Covid-19 News Data

Team Torreys Peak
Dylan Bernstein, Jennifer Dickson, Katie Greenfield, Madison Moye, and Yongbo Shu

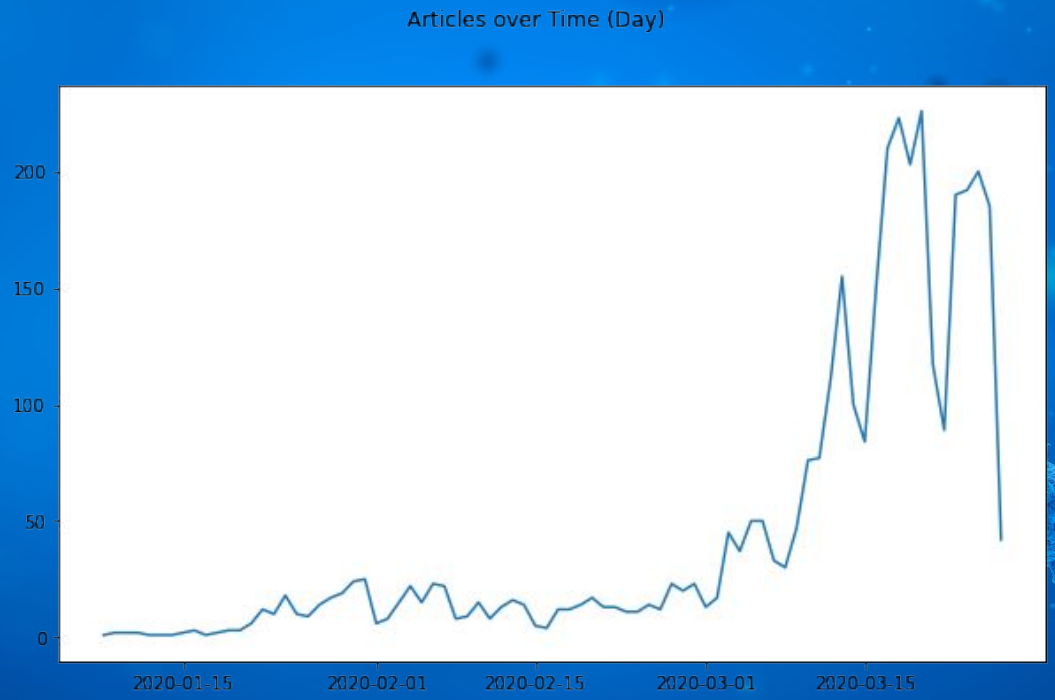
The Data

COVID-19 News Articles

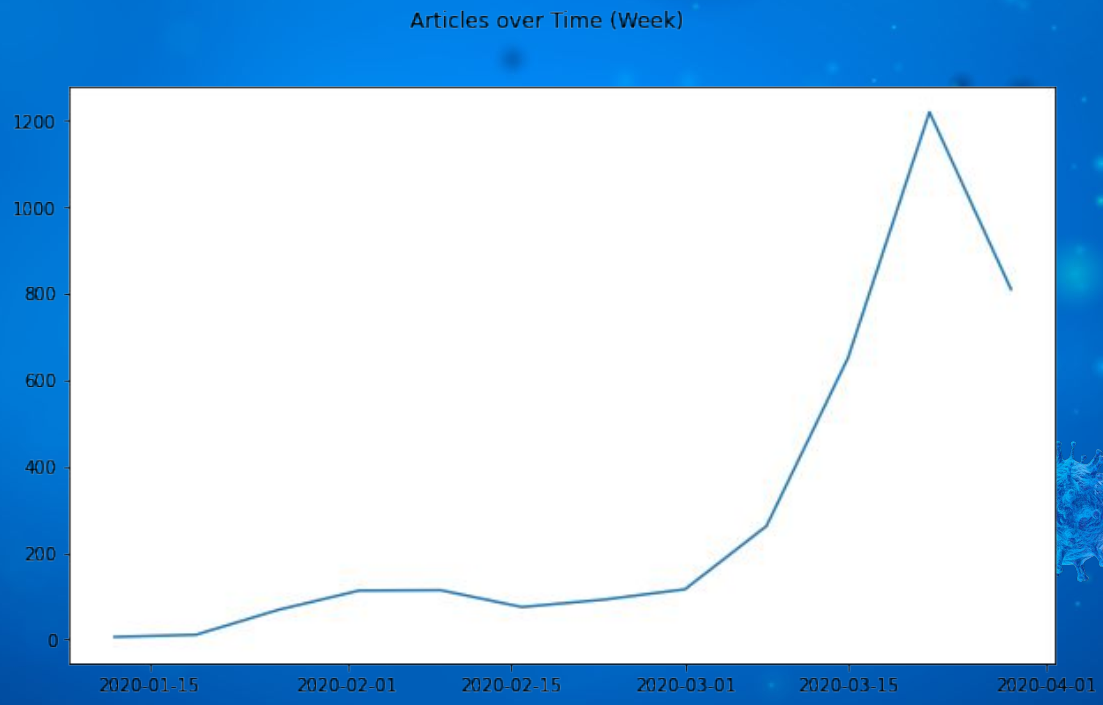
- 3500+ CBC News Articles
  - CBC = Canadian Broadcasting Company
- Contains authors, title, publish date, description, full article, and URL.
- Articles date from Dec, 2019 - Mar 27, 2020

https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26

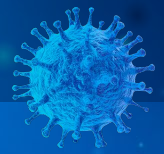
The Data


Articles over Time (Day)
200
150
100
50
0
2020-01-15
2020-02-01
2020-02-15
2020-03-01
2020-03-15

The Data


Articles over Time (Week)
1200
1000
800
600
400
200
0
2020-01-15
2020-02-01
2020-02-15
2020-03-01
2020-03-15
2020-04-01

Three Approaches



Word Cloud

Data Visualization tool that summarizes key topics or words in data in a visually intuitive manner.

Topic Modeling

Unsupervised machine learning model that deciphers topics in a corpus of data or documents.

Word Count

Using a pySpark RDD and pySpark Dataframe to count all the occurrences of words used in the articles.

1. Word Cloud

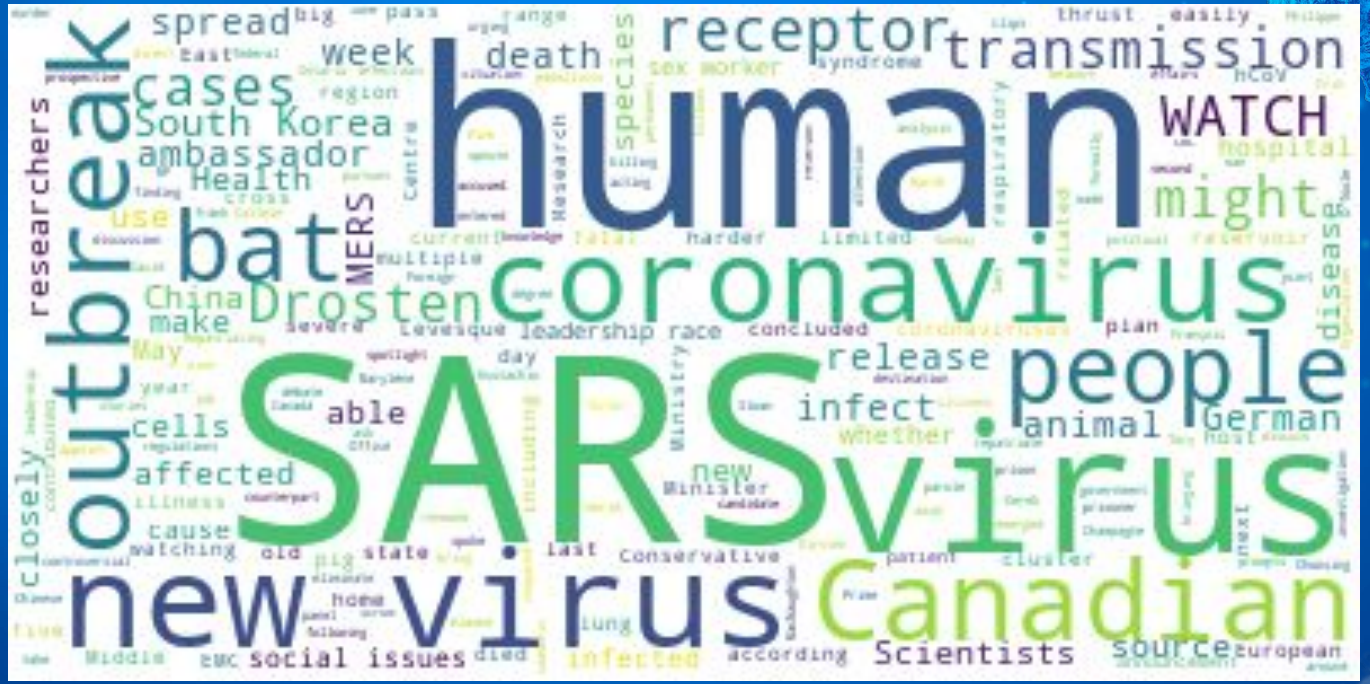
Word Cloud

Created word clouds for all text in the articles and also created word clouds for December, January, February and March to see if the articles had changed over time.

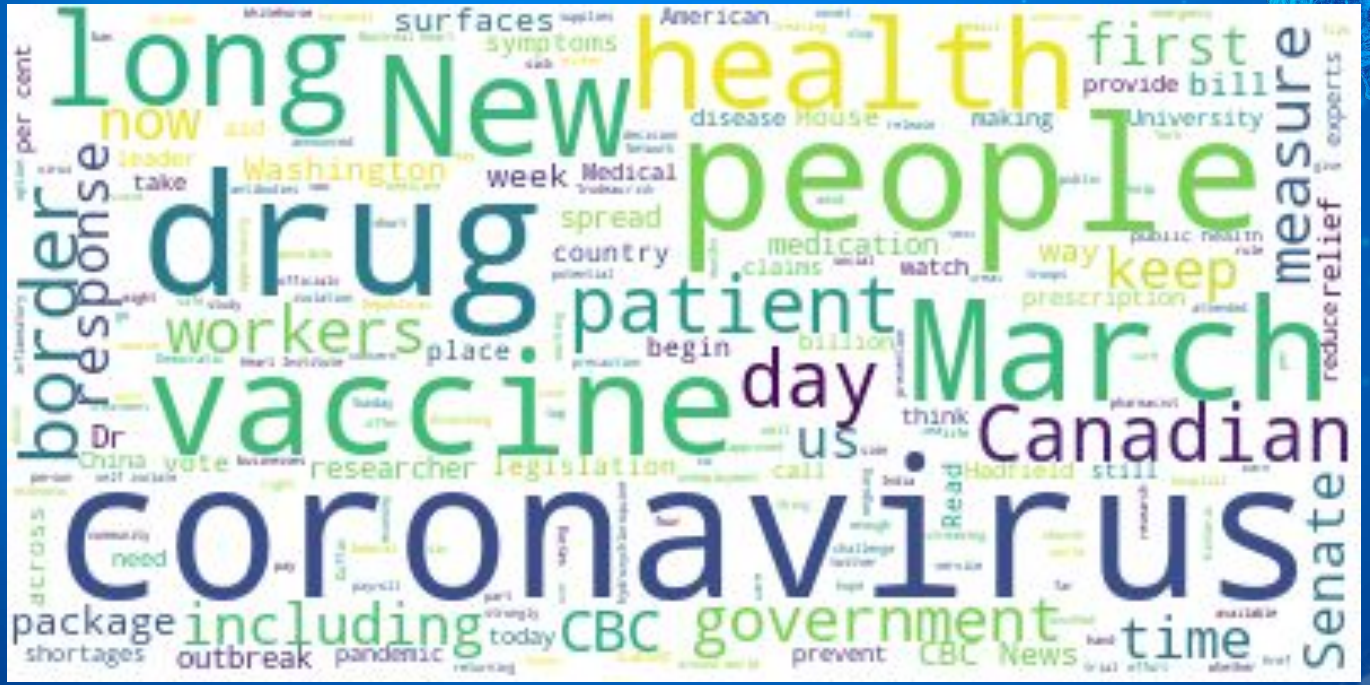
Found 15.8M words in the text of the articles.

Word Cloud - All Articles

Word Cloud - December 2019

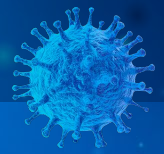


Word Cloud - March 2020



2. Topic Modeling

Topic Modeling – Data Preprocessing



Tokenizing

“Bob walked the dog” →

[Bob, walked, the, dog]

Position Tagging

Noun, verb, adj

Lemmatizing

“Walked” → “walk”

Remove Special Characters

!,”;-?- etc.

Remove Stopwords

‘http’ or ‘nt’

Two Topic Models

”Bigger” – less data preprocessing

“Smaller” – much more data preprocessing

“Bigger” Topic Model

Topic 1: #1: pei #2: islanders #3: morrison	Topic 2: #1: #2: the #3: be	Topic 3: #1: #2: the #3: to	Topic 4: #1: #2: the #3: be	Topic 5: #1: per #2: cent #3: price	Topic 6: #1: flight #2: travel #3: airline
Topic 7: #1: #2: the #3: be	Topic 8: #1: community #2: territory #3: nwt	Topic 9: #1: http #2: href #3: 5ewtf	Topic 10: #1: mask #2: hand #3: supply	Topic 11: #1: sanders #2: biden #3: read	Topic 12: #1: party #2: budget #3: election

“Smaller” Preprocessing

- Remove numbers and short words
- Remove common tokens (Threshold = 0.9)
- Remove Uncommon Tokens (Threshold=0.01)

- Beta of 0.5

“Smaller” Topic Model

Topic 1: #1: people #2: masks #3: health	Topic 2: #1: bc #2: people #3: health	Topic 3: #1: pei #2: nova #3: scotia	Topic 4: #1: ship #2: cruise #3: passenger	Topic 5: #1: government #2: trump #3: party	Topic 6: #1: health #2: case #3: hospital
Topic 7: #1: people #2: city #3: service	Topic 8: #1: canada #2: flight #3: canadians	Topic 9: #1: case #2: virus #3: china	Topic 10: #1: school #2: student #3: time	Topic 11: #1: health #2: case #3: province	Topic 12: #1: business #2: cent #3: government

3. Word Count

Word Count

Counted occurrences of all words that were used in all these news articles.

Two ways:

- pySpark rdd
- pySpark dataframe.

Word Count

complying with the Canadian Pharmacists Association call to limit the amount of medications given to patients to 30-d
ay\xa0supplies. The goal is to stop people from refilling prescriptions early and to ensure life-saving drugs don\'t
run short when supply chains are\xa0vulnerable. Mina Tadrous is a pharmacist and researcher in Toronto who monitors p

```
1 newdf = df.withColumn('text', regexp_replace('text', '\xa0', ''))
2 newdf = newdf.withColumn('text', regexp_replace('text', '\\\'', ''))
```

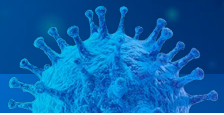
```
[('\U0001f9d0)', 1),
 ('\U0001f970🎉', 1),
 ('😳 &mdash;@RubyTuesday_72Tried', 1),
 ('😄 <a', 1),
 ('😂 &mdash;@hannahmaryr', 1),
 ('😂', 1),
 ('🔅 &mdash;@jdutchermusic', 2),
 ('💰', 1),
 ('💚', 1),
 ('💖', 3),
 ('💔 &mdash;@DocSchmadia', 2),
 ('💫', 3),
 ('👊🏾👊🏾', 1),
 ('👉 <a', 1),
 ('👇🏻 <br><br>We', 1),
 ('🏆', 1),
 ('🇨🇦 <br><br>Meet', 1),
 ('\ufeffCorriveau\ufeff', 1),
 (' - ', 2),
 ('➡️ <a', 1)]
```

Word Count

```
--- rdd running time (without sorting) is 0.030278921127319336 seconds ---
--- DataFrame running time (without sorting) is 0.183934688856811523 seconds ---
--- rdd running time (with sorting) is 8.592974662780762 seconds ---
--- DataFrame running time (with sorting) is 0.2873249053955078 seconds ---
```

```
--- rdd running time is 0.7412478923797607 seconds ---
--- DataFrame running time is 8.040996551513672 seconds ---
--- DataFrame running time is 1.1875431537628174 seconds ---
```

Word Count



authors	title	publish_date	text	url	wordCount
['Cbc News']	Coronavirus a 'wa...	2020-03-27 08:00:00	Canadian pharmaci...	https://www.cbc.c...	395
['Cbc News']	Yukon gov't names...	2020-03-27 01:45:00	The Yukon governm...	https://www.cbc.c...	303
['The Associated ...	U.S. Senate passe...	2020-03-26 05:13:00	The Senate late W...	https://www.cbc.c...	974
['Cbc News']	Coronavirus: The ...	2020-03-27 00:36:00	Scientists around...	https://www.cbc.c...	721
['Cbc News']	The latest on the...	2020-03-26 20:57:00	Trudeau says r...	https://www.cbc.c...	1388
['Mark Gollom Is ...	'Worse' pandemic ...	2020-03-27 08:00:00	The continued exi...	https://www.cbc.c...	1017
['Cbc News']	What you need to ...	2020-03-27 08:00:00	Recent developmen...	https://www.cbc.c...	1821
['The Associated ...	Michigan hospital...	2020-03-26 11:02:00	Michigan hospital...	https://www.cbc.c...	131
['Thomson Reuters']	U.S. coronavirus ...	2020-03-26 14:55:00	The number of con...	https://www.cbc.c...	1407
['Leah Hendry Is ...	'Avoid the emerge...	2020-03-27 08:00:00	The Jewish Genera...	https://www.cbc.c...	652
['Reporter', 'Web...	COVID-19 in Sask:...	2020-03-26 14:18:00	Three Saskatchew...	https://www.cbc.c...	1490
['Jorge Barrera I...	Manitoba chiefs o...	2020-03-27 08:01:00	A Manitoba chiefs...	https://www.cbc.c...	789
['Colleen M. Floo...	How invoking the ...	2020-03-27 08:00:00	This column is an...	https://www.cbc.c...	1051
['Producer', 'Cbc...	In Ontario, const...	2020-03-27 08:00:00	Construction is o...	https://www.cbc.c...	1557
['Dan Mcgarvey Is...	Alberta's film in...	2020-03-26 12:00:00	Albertas TV and m...	https://www.cbc.c...	687
['Cbc News New Yo...	'Like a war zone'...	2020-03-27 08:00:00	The first wave hi...	https://www.cbc.c...	908
[]	N.L. fisheries re...	2020-03-27 08:30:00	A Memorial Univer...	https://www.cbc.c...	484
[]	1st death, 3 new ...	2020-03-11 00:15:00	Manitoba is under...	https://www.cbc.c...	4783
['Investigative R...	Medical experts w...	2020-03-26 08:00:00	Medical experts a...	https://www.cbc.c...	1745
['Hadeel Ibrahim ...	It's 'too late' f...	2020-03-27 08:00:00	New Brunswick has...	https://www.cbc.c...	1239

only showing top 20 rows

Ingest data to S3 bucket

```
1  counts_rdd.write.parquet('s3://msbx5420-2020/Team-Torreys-Peak/counts.parquet')
```

```
1  nnewdf.write.parquet('s3://msbx5420-2020/Team-Torreys-Peak/news_countnews.parquet')
```

```
[hadoop@ip-172-16-1-40 ~]$ aws s3 ls s3://msbx5420-2020/Team-Torreys-Peak/
                           PRE counts.parquet/
                           PRE news_countnews.parquet/
2020-04-26 04:38:51   17478084 news.csv
```

Thank you!

Questions?