Anna Wang and Maxwell Du

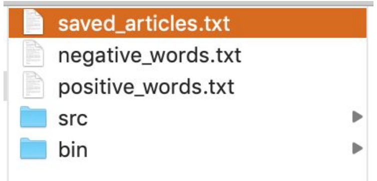
COVID-19 News Outlet Analyzer

Welcome to our final project! In our current times, in the throes of a global pandemic, news and media outlets are more salient than ever. While citizens around the world follow stay at home orders, news acts as a source of vital information which we can react to and predict our future livelihoods off of. Media can also show the current mindset and effect of corona on certain regions.

Thus, we thought that it would be interesting to apply the knowledge we learned from NETS 150 such as document search and information networks to parse through Google News's new section COVID-19 to compare articles of certain dates, regions, and publishers.

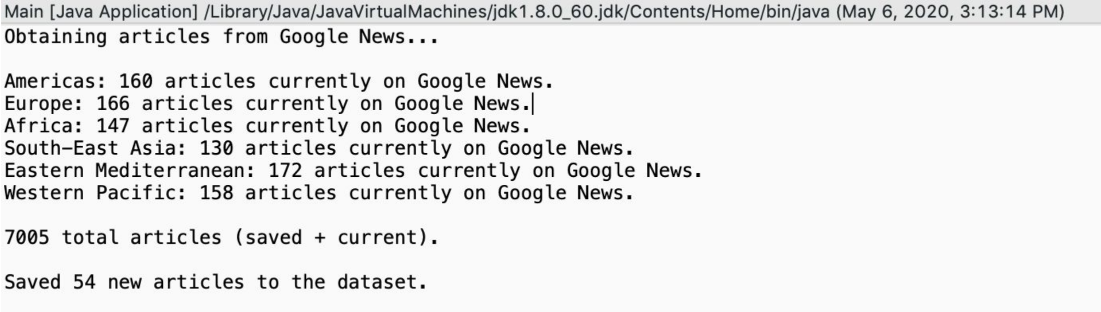
User Manual & Functionalities

Google News only shows recent articles from the past few days, so we have a saved_document txt file which contains saved COVID-19 articles from google news since we started our project (articles in this txt file range back to February 5th).



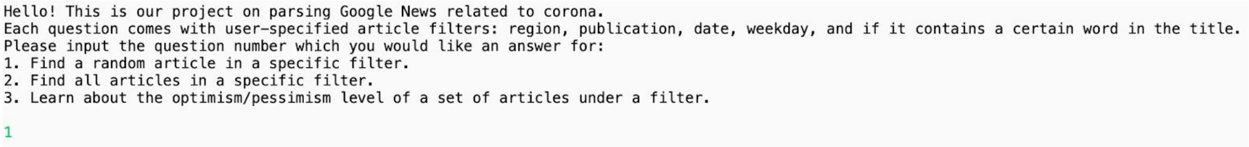
saved_articles.txt
negative_words.txt
positive_words.txt
src
bin

Everytime you run the program, we call ArticleSorter.saveAllArticles() which will go through Google News and tell you the number of current articles that are displayed from different regions. Then, we will display the number of saved articles and the number of new articles saved since you last ran the program. We prevented duplicates in the text file in our saveArticle() method.



```
Main [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_60.jdk/Contents/Home/bin/java (May 6, 2020, 3:13:14 PM)
Obtaining articles from Google News...

Americas: 160 articles currently on Google News.
Europe: 166 articles currently on Google News.
Africa: 147 articles currently on Google News.
South-East Asia: 130 articles currently on Google News.
Eastern Mediterranean: 172 articles currently on Google News.
Western Pacific: 158 articles currently on Google News.

7005 total articles (saved + current).

Saved 54 new articles to the dataset.
```

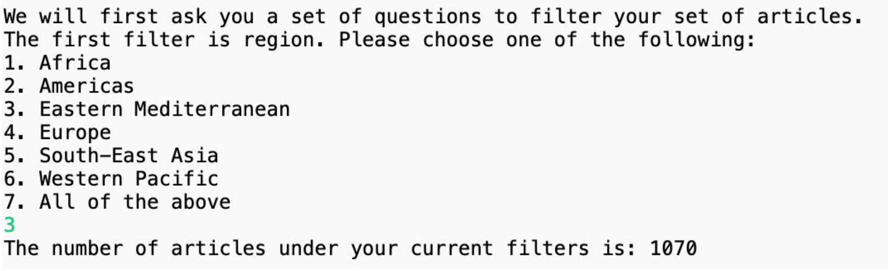
Our program has 3 main functionalities. At the offset, our UI will prompt the user to pick what that they want to learn about: 1. Finding a **random article** under a specific filter, 2. Finding **all articles** under a specific filter, and 3. Learn about the **optimism/pessimism** level of articles under a specific filter. Please enter 1, 2, or 3.



```
Hello! This is our project on parsing Google News related to corona.
Each question comes with user-specified article filters: region, publication, date, weekday, and if it contains a certain word in the title.
Please input the question number which you would like an answer for:
1. Find a random article in a specific filter.
2. Find all articles in a specific filter.
3. Learn about the optimism/pessimism level of a set of articles under a filter.

1
```

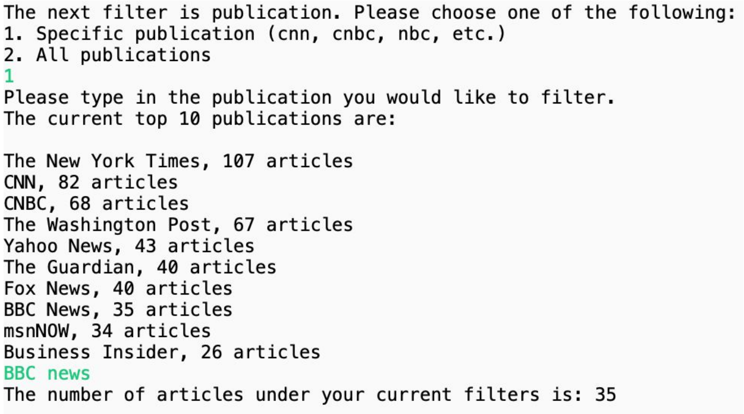
After picking one of these, our UI will prompt the user with a list of questions to filter down the articles they want based on 1. **region of publication**, 2. **publisher**, 3. **date of publication**, and 4. **keywords in title**. When filtering down the articles, our UI helps the user make smarter decisions by giving a list of regions to choose from, example publishers (lists out the top 10 frequent publishers in article set), and the earliest publication date. Each time you filter the set, we will tell you how many articles are currently in our set.

1. Filtering on region of publication
   Please enter a number 1 to 7 for which region you would like your articles from.



```
We will first ask you a set of questions to filter your set of articles.
The first filter is region. Please choose one of the following:
1. Africa
2. Americas
3. Eastern Mediterranean
4. Europe
5. South-East Asia
6. Western Pacific
7. All of the above
3
The number of articles under your current filters is: 1070
```

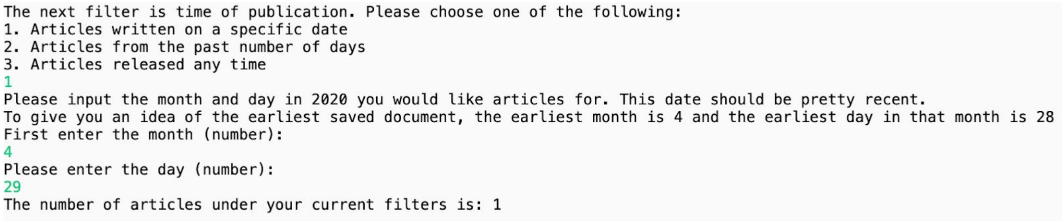
2. Filtering on publisher
   If you'd like to filter by publisher of the article, enter 1. We will then give you a list of the top 10 frequent publishers based off of your current set of articles to help you out. You can then enter which publisher you'd like as a filter.



```
The next filter is publication. Please choose one of the following:
1. Specific publication (cnn, cnbc, nbc, etc.)
2. All publications
1
Please type in the publication you would like to filter.
The current top 10 publications are:

The New York Times, 107 articles
CNN, 82 articles
CNBC, 68 articles
The Washington Post, 67 articles
Yahoo News, 43 articles
The Guardian, 40 articles
Fox News, 40 articles
BBC News, 35 articles
msnNOW, 34 articles
Business Insider, 26 articles
BBC news
The number of articles under your current filters is: 35
```

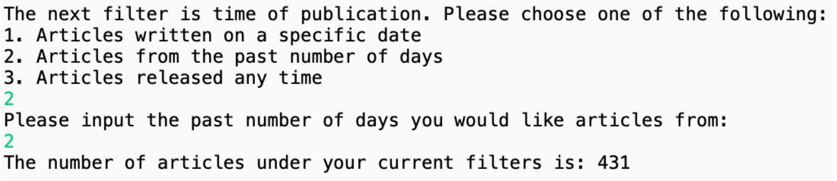
   Sometimes, the number of articles under a publisher may be more than the number of articles shown next to that publisher in the top 10 list. This is because if you type in "yahoo" for example, it will get you all yahoo related articles from yahoo UK to yahoo india. Thus, we'd recommend you to be as specific as possible.

3. Filtering on date of publication
   Next, you can filter the date of publication either on a specific date or on the past number of days. If you'd like to find articles on a specific date, we will tell you the earliest article's date of publication. Please enter a month number and day number *after* this given earliest date.



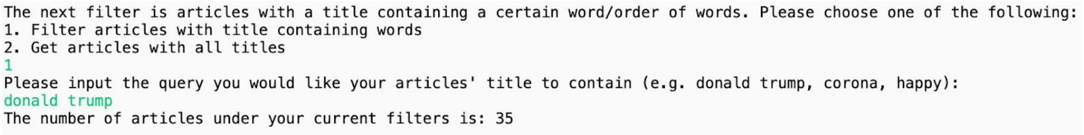
```
The next filter is time of publication. Please choose one of the following:
1. Articles written on a specific date
2. Articles from the past number of days
3. Articles released any time
1
Please input the month and day in 2020 you would like articles for. This date should be pretty recent.
To give you an idea of the earliest saved document, the earliest month is 4 and the earliest day in that month is 28
First enter the month (number):
4
Please enter the day (number):
29
The number of articles under your current filters is: 1
```

If you'd like articles from the past __ days, enter the __ days you'd like articles from.



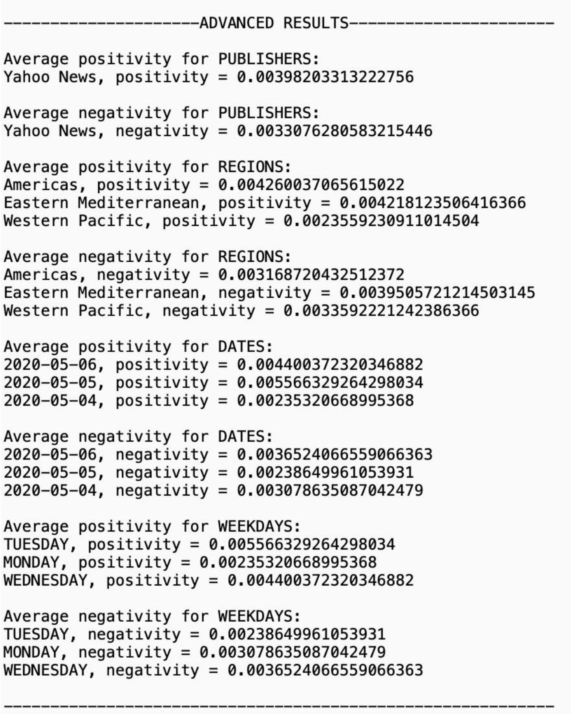
```
The next filter is time of publication. Please choose one of the following:
1. Articles written on a specific date
2. Articles from the past number of days
3. Articles released any time
2
Please input the past number of days you would like articles from:
2
The number of articles under your current filters is: 431
```

4. Filtering on keywords in title
   Finally, you can filter based on keywords you would like in your title. Simply type in 1 and then the phrase/word you would like your articles' titles to contain.

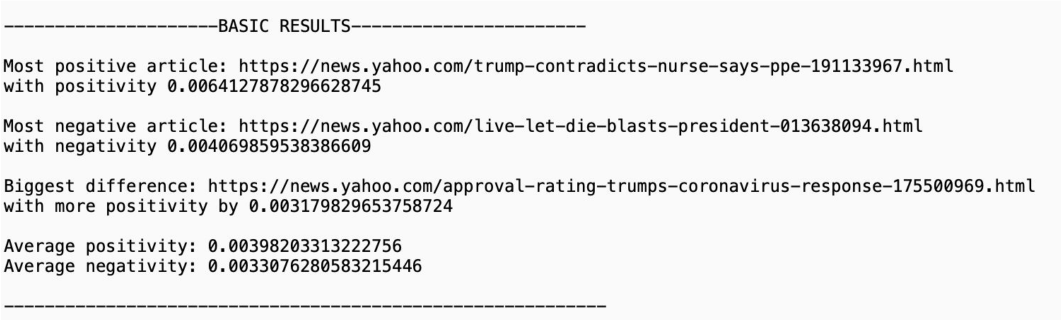


```
The next filter is articles with a title containing a certain word/order of words. Please choose one of the following:
1. Filter articles with title containing words
2. Get articles with all titles
1
Please input the query you would like your articles' title to contain (e.g. donald trump, corona, happy):
donald trump
The number of articles under your current filters is: 35
```

At the end, it will show you results based on your original query and the filters you put in. If you asked for positivity/negativity of a set of articles, our optimism/pessimism calculator will release a set of basic results and advanced results. For example:

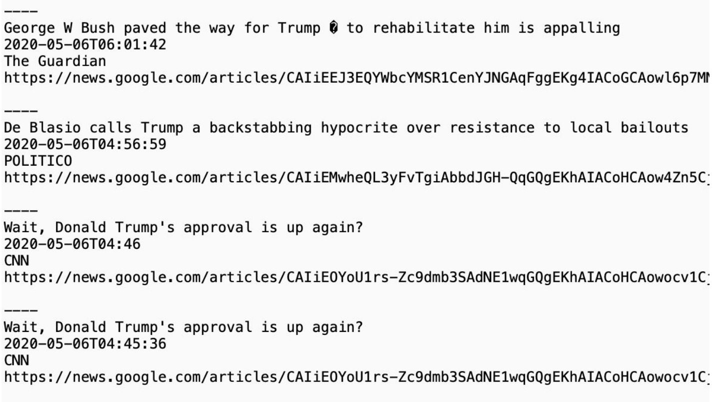


```
---------------------ADVANCED RESULTS----------------------

Average positivity for PUBLISHERS:
Yahoo News, positivity = 0.00398203313222756

Average negativity for PUBLISHERS:
Yahoo News, negativity = 0.0033076280583215446

Average positivity for REGIONS:
Americas, positivity = 0.004260037065615022
Eastern Mediterranean, positivity = 0.004218123506416366
Western Pacific, positivity = 0.0023559230911014504

Average negativity for REGIONS:
Americas, negativity = 0.003168720432512372
Eastern Mediterranean, negativity = 0.0039505721214503145
Western Pacific, negativity = 0.0033592221242386366

Average positivity for DATES:
2020-05-06, positivity = 0.004400372320346882
2020-05-05, positivity = 0.005566329264298034
2020-05-04, positivity = 0.00235320668995368

Average negativity for DATES:
2020-05-06, negativity = 0.0036524066559066363
2020-05-05, negativity = 0.00238649961053931
2020-05-04, negativity = 0.003078635087042479

Average positivity for WEEKDAYS:
TUESDAY, positivity = 0.005566329264298034
MONDAY, positivity = 0.00235320668995368
WEDNESDAY, positivity = 0.004400372320346882

Average negativity for WEEKDAYS:
TUESDAY, negativity = 0.00238649961053931
MONDAY, negativity = 0.003078635087042479
WEDNESDAY, negativity = 0.0036524066559066363

-----------------------------------------------------------
```



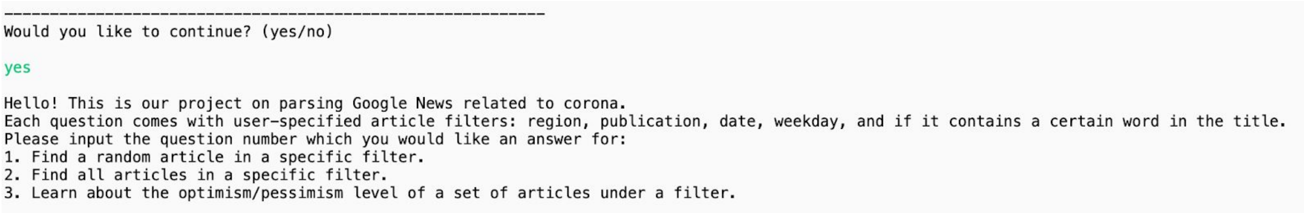
```
---------------------BASIC RESULTS-----------------------

Most positive article: https://news.yahoo.com/trump-contradicts-nurse-says-ppe-191133967.html
with positivity 0.0064127878296628745

Most negative article: https://news.yahoo.com/live-let-die-blasts-president-013638094.html
with negativity 0.004069859538386609

Biggest difference: https://news.yahoo.com/approval-rating-trumps-coronavirus-response-175500969.html
with more positivity by 0.003179829653758724

Average positivity: 0.00398203313222756
Average negativity: 0.0033076280583215446

-----------------------------------------------------------
```

If you selected all articles or random article under a specific filter, the UI will print out the article(s) and their url & information. For example:



```
----
George W Bush paved the way for Trump � to rehabilitate him is appalling
2020-05-06T06:01:42
The Guardian
https://news.google.com/articles/CAIiEEJ3EQYWbcYMSR1CenYJNGAqFggEKg4IACoGCAowl6p7M

----
De Blasio calls Trump a backstabbing hypocrite over resistance to local bailouts
2020-05-06T04:56:59
POLITICO
https://news.google.com/articles/CAIiEMwheQL3yFvTgiAbbdJGH-QqGQgEKhAIACoHCAow4Zn5C

----
Wait, Donald Trump's approval is up again?
2020-05-06T04:46
CNN
https://news.google.com/articles/CAIiEOYoU1rs-Zc9dmb3SAdNE1wqGQgEKhAIACoHCAowocv1C

----
Wait, Donald Trump's approval is up again?
2020-05-06T04:45:36
CNN
https://news.google.com/articles/CAIiEOYoU1rs-Zc9dmb3SAdNE1wqGQgEKhAIACoHCAowocv1C
```

Finally, our UI will prompt you to choose to continue or not. Please enter yes or no.



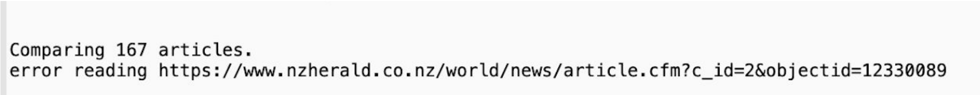
```
-----------------------------------------------------------
Would you like to continue? (yes/no)

yes

Hello! This is our project on parsing Google News related to corona.
Each question comes with user-specified article filters: region, publication, date, weekday, and if it contains a certain word in the title.
Please input the question number which you would like an answer for:
1. Find a random article in a specific filter.
2. Find all articles in a specific filter.
3. Learn about the optimism/pessimism level of a set of articles under a filter.
```

Congratulations! You have learned a bit more about COVID-19 news articles under a specific filter.

Considerations

If you don't type in an existing option (e.g. asked to type in 1-7 for regions but you type in 8), your data set will be empty and we will terminate the filtering and ask if you'd like to start over.

Very rarely when comparing articles, there will be a reading error that occurs like the following:



```
Comparing 167 articles.
error reading https://www.nzherald.co.nz/world/news/article.cfm?c_id=2&objectid=12330089
```

This occurs when google news has just updated a url and there is an error when jsoup tries to get the html file. This happens very rarely, so please ignore it and rerun if it occurs.

Acknowledgements

We would like to thank our TA Kunal for his thorough guidance and suggestions through this project. We would also like to thank Swap for everything he taught us in this class to allow us to make this project. Finally, we would like to thank The University of Pennsylvania for all of its resources.