Assignment 3

PROBLEM1-TASK2

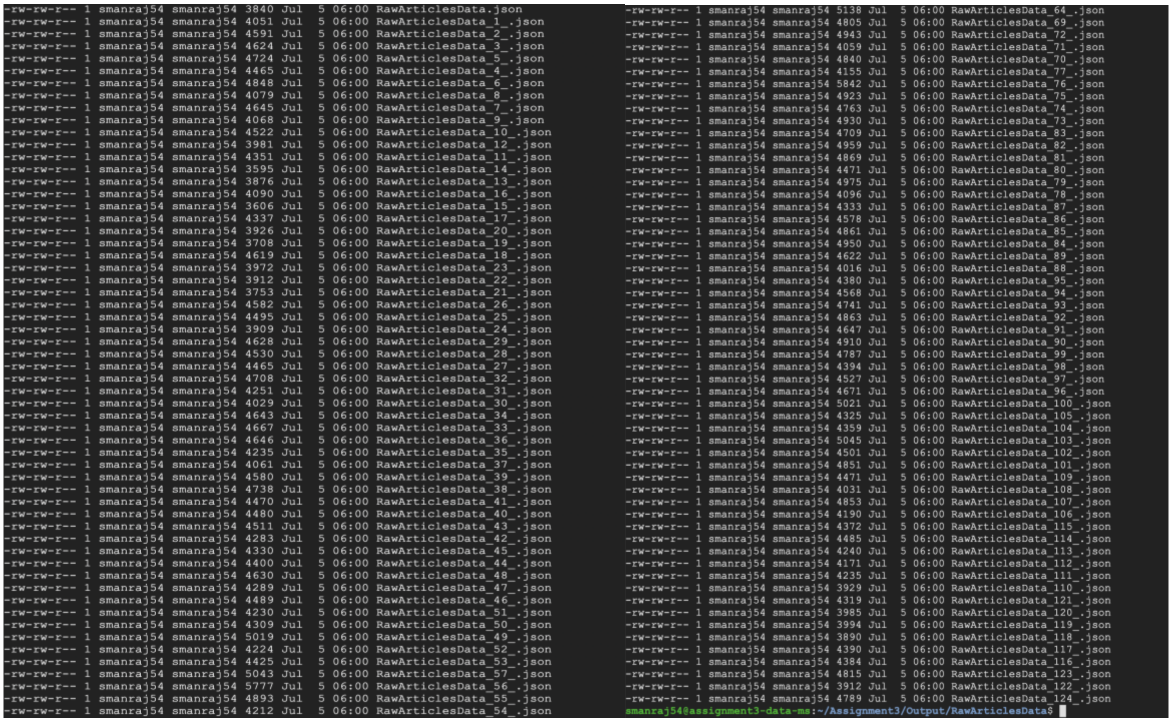
Manraj Singh

B00877934

Problem-1

Task-2

- Pseudocode for data extraction:
  - Recursive delete all the pre-stored files in the Output/ directory.
  - Generate the API link for all the keywords using link [4] : *https://newsapi.org/v2/everything?q=<KeywordName>&sortBy=popularity&apiKey=6e7ddac0bd2e44ec82aa90073f939e16&language=en&pageSize=90*
  - Replace the <keywordName> in ablove mentioned link with the keywords mentioned in the problem task 2.
  - Hit the new generated urls using HttpURLConnection and InputStreamReader and BufferedReader.
  - Fetch all the articles from each API hit in the loop.
  - To get the articles only from the raw data calculated startPoint and endpoint of the new SubString
    - StartPoint = rawData.indexOf("/"articles/":[");
    - EndPoint = rawData.lastIndexOf("]");
    - Create substring from startPoint and endpoint
  - Create folder Structure to store raw files. Path : "Output/RawArticlesData/"
  - Break the data into group of 5 articles.
  - Write each group of 5 articles in a file inside the created folder structure.
  - Create new file with an incremented name *"RawArticlesData_<number++>_"* for further articles entry.



```
-rw-rw-r-- 1 smanraj54 smanraj54 3840 Jul  5 06:00 RawArticlesData.json
-rw-rw-r-- 1 smanraj54 smanraj54 4051 Jul  5 06:00 RawArticlesData_1_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4591 Jul  5 06:00 RawArticlesData_2_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4624 Jul  5 06:00 RawArticlesData_3_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4724 Jul  5 06:00 RawArticlesData_5_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4465 Jul  5 06:00 RawArticlesData_4_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4848 Jul  5 06:00 RawArticlesData_6_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4079 Jul  5 06:00 RawArticlesData_8_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4645 Jul  5 06:00 RawArticlesData_7_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4068 Jul  5 06:00 RawArticlesData_9_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4522 Jul  5 06:00 RawArticlesData_10_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3981 Jul  5 06:00 RawArticlesData_12_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4351 Jul  5 06:00 RawArticlesData_11_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3595 Jul  5 06:00 RawArticlesData_14_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3876 Jul  5 06:00 RawArticlesData_13_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4090 Jul  5 06:00 RawArticlesData_16_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3606 Jul  5 06:00 RawArticlesData_15_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4337 Jul  5 06:00 RawArticlesData_17_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3926 Jul  5 06:00 RawArticlesData_20_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3708 Jul  5 06:00 RawArticlesData_19_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4619 Jul  5 06:00 RawArticlesData_18_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3972 Jul  5 06:00 RawArticlesData_23_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3912 Jul  5 06:00 RawArticlesData_22_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3753 Jul  5 06:00 RawArticlesData_21_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4582 Jul  5 06:00 RawArticlesData_26_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4495 Jul  5 06:00 RawArticlesData_25_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3909 Jul  5 06:00 RawArticlesData_24_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4628 Jul  5 06:00 RawArticlesData_29_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4530 Jul  5 06:00 RawArticlesData_28_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4465 Jul  5 06:00 RawArticlesData_27_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4708 Jul  5 06:00 RawArticlesData_32_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4251 Jul  5 06:00 RawArticlesData_31_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4029 Jul  5 06:00 RawArticlesData_30_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4643 Jul  5 06:00 RawArticlesData_34_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4667 Jul  5 06:00 RawArticlesData_33_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4646 Jul  5 06:00 RawArticlesData_36_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4235 Jul  5 06:00 RawArticlesData_35_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4061 Jul  5 06:00 RawArticlesData_37_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4580 Jul  5 06:00 RawArticlesData_39_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4738 Jul  5 06:00 RawArticlesData_38_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4470 Jul  5 06:00 RawArticlesData_41_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4480 Jul  5 06:00 RawArticlesData_40_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4511 Jul  5 06:00 RawArticlesData_43_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4283 Jul  5 06:00 RawArticlesData_42_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4330 Jul  5 06:00 RawArticlesData_45_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4400 Jul  5 06:00 RawArticlesData_44_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4630 Jul  5 06:00 RawArticlesData_48_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4289 Jul  5 06:00 RawArticlesData_47_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4489 Jul  5 06:00 RawArticlesData_46_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4230 Jul  5 06:00 RawArticlesData_51_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4309 Jul  5 06:00 RawArticlesData_50_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5019 Jul  5 06:00 RawArticlesData_49_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4224 Jul  5 06:00 RawArticlesData_52_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4425 Jul  5 06:00 RawArticlesData_53_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5043 Jul  5 06:00 RawArticlesData_57_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5777 Jul  5 06:00 RawArticlesData_56_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4893 Jul  5 06:00 RawArticlesData_55_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4212 Jul  5 06:00 RawArticlesData_54_.json
```

```
-rw-rw-r-- 1 smanraj54 smanraj54 5138 Jul  5 06:00 RawArticlesData_64_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4805 Jul  5 06:00 RawArticlesData_69_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4943 Jul  5 06:00 RawArticlesData_72_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4059 Jul  5 06:00 RawArticlesData_71_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4840 Jul  5 06:00 RawArticlesData_70_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4155 Jul  5 06:00 RawArticlesData_77_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5842 Jul  5 06:00 RawArticlesData_76_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4923 Jul  5 06:00 RawArticlesData_75_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4763 Jul  5 06:00 RawArticlesData_74_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4930 Jul  5 06:00 RawArticlesData_73_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4709 Jul  5 06:00 RawArticlesData_83_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4959 Jul  5 06:00 RawArticlesData_82_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4869 Jul  5 06:00 RawArticlesData_81_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4471 Jul  5 06:00 RawArticlesData_80_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4975 Jul  5 06:00 RawArticlesData_79_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4096 Jul  5 06:00 RawArticlesData_78_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4333 Jul  5 06:00 RawArticlesData_87_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4578 Jul  5 06:00 RawArticlesData_86_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4861 Jul  5 06:00 RawArticlesData_85_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4950 Jul  5 06:00 RawArticlesData_84_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4622 Jul  5 06:00 RawArticlesData_89_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4016 Jul  5 06:00 RawArticlesData_88_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4380 Jul  5 06:00 RawArticlesData_95_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4568 Jul  5 06:00 RawArticlesData_94_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4741 Jul  5 06:00 RawArticlesData_93_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4863 Jul  5 06:00 RawArticlesData_92_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4647 Jul  5 06:00 RawArticlesData_91_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4910 Jul  5 06:00 RawArticlesData_90_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4787 Jul  5 06:00 RawArticlesData_99_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4394 Jul  5 06:00 RawArticlesData_98_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4527 Jul  5 06:00 RawArticlesData_97_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4671 Jul  5 06:00 RawArticlesData_96_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5021 Jul  5 06:00 RawArticlesData_100_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4325 Jul  5 06:00 RawArticlesData_105_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4359 Jul  5 06:00 RawArticlesData_104_.json
-rw-rw-r-- 1 smanraj54 smanraj54 5045 Jul  5 06:00 RawArticlesData_103_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4501 Jul  5 06:00 RawArticlesData_102_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4851 Jul  5 06:00 RawArticlesData_101_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4471 Jul  5 06:00 RawArticlesData_109_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4031 Jul  5 06:00 RawArticlesData_108_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4853 Jul  5 06:00 RawArticlesData_107_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4190 Jul  5 06:00 RawArticlesData_106_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4372 Jul  5 06:00 RawArticlesData_115_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4485 Jul  5 06:00 RawArticlesData_114_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4240 Jul  5 06:00 RawArticlesData_113_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4171 Jul  5 06:00 RawArticlesData_112_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4235 Jul  5 06:00 RawArticlesData_111_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3929 Jul  5 06:00 RawArticlesData_110_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4319 Jul  5 06:00 RawArticlesData_121_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3985 Jul  5 06:00 RawArticlesData_120_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3994 Jul  5 06:00 RawArticlesData_119_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3890 Jul  5 06:00 RawArticlesData_118_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4390 Jul  5 06:00 RawArticlesData_117_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4384 Jul  5 06:00 RawArticlesData_116_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4815 Jul  5 06:00 RawArticlesData_123_.json
-rw-rw-r-- 1 smanraj54 smanraj54 3912 Jul  5 06:00 RawArticlesData_122_.json
-rw-rw-r-- 1 smanraj54 smanraj54 4789 Jul  5 06:00 RawArticlesData_124_.json
smanraj54@assignment3-data-ms:~/Assignment3/Output/RawArticlesData$ 
```

Figure 1: RawArticlesData.json files [1][3]

- As shown in figure:1 [1][3] 125 raw files generated with 5 articles each.
- To get the articles segregated from the raw data I used the index of "{" and "}" brackets.

- FlowChart for MongoDB connection and uploading cleaned data:[2]

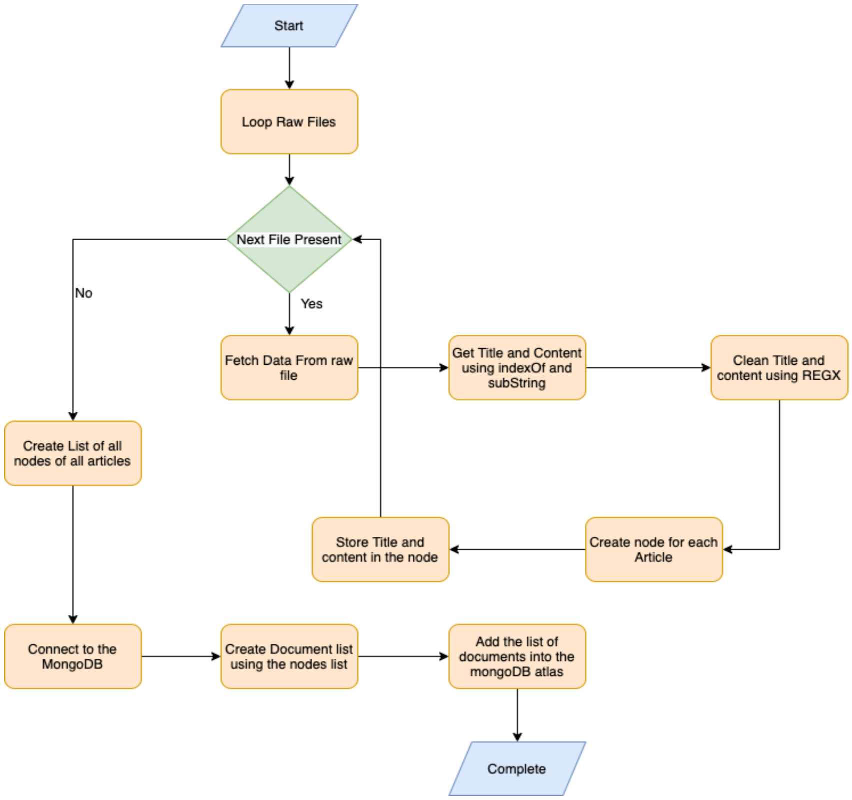
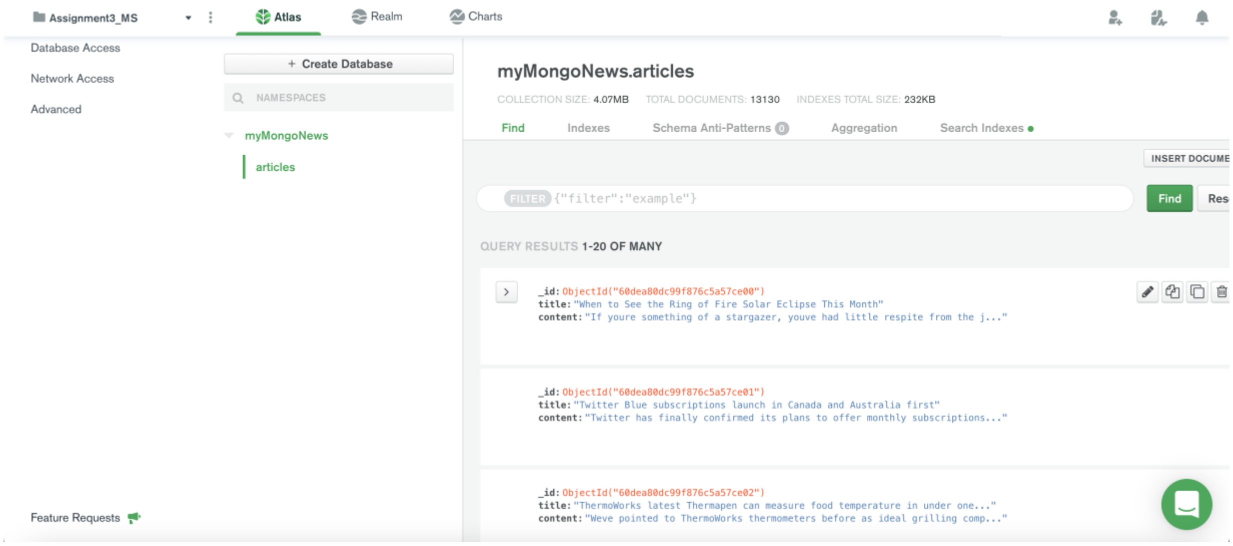

Start
Loop Raw Files
Next File Present
No
Yes
Fetch Data From raw file
Get Title and Content using indexOf and subString
Clean Title and content using REGX
Create List of all nodes of all articles
Store Title and content in the node
Create node for each Article
Connect to the MongoDB
Create Document list using the nodes list
Add the list of documents into the mongoDB atlas
Complete

Figure 2: Mongo Connection flowchart and data cleaning [5]

- Data is fetched from the raw files containing 5 articles each.
- Title and content fields are fetched for further processing using indexOf and substring.
- To clean the data REGX is used:
  - Regex variable = "[^0-9a-zA-Z:,\\s?!()\\\/\\.]+" is used to remove all the special characters and emoticons.
  - Regex variable = "\\<.*?\\>" is used to remove all the URLs and Http codes.
  - Regex variable = "(\r\n|\n)" is used to remove the special characters like "\r" or "\n"
- Nodes are created one per each article to contain the cleaned title and content data for each article.
- These nodes are stored in a list of nodes for further use.
- A json file is maintained containing all the the nodes data in json format (./Output/MongoArticlesProcessed/ MongoArticlesProcessed.json).
- MongoDB client connection is made.

- MongoDB Database named : “MyMongoNews” is made.
- MongoCollection named : “articles” is made
- Mongo DB connection is created using MongoDB atlas credentials.
- Documents list is created which contains Documents one per each node in the node list.
- Documents contain title and content in the json form.
- This Documents list is added to the MongoDB



myMongoNews.articles

COLLECTION SIZE: 4.07MB   TOTAL DOCUMENTS: 13130   INDEXES TOTAL SIZE: 232KB

{"filter":"example"}

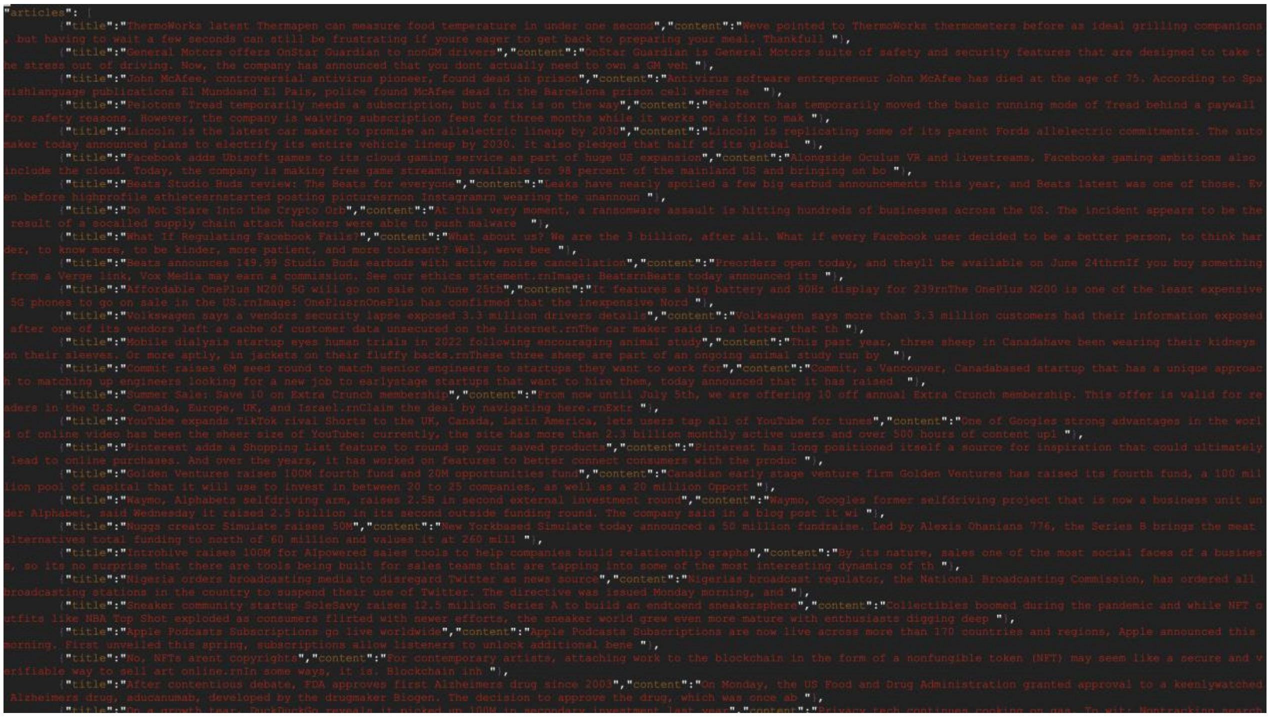
QUERY RESULTS 1-20 OF MANY

```
_id: ObjectId("60dea80dc99f876c5a57ce00")
title: "When to See the Ring of Fire Solar Eclipse This Month"
content: "If youre something of a stargazer, youve had little respite from the j..."

_id: ObjectId("60dea80dc99f876c5a57ce01")
title: "Twitter Blue subscriptions launch in Canada and Australia first"
content: "Twitter has finally confirmed its plans to offer monthly subscriptions..."

_id: ObjectId("60dea80dc99f876c5a57ce02")
title: "ThermoWorks latest Thermapen can measure food temperature in under one..."
content: "Weve pointed to ThermoWorks thermometers before as ideal grilling comp..."
```

Figure 3: MongoDB Atlas View after adding all the articles [2]



```
"articles": [
        {"title":"ThermoWorks latest Thermapen can measure food temperature in under one second","content":"Weve pointed to ThermoWorks thermometers before as ideal grilling companions
, but having to wait a few seconds can still be frustrating if youre eager to get back to preparing your meal. Thankfull "},
        {"title":"General Motors offers OnStar Guardian to nonGM drivers","content":"OnStar Guardian is General Motors suite of safety and security features that are designed to take t
he stress out of driving. Now, the company has announced that you dont actually need to own a GM veh "},
        {"title":"John McAfee, controversial antivirus pioneer, found dead in prison","content":"Antivirus software entrepreneur John McAfee has died at the age of 75. According to Spa
nishlanguage publications El Mundoand El Pais, police found McAfee dead in the Barcelona prison cell where he  "},
        {"title":"Pelotons Tread temporarily needs a subscription, but a fix is on the way","content":"Pelotonrn has temporarily moved the basic running mode of Tread behind a paywall
for safety reasons. However, the company is waiving subscription fees for three months while it works on a fix to mak "},
        {"title":"Lincoln is the latest car maker to promise an allelectric lineup by 2030","content":"Lincoln is replicating some of its parent Fords allelectric commitments. The auto
maker today announced plans to electrify its entire vehicle lineup by 2030. It also pledged that half of its global  "},
        {"title":"Facebook adds Ubisoft games to its cloud gaming service as part of huge US expansion","content":"Alongside Oculus VR and livestreams, Facebooks gaming ambitions also
include the cloud. Today, the company is making free game streaming available to 98 percent of the mainland US and bringing on bo "},
        {"title":"Beats Studio Buds review: The Beats for everyone","content":"Leaks have nearly spoiled a few big earbud announcements this year, and Beats latest was one of those. Ev
en before highprofile athletesrnstarted posting picturesrnon Instagramrn wearing the unannoun "},
        {"title":"Do Not Stare Into the Crypto Orb","content":"At this very moment, a ransomware assault is hitting hundreds of businesses across the US. The incident appears to be the
 result of a socalled supply chain attack hackers were able to push malware  "},
        {"title":"What If Regulating Facebook Fails?","content":"What about us? We are the 3 billion, after all. What if every Facebook user decided to be a better person, to think har
der, to know more, to be kinder, more patient, and more tolerant? Well, weve bee "},
        {"title":"Beats announces 149.99 Studio Buds earbuds with active noise cancellation","content":"Preorders open today, and theyll be available on June 24thrnIf you buy something
 from a Verge link, Vox Media may earn a commission. See our ethics statement.rnImage: BeatsrnBeats today announced its "},
        {"title":"Affordable OnePlus N200 5G will go on sale on June 25th","content":"It features a big battery and 90Hz display for 239rnThe OnePlus N200 is one of the least expensive
 5G phones to go on sale in the US.rnImage: OnePlusrnOnePlus has confirmed that the inexpensive Nord "},
        {"title":"Volkswagen says a vendors security lapse exposed 3.3 million drivers details","content":"Volkswagen says more than 3.3 million customers had their information exposed
 after one of its vendors left a cache of customer data unsecured on the internet.rnThe car maker said in a letter that th "},
        {"title":"Mobile dialysis startup eyes human trials in 2022 following encouraging animal study","content":"This past year, three sheep in Canadahave been wearing their kidneys
on their sleeves. Or more aptly, in jackets on their fluffy backs.rnThese three sheep are part of an ongoing animal study run by  "},
        {"title":"Commit raises 6M seed round to match senior engineers to startups they want to work for","content":"Commit, a Vancouver, Canadabased startup that has a unique approac
h to matching up engineers looking for a new job to earlystage startups that want to hire them, today announced that it has raised  "},
        {"title":"Summer Sale: Save 10 on Extra Crunch membership","content":"From now until July 5th, we are offering 10 off annual Extra Crunch membership. This offer is valid for re
aders in the U.S., Canada, Europe, UK, and Israel.rnClaim the deal by navigating here.rnExtr "},
        {"title":"YouTube expands TikTok rival Shorts to the UK, Canada, Latin America, lets users tap all of YouTube for tunes","content":"One of Googles strong advantages in the worl
d of online video has been the sheer size of YouTube: currently, the site has more than 2.3 billion monthly active users and over 500 hours of content upl "},
        {"title":"Pinterest adds a Shopping List feature to round up your saved products","content":"Pinterest has long positioned itself a source for inspiration that could ultimately
 lead to online purchases. And over the years, it has worked on features to better connect consumers with the produc "},
        {"title":"Golden Ventures raises 100M fourth fund and 20M opportunities fund","content":"Canadian early stage venture firm Golden Ventures has raised its fourth fund, a 100 mil
lion pool of capital that it will use to invest in between 20 to 25 companies, as well as a 20 million Opport "},
        {"title":"Waymo, Alphabets selfdriving arm, raises 2.5B in second external investment round","content":"Waymo, Googles former selfdriving project that is now a business unit un
der Alphabet, said Wednesday it raised 2.5 billion in its second outside funding round. The company said in a blog post it wi "},
        {"title":"Nuggs creator Simulate raises 50M","content":"New Yorkbased Simulate today announced a 50 million fundraise. Led by Alexis Ohanians 776, the Series B brings the meat
alternatives total funding to north of 60 million and values it at 260 mill "},
        {"title":"Introhive raises 100M for AIpowered sales tools to help companies build relationship graphs","content":"By its nature, sales one of the most social faces of a busines
s, so its no surprise that there are tools being built for sales teams that are tapping into some of the most interesting dynamics of th "},
        {"title":"Nigeria orders broadcasting media to disregard Twitter as news source","content":"Nigerias broadcast regulator, the National Broadcasting Commission, has ordered all
broadcasting stations in the country to suspend their use of Twitter. The directive was issued Monday morning, and "},
        {"title":"Sneaker community startup SoleSavy raises 12.5 million Series A to build an endtoend sneakersphere","content":"Collectibles boomed during the pandemic and while NFT o
utfits like NBA Top Shot exploded as consumers flirted with newer efforts, the sneaker world grew even more mature with enthusiasts digging deep "},
        {"title":"Apple Podcasts Subscriptions go live worldwide","content":"Apple Podcasts Subscriptions are now live across more than 170 countries and regions, Apple announced this
morning. First unveiled this spring, subscriptions allow listeners to unlock additional bene "},
        {"title":"No, NFTs arent copyrights","content":"For contemporary artists, attaching work to the blockchain in the form of a nonfungible token (NFT) may seem like a secure and v
erifiable way to sell art online.rnIn some ways, it is. Blockchain inh "},
        {"title":"After contentious debate, FDA approves first Alzheimers drug since 2003","content":"On Monday, the US Food and Drug Administration granted approval to a keenlywatched
 Alzheimers drug, aducanumab, developed by the drugmaker Biogen. The decision to approve the drug, which was once ab "},
        {"title":"On a growth tear, DuckDuckGo reveals it picked up 100M in secondary investment last year","content":"Privacy tech continues cooking on gas. To wit: Nontracking search
```

Figure 4: MongoDB Data entry in local file (./Output/MongoArticlesProcessed/ MongoArticlesProcessed.json) [3]

References

[1] 35.226.207.33. 2021. *Spark Master at spark://assignment3-data-ms.us-central1-a.c.tribal-quasar-316422.internal:7077*. [online] Available at: <http://35.226.207.33:8080/> [Accessed 5 July 2021].

[2] Cloud.mongodb.com. 2021. *Cloud: MongoDB Cloud*. [online] Available at: <https://cloud.mongodb.com/v2/60da206dda3bb271226e1ead#metrics/replicaSet/60da21bb39e2ea2b93467554/explorer/myMongoNews/articles/find> [Accessed 5 July 2021].

[3] Console.cloud.google.com. 2021. *Google Cloud Platform*. [online] Available at: <https://console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/assignment3-data-ms?project=tribal-quasar-316422&rif_reserved> [Accessed 5 July 2021].

[4] Newsapi.org. 2021. *News API – Search News and Blog Articles on the Web*. [online] Available at: <https://newsapi.org/> [Accessed 5 July 2021].

[5] App.diagrams.net. 2021. *Flowchart Maker & Online Diagram Software*. [online] Available at: <https://app.diagrams.net/> [Accessed 5 July 2021].