Realtime and Big Data Analytics Project

Jiraphon Yenphraphai (jy3694)

November 23th, 2022

1 Introduction

I was assigned to work on Health Data, and I chose to work on Google Health COVID-19 Open Data, with the goal of determining how sentiment analysis on Tweets relates to COVID-19 and weather. For example, we can learn how frequently specific words are used in relation to COVID-19 and its symptoms. This Google data set contains the most comprehensive set of COVID-19 information. It contained over 20,000 locations gathered from governments, universities, media, and publications between the January 1st, 2020 to September 15th, 2022. This public dataset provides several aspects, but we chose only three: epidemiology, Google search trends, and government response.

2 Datasets

2.1 Epidemiology

This dataset contains the information related to the COVID-19 infections for each location at each date and can be downloaded here. The table is in csv format and its schema is defined as follows:

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location (country_state_county)	US_NY_36029
new_confirmed	new cases confirmed	1234
new_deceased	new death	12
new_recovered	new recoveries	23
new_tested	new COVID-19 tests	43
cumulative_confirmed	cumulative sum of confirmed cases	1234
cumulative_deceased	cumulative sum of death cases	1234
cumulative_recovered	cumulative sum of recovery cases	1234

Table 1: Table Schema

date	location_key	new_confirmed	new_deceased	new_recovered	new_tested	cumulative_confirmed	cumulative_deceased	cumulative_recovered	cumulative_tested
2020-01-01	AD	0	0			0	0		
2020-01-02	AD	0	0			0	0		
2020-01-03	AD	0	0			0	0		
2020-01-04	AD	0	0			0	0		
2020-01-05	AD	0	0			0	0		
2020-01-06	AD	0	0			0	0		
2020-01-07	AD	0	0			0	0		
2020-01-08	AD	0	0			0	0		
2020-01-09	AD	0	0			0	0		

Figure 1: Snippet of epidemiology dataset

This dataset is depicted in Fig. 1. This file is 520.9 MB in size and contains 12,525,826 records.

2.1.1 Data cleaning

The dataset is mostly cleaned, but some data is missing. We mark the missing data with the value -1 to avoid including it in the reducer code.

2.1.2 Data profiling

We only use two columns, new confirmed cases and new recovered cases, because we believe these data are most closely related to the tweets data, where people begin to complain when the number of new cases increases rapidly and people begin to react positively when there are new recovered cases.

We only focus on cases in the United States, so our mapper filters out information from other countries and only keeps data from the United States. Our mapper reads the csv file line by line, just like a text file. Each line is separated by a comma. If any of the attributes are missing, we mark them as -1. The mapper returns date as the key and a tuple of all record attributes as the value. Within a day, our reducer must summarize all confirmed cases and recovery cases from all states in the United States.

```
2020-03-05      214,0
2020-03-06      350,0
2020-03-07      393,0
2020-03-08      522,0
2020-03-09      989,0
2020-03-10      1265,0
2020-03-11      1403,0
2020-03-12      2362,0
2020-03-13      3344,0
2020-03-14      3715,0
2020-03-15      5484,0
2020-03-16      8525,0
2020-03-17      10477,0
2020-03-18      13836,0
2020-03-19      18617,0
2020-03-20      22376,0
2020-03-21      22122,0
2020-03-22      26764,0
2020-03-23      33296,0
2020-03-24      36605,0
2020-03-25      42130,4
2020-03-26      54800,44
2020-03-27      59386,89
2020-03-28      57405,75
2020-03-29      62617,110
2020-03-30      61826,32
2020-03-31      75919,100
2020-04-01      80394,240
2020-04-02      88976,162
2020-04-03      97250,413
```

Figure 2: Snippet of mapreduce output

Figure 2 depicts a portion of the output of our mapreduce task. The first column contains the date, the second column contains the number of infected cases per day, and the last column contains the number of recovered cases. This corresponds to the news that the number of cases began to rise rapidly in March 2020.

2.2 Google search trends

This dataset is available for download here. It's a compilation of Google searches for symptoms and health problems. The information includes hundreds of symptoms ranging from abdominal obesity to Hepatitis. The searches, like the epidemiology dataset, are mapped to each symptom and organized by date and region. The table schema (over 200 columns) is too large to fit in this report. Only the following columns are displayed:

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location (country_state_county)	US_NY_36029
symptom_name	normalized search volumes of a symptoms	21.01

Table 2: Table Schema

This dataset is illustrated in Fig. 3. This file is 1.9 GB, and it contains 2,713,930 records. What does the score really means? It is the relative popularity of symptoms in searches within in that region for that day, compared to the maximum value.

date	location_key	search_trends_abdominal_obesity	search_trends_abdominal_pain	search_trends_acne	search_trends_actinic_keratosis	search_trends_acute_bronchitis
2020-01-01	AU	3.56	5.38	10.76	0.52	0.35
2020-01-02	AU	3.46	5.35	11.3	0.55	0.36
2020-01-03	AU	3.4	5.35	11.23	0.53	0.34
2020-01-04	AU	3.43	5.23	10.68	0.42	0.35
2020-01-05	AU	3.29	5.14	10.2	0.49	0.38
2020-01-06	AU	3.52	5.23	10.83	0.49	0.36
2020-01-07	AU	3.62	5.21	11.17	0.55	0.41
2020-01-08	AU	3.54	5.17	11.39	0.54	0.36
2020-01-09	AU	3.4	5.11	11.09	0.52	0.33

Figure 3: Snippet of google search trends dataset

```
2020-02-06      {alcoholism=1437, otitis_externa=2740, constipation=787, colitis=2124, impulsivity=2612}
2020-02-07      {alcoholism=1484, otitis_externa=2719, constipation=894, colitis=2123, impulsivity=2614}
2020-02-08      {alcoholism=1478, otitis_externa=2677, constipation=969, colitis=2108, impulsivity=2556}
2020-02-09      {alcoholism=1465, otitis_externa=2725, constipation=844, colitis=2105, impulsivity=2579}
2020-02-10      {alcoholism=1575, otitis_externa=2753, constipation=794, colitis=2132, impulsivity=2628}
2020-02-11      {alcoholism=1540, otitis_externa=2750, colitis=2121, impulsivity=2627, desquamation=730}
2020-02-12      {alcoholism=1509, otitis_externa=2736, colitis=2116, impulsivity=2612, desquamation=1062}
2020-02-13      {alcoholism=1454, otitis_externa=2725, colitis=2119, impulsivity=2606, desquamation=1245}
2020-02-14      {alcoholism=1549, otitis_externa=2688, colitis=2096, impulsivity=2547, desquamation=921}
2020-02-15      {alcoholism=1592, otitis_externa=2672, constipation=794, colitis=2097, impulsivity=2515}
2020-02-16      {alcoholism=1508, otitis_externa=2710, constipation=788, colitis=2112, impulsivity=2546}
2020-02-17      {alcoholism=1613, otitis_externa=2742, colitis=2117, impulsivity=2609, desquamation=759}
2020-02-18      {alcoholism=1560, otitis_externa=2755, colitis=2121, impulsivity=2619, desquamation=819}
2020-02-19      {alcoholism=1573, otitis_externa=2755, colitis=2119, impulsivity=2641, desquamation=905}
2020-02-20      {alcoholism=1520, otitis_externa=2740, colitis=2103, impulsivity=2614, desquamation=1020}
2020-02-21      {alcoholism=1618, otitis_externa=2734, colitis=2086, impulsivity=2588, desquamation=904}
2020-02-22      {alcoholism=1580, otitis_externa=2690, colitis=2070, abdominal_obesity=606, impulsivity=2518}
2020-02-23      {alcoholism=1563, otitis_externa=2709, colitis=2099, abdominal_obesity=703, impulsivity=2545}
```

Figure 4: Snippet of MapReduce output

2.2.1 Data cleaning

Again, the dataset is mostly cleaned, but some data is missing. We mark the missing data with the value -1 to avoid including it in the reducer code.

2.2.2 Data profiling

Top-k is the most commonly used metric for search trends. We're doing the same thing. Because each symptom's score is normalized by a different value, the score is meaningless across regions and dates. We can't really sum up each symptom and rank them from most searched to least searched. The ranking in each day, however, is what matters. Because we are only interested in data from the United States, the mapper filters out unrelated data and uses max Heap to select the top-k symptoms within each day at each location. Our reducer combines the top-k symptoms across states in each day, counts their frequency, and only selects the top-k most frequent symptoms.

Figure 4 depicts a portion of the mapreduce output. We output the top five symptoms people googled the most for each date. Along with the symptoms, we output the number of times these top keywords were searched. One interesting finding is that even during the COVID peak, the symptom of covid does not make it to the top 5, whereas alcoholism always does.

2.3 Government Response

This dataset is the summary of the government response to the COVID-19. It can be downloaded here. Only some part of table schema is shown as follows in Table 2.3:

Let's take a look at the number that corresponds to each column. The closing of schools and universities is referred to as school closing. 0 denotes no measurements. 1 denotes suggested closings. 2 denotes the need for closure. Three means there is no data. The rest of the numbers in the columns have similar meanings depending on the context.

This dataset is depicted in Fig. 5. This dataset has 303,970 records and is 17 MB in size.

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location(country_state_county)	US_NY_36029
school_closing	Record closings of schools [0-3]	2
workplace_closing	Record closings of workplaces [0-3]	1
stay_at_home	Record orders to stay at home [0-3]	3
debt_relief	Record if govt. is freezing financial obligations	2
testing_policy	COVID testing policy [0-3]	2

Table 3: Table Schema

date	location_key	school_closing	workplace_closing	cancel_public_events	restrictions_on_gatherings	public_transport_closing	stay_at_home_requirements
2020-01-01	AD	0	0	0	0	0	0
2020-01-02	AD	0	0	0	0	0	0
2020-01-03	AD	0	0	0	0	0	0
2020-01-04	AD	0	0	0	0	0	0
2020-01-05	AD	0	0	0	0	0	0
2020-01-06	AD	0	0	0	0	0	0
2020-01-07	AD	0	0	0	0	0	0
2020-01-08	AD	0	0	0	0	0	0
2020-01-09	AD	0	0	0	0	0	0

Figure 5: Snippet of government response dataset

2.3.1 Data cleaning

Again, the dataset is mostly cleaned, but some data is missing. We mark the missing data with the value -1 to avoid including it in the reducer code.

2.3.2 Data profiling

Because the data is discrete, taking the average makes little sense. Assume the school closing score is 2.345. What exactly does it mean? Is it means 2 or 3? The median, in my opinion, is the better numerical summation. So our mapper is basically doing the same thing as before, except this time it is filtering US data and outputting the attributes of each record. The reducer determines the medians for the attributes of interest. Figure 6 depicts a small portion of the mapreduce output. The information shown here is from when the covid first became popular. We begin to see income support and schools being forced to close.

```
2020-03-07      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-08      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-09      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-10      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-11      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-12      {school_closing=0.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-13      {school_closing=1.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-14      {school_closing=1.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-15      {school_closing=1.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-16      {school_closing=2.0, contact_policy=1.0, debt_support=0.0, income_support=0.0,
2020-03-17      {school_closing=2.0, contact_policy=1.0, debt_support=1.0, income_support=0.0,
2020-03-18      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=0.0,
2020-03-19      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=0.5,
2020-03-20      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=1.0,
2020-03-21      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=1.0,
2020-03-22      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=1.0,
2020-03-23      {school_closing=3.0, contact_policy=1.0, debt_support=1.0, income_support=1.0,
```

Figure 6: Snippet of MapReduce output