mathcal{O}(n)$
Plot Results	Plots the similarities and tone progressions using MatPlotLib	$\mathcal{O}(n)$

6 — Reflection

The most useful tools from the course were those of the Map-Reduce algorithms learned in class. They allowed the CSV files to be more efficiently analysed without having as much iteration. The other main tool used in this project was cosine similarity. This was essential in computing how similar the progression of the sentiment was between New Zealand and the other countries.

The most significant takeaway from this project was the investigation around the best way to parallelise code. The method used in this project worked for the question I wanted to answer but had poor scalability and did not respond well to parallelisation. Upon reflection, it was clearly not the best method; it was, however, a good insight into code construction and did force me to really think about what I was doing and how it could be done better. It made me think about how not every algorithm or program benefits from parallelisation, and, while it is a useful concept that often reduces the time taken, it still requires careful thought and planning on the construction of the program to make best use of the concept.

References

Atlas J.(2022). *Sample GDELT Project*, Accessed: 11 May 2022
Available: https://colab.research.google.com/drive/1sTsl_-f2ipgzqM6htsVdKjf4MZ3Ds2CW

Atlas J.(2022). *Sample GDELT Starter Only*, Accessed: 13 May 2022
Available: https://colab.research.google.com/drive/1hXAeG6yheFUQiHfc9Z5ISfNBqAQw47Dq

timpone (User) (2012, January 17). *Convert UTF-8 with BOM to UTF-8 with no BOM in Python*, Accessed: 31 May 2022
Available: https://stackoverflow.com/questions/8898294/convert-utf-8-with-bom-to-utf-8-with-no-bom-in-python

The GDELT project. (n.d.). *Google BigQuery Sample Queries.* Accessed: May 3, 2022
Available: https://blog.gdeltproject.org/google-bigquery-gkg-2-0-sample-queries/

Gupta, S. (2018, January 19). *Sentiment analysis: Concept, analysis and applications.* Accessed: May 3, 2022
Available: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Lærd Dissertation. (n.d.). *How to structure quantitative research questions* Accessed: May 3, 2022
Available: https://dissertation.laerd.com/how-to-structure-quantitative-research-questions-p2.php

Leetaru, K. (2016, January 13). *Mapping world happiness and conflict through Global News and image mining.* Forbes. Accessed: May 3, 2022
Available: https://www.forbes.com/sites/kalevleetaru/2016/01/13/mapping-world-happiness-and-conflict-through-global-news-and-image-mining/?sh=78c05923e224

MonkeyLearn. (n.d.). *Sentiment Analysis Guide.* Accessed: May 3, 2022
Available: https://monkeylearn.com/sentiment-analysis/

Thematic. (n.d.). *Sentiment Analysis: Comprehensive Beginners Guide.* Accessed: May 3, 2022
Available: https://getthematic.com/sentiment-analysis/

GDELT (2022). Accessed: 10 May, 2022
Available: https://api.gdeltproject.org/api/v2/summary/summary?d=webt=summaryk=health+-mentalts=customsdt=20190301000000edt=20190302235959fsc=NZsvt=zoomstc=yessta=listc=1