Putting Twitter Data into the Cloud

DS 3010 - Case Study 1 Report

Team 9: Abigail Albuquerque, Aria Yan, Isabel Herrero Estrada, Megan Sin, Sandra Phan

Formula 1 is one of the most premier forms of car racing in the world. It has a worldwide viewership of over 430 million in 2020, gathering about a million viewers per race from just the United States alone. The industry is worth over $4 billion dollars. What makes F1 unique is its global following. Unlike say the NFL or the NBA in the US, F1 competitions occur globally and are followed by a wide range of people. With high revenue and high viewership, a large media following naturally follows suit. From Netflix shows to household names of world famous drivers such as Hamilton and Schumacher, the media cloud is filled with F1 data that can be easily accessed and analyzed to learn something new. Formula 1 is also a great area to look into as it's fairly narrowed down already. In fact, there are only ten teams competing in these races each season, making it much simpler for us as data scientists to look into. When sports like soccer have thousands of teams and hunderns of famous active players, F1 is able to provide us with an advantageous position. We have therefore decided to look into the Twitter data of Formula 1 and it's most known race drivers.

Problem 1:

We were tasked with getting a cloud database using MongoDB atlas. In our case, we wanted to gain access to the twitter database in order to complete our investigation. This was achieved fairly easily using the twitter login API. After installing and importing the twitter API we set up a twitter login function, *oauth_login*, with a developer account provided to us.

Problem 2:

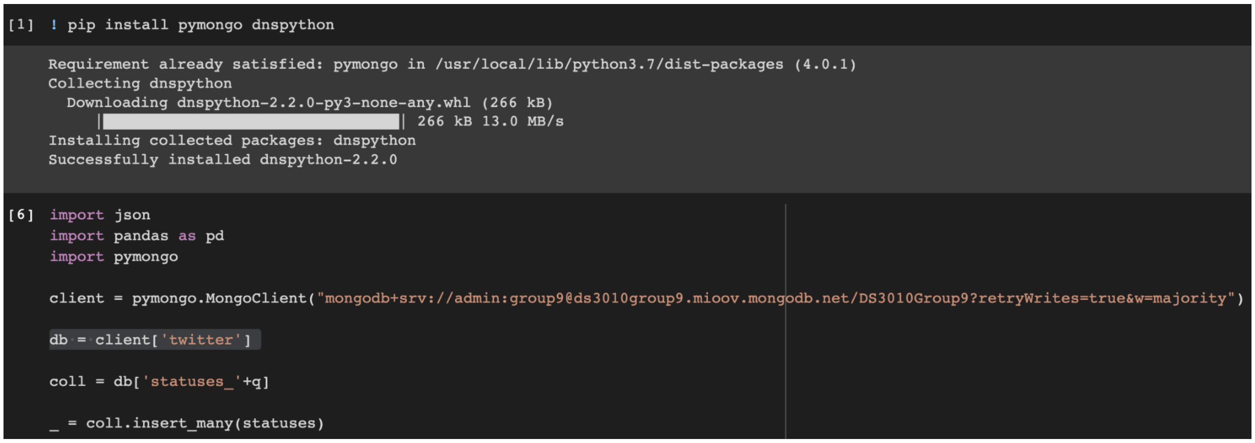
We were then tasked with reading twitter data into the cloud MongoDB data. We collected 1000 tweets. The list of keywords we chose to search for tweets from are:

> formula one, formula 1, f1, max verstappen, lewis hamilton, valterri bottas, sergio perez, carlos sainz, lando norris, charles leclerc, daniel ricciardo, pierre gasly, fernando alonso, esteban ocon,

sebastian vettel, lance stroll, yuki tsunoda, george russell, kimi raikkonen, nicholas latifi, antonio giovinazzi, mick schumacher, nikita mazepin, alex albon, guanyu zhou, fia

Since we are between seasons for the Formula 1, the keywords include the names of all drivers in last year's season as well as all driver’s in this year’s season. We also have a keyword ‘fia’ which stands for Fédération Internationale de l'Automobile, which is the governing body for the Formula One races.

After gathering tweets, we created a free cloud MongoDB database and connected it to our virtual machine so that we could read our Twitter data into it. We then installed and used the packages pymongo and dnspython to read and store all of the tweets we gathered into our database permanently. We could now access these gathered tweets anytime regardless if our virtual machine crashes or restarts.



```
[1]  ! pip install pymongo dnspython

     Requirement already satisfied: pymongo in /usr/local/lib/python3.7/dist-packages (4.0.1)
     Collecting dnspython
       Downloading dnspython-2.2.0-py3-none-any.whl (266 kB)
          |████████████████████████████████| 266 kB 13.0 MB/s
     Installing collected packages: dnspython
     Successfully installed dnspython-2.2.0

[6]  import json
     import pandas as pd
     import pymongo

     client = pymongo.MongoClient("mongodb+srv://admin:group9@ds3010group9.mioov.mongodb.net/DS3010Group9?retryWrites=true&w=majority")

     db = client['twitter']

     coll = db['statuses_'+q]

     _ = coll.insert_many(statuses)
```

Figure 1: Installing pymongo and dnspython to read and store tweets in the database permanently.

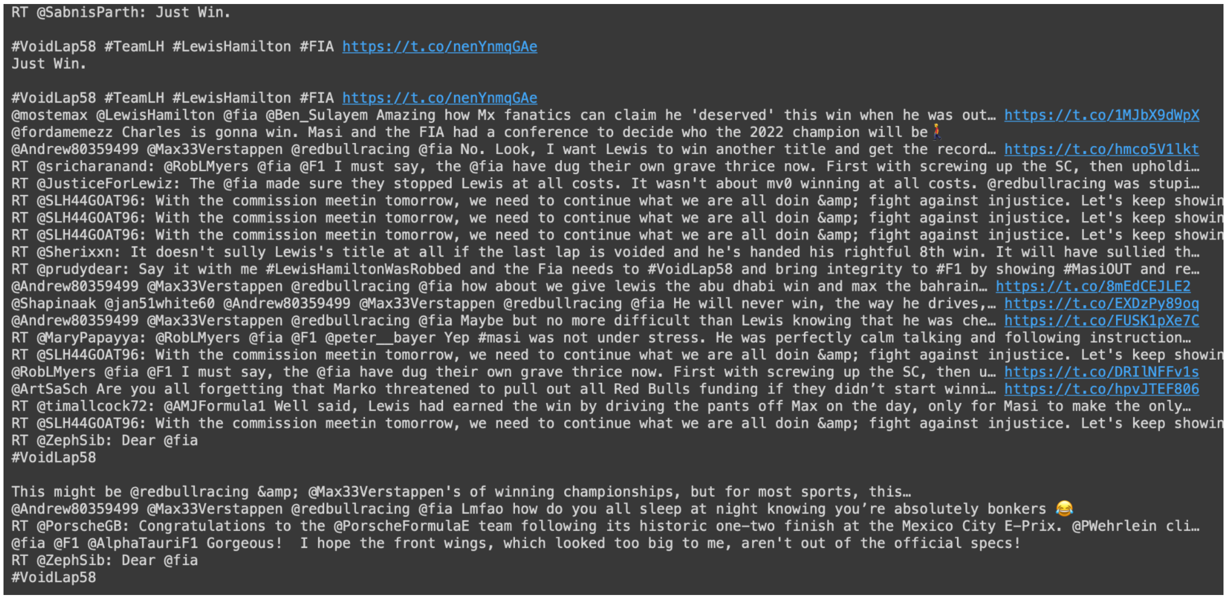
Problem 3:

Reading the data in MongoDB

To read the database, we first get a reference to a particular database, in our case it would be ‘twitter’. We reference a particular collection in the database, which would be ‘statuses_’.

Searching the data in MongoDB

We then searched the data in MongoDB for retweets. We also made a rough estimate of the number of tweets in our sample data. The number is ~3400. However, this is an overestimate as



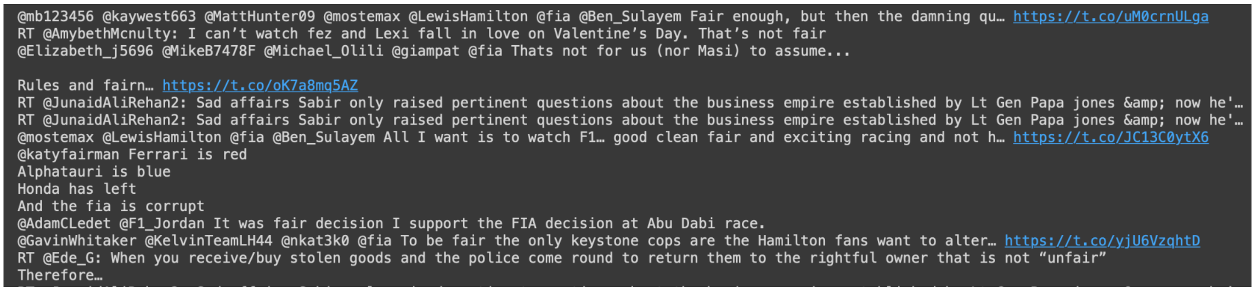
```
RT @SabnisParth: Just Win.

#VoidLap58 #TeamLH #LewisHamilton #FIA https://t.co/nenYnmqGAe
Just Win.

#VoidLap58 #TeamLH #LewisHamilton #FIA https://t.co/nenYnmqGAe
@mostemax @LewisHamilton @fia @Ben_Sulayem Amazing how Mx fanatics can claim he 'deserved' this win when he was out… https://t.co/1MJbX9dWpX
@fordamemezz Charles is gonna win. Masi and the FIA had a conference to decide who the 2022 champion will be🚶
@Andrew80359499 @Max33Verstappen @redbullracing @fia No. Look, I want Lewis to win another title and get the record… https://t.co/hmco5V1lkt
RT @sricharanand: @RobLMyers @fia @F1 I must say, the @fia have dug their own grave thrice now. First with screwing up the SC, then upholdi…
RT @JusticeForLewiz: The @fia made sure they stopped Lewis at all costs. It wasn't about mv0 winning at all costs. @redbullracing was stupi…
RT @SLH44GOAT96: With the commission meetin tomorrow, we need to continue what we are all doin &amp; fight against injustice. Let's keep showin
RT @SLH44GOAT96: With the commission meetin tomorrow, we need to continue what we are all doin &amp; fight against injustice. Let's keep showin
RT @SLH44GOAT96: With the commission meetin tomorrow, we need to continue what we are all doin &amp; fight against injustice. Let's keep showin
RT @Sherixxn: It doesn't sully Lewis's title at all if the last lap is voided and he's handed his rightful 8th win. It will have sullied th…
RT @prudydear: Say it with me #LewisHamiltonWasRobbed and the Fia needs to #VoidLap58 and bring integrity to #F1 by showing #MasiOUT and re…
@Andrew80359499 @Max33Verstappen @redbullracing @fia how about we give lewis the abu dhabi win and max the bahrain… https://t.co/8mEdCEJLE2
@Shapinaak @jan51white60 @Andrew80359499 @Max33Verstappen @redbullracing @fia He will never win, the way he drives,… https://t.co/EXDzPy89oq
@Andrew80359499 @Max33Verstappen @redbullracing @fia Maybe but no more difficult than Lewis knowing that he was che… https://t.co/FUSK1pXe7C
RT @MaryPapayya: @RobLMyers @fia @F1 @peter__bayer Yep #masi was not under stress. He was perfectly calm talking and following instruction…
RT @SLH44GOAT96: With the commission meetin tomorrow, we need to continue what we are all doin &amp; fight against injustice. Let's keep showin
@RobLMyers @fia @F1 I must say, the @fia have dug their own grave thrice now. First with screwing up the SC, then u… https://t.co/DRIlNFFv1s
@ArtSaSch Are you all forgetting that Marko threatened to pull out all Red Bulls funding if they didn’t start winni… https://t.co/hpvJTEF806
RT @timallcock72: @AMJFormula1 Well said, Lewis had earned the win by driving the pants off Max on the day, only for Masi to make the only…
RT @SLH44GOAT96: With the commission meetin tomorrow, we need to continue what we are all doin &amp; fight against injustice. Let's keep showin
RT @ZephSib: Dear @fia
#VoidLap58

This might be @redbullracing &amp; @Max33Verstappen's of winning championships, but for most sports, this…
@Andrew80359499 @Max33Verstappen @redbullracing @fia Lmfao how do you all sleep at night knowing you’re absolutely bonkers 😂
RT @PorscheGB: Congratulations to the @PorscheFormulaE team following its historic one-two finish at the Mexico City E-Prix. @PWehrlein cli…
@fia @F1 @AlphaTauriF1 Gorgeous!  I hope the front wings, which looked too big to me, aren't out of the official specs!
RT @ZephSib: Dear @fia
#VoidLap58
```

Figure 2: Retweets collected while searching MongoDB



```
@mb123456 @kaywest663 @MattHunter09 @mostemax @LewisHamilton @fia @Ben_Sulayem Fair enough, but then the damning qu… https://t.co/uM0crnULga
RT @AmybethMcnulty: I can’t watch fez and Lexi fall in love on Valentine’s Day. That’s not fair
@Elizabeth_j5696 @MikeB7478F @Michael_Olili @giampat @fia Thats not for us (nor Masi) to assume...

Rules and fairn… https://t.co/oK7a8mq5AZ
RT @JunaidAliRehan2: Sad affairs Sabir only raised pertinent questions about the business empire established by Lt Gen Papa jones &amp; now he'…
RT @JunaidAliRehan2: Sad affairs Sabir only raised pertinent questions about the business empire established by Lt Gen Papa jones &amp; now he'…
@mostemax @LewisHamilton @fia @Ben_Sulayem All I want is to watch F1… good clean fair and exciting racing and not h… https://t.co/JC13C0ytX6
@katyfairman Ferrari is red
Alphatauri is blue
Honda has left
And the fia is corrupt
@AdamCLedet @F1_Jordan It was fair decision I support the FIA decision at Abu Dabi race.
@GavinWhitaker @KelvinTeamLH44 @nkat3k0 @fia To be fair the only keystone cops are the Hamilton fans want to alter… https://t.co/yjU6VzqhtD
RT @Ede_G: When you receive/buy stolen goods and the police come round to return them to the rightful owner that is not “unfair”
Therefore…
```

Figure 3: Retweets collected while searching MongoDB, duplication because MongoDB 3 does not allow for removal of duplicates while searching.

Problem 4:

- Come up with a business question that Twitter data and MongoDB could help answer.

Our business question was: How can we use the frequency of the keywords, specifically the driver’s names, to develop a product or solution that can be sold to companies?

- Describe the business case.

The thinking behind this was fairly straightforward. Our first example that sparked this idea for this question was the concept of two names of drivers frequently occurring together. Using MongoDB and the twitter API we would be able to identify the frequency of combinations of names, such as Verstappen and Hamilton. This would be extremely valuable data for certain groups of interest. Media outlets such as tabloids would find a lot of use knowing this as it would give them the upper hand on what combination of names the public is most interested in. Using this information they would then be able to spark drama between these specific names and theoretically achieve a high success rate with their media campaigns targeting this, thus having profits. Drama sells, and in the right hands this information could be very useful in achieving it.

Another great example of a group of interest in this business case would be Formula 1 teams themselves. By knowing the frequency of the drivers showing up on social media, they can gauge the popularity of their drivers. Whether winning or losing, at the end of the day, fans are what drive in money for these teams. Ferrari does not only pick its race drivers because they perform best on paper, but also due to how likable and popular they are among F1 fans. In our current day and age, this type of data can be found in media databases such as twitter. In this case, the product could be a Twitter analysis of a teams roster of drivers, to help decide whether it is worth it to resign driver contracts or to start looking at other options.

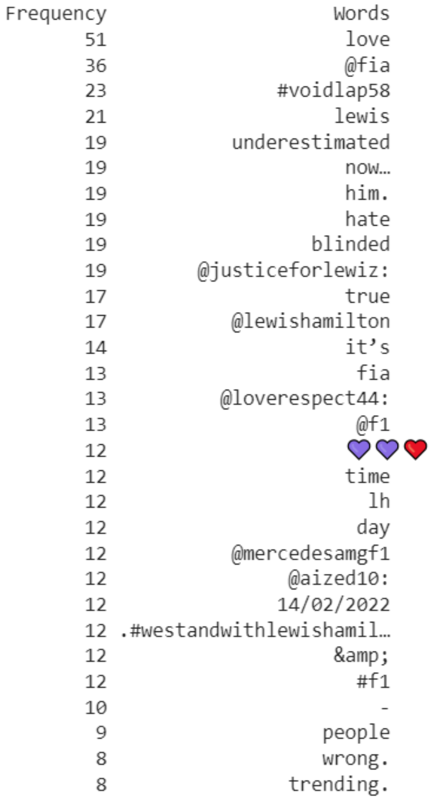
- How could MongoDB help a company to scale up its computation environment?

Historically, scaling databases has been a well known issue for companies. The number of options for scalability have traditionally been quite limited and also often too expensive to be worth the investment. Fortunately, MongoDB has provided advancements in the area of scalability to help companies achieve their goals as they grow. Firstly, vertical scaling can be performed in companies starting to expand their database through MongoDB. An example of this could be that a company increases their server power by adding more CPUs. While vertical scaling is a great feature initially, it can get quite expensive. By being non-SQL, MongoDB can provide more advanced methods of scaling. In MongoDB, data can be spread out across many nodes to achieve scalability, known as sharding. By virtually scaling the capacity, each node has much less work to do, so reads and writes overall will be processed at a much greater capacity.

- Business Question Exploration

Some additional experiments we ran with our data to gain familiarity with MongoDB and see how our business idea could be implemented.

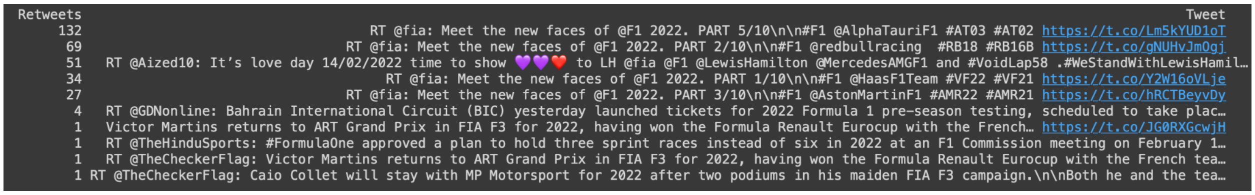
We filtered the stopwords to conduct a frequency analysis based off of tweets that have the word 'love' in them.



```
Frequency                    Words
       51                     love
       36                     @fia
       23               #voidlap58
       21                    lewis
       19           underestimated
       19                     now…
       19                     him.
       19                     hate
       19                  blinded
       19        @justiceforlewiz:
       17                     true
       17           @lewishamilton
       14                     it's
       13                      fia
       13            @loverespect44:
       13                      @f1
       12                      💜💜❤️
       12                     time
       12                       lh
       12                      day
       12           @mercedesamgf1
       12                @aized10:
       12               14/02/2022
       12 .#westandwithlewishamil…
       12                     &amp;
       12                      #f1
       10                        -
        9                   people
        8                   wrong.
        8                trending.
```

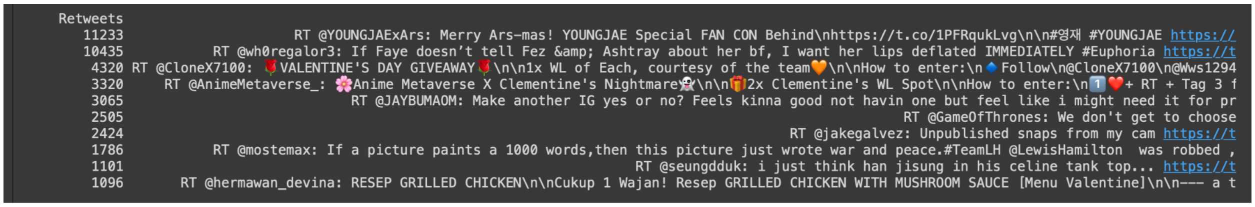
Figure 3: Plot of top 30 words based on frequency

We also collected different samples of retweets and sorted them by frequency to see how frequency could be used to find out trends, popular drivers and popular news topics.



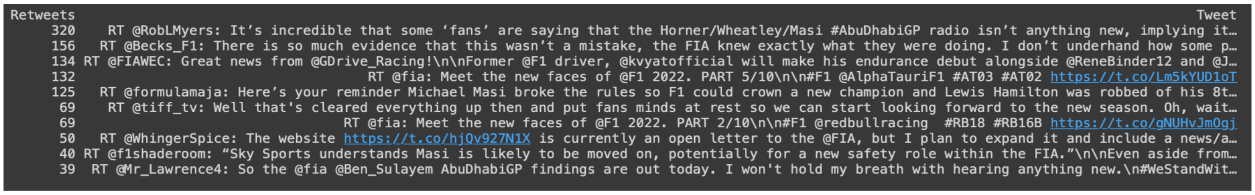
```
Retweets                                                                                                                      Tweet
     132                                        RT @fia: Meet the new faces of @F1 2022. PART 5/10\n\n#F1 @AlphaTauriF1 #AT03 #AT02 https://t.co/Lm5kYUD1oT
      69                                     RT @fia: Meet the new faces of @F1 2022. PART 2/10\n\n#F1 @redbullracing  #RB18 #RB16B https://t.co/gNUHvJmOgj
      51   RT @Aized10: It's love day 14/02/2022 time to show 💜💜❤️ to LH @fia @F1 @LewisHamilton @MercedesAMGF1 and #VoidLap58 .#WeStandWithLewisHamil…
      34                                          RT @fia: Meet the new faces of @F1 2022. PART 1/10\n\n#F1 @HaasF1Team #VF22 #VF21 https://t.co/Y2W16oVLje
      27                                     RT @fia: Meet the new faces of @F1 2022. PART 3/10\n\n#F1 @AstonMartinF1 #AMR22 #AMR21 https://t.co/hRCTBeyvDy
       4   RT @GDNonline: Bahrain International Circuit (BIC) yesterday launched tickets for 2022 Formula 1 pre-season testing, scheduled to take plac…
       1   Victor Martins returns to ART Grand Prix in FIA F3 for 2022, having won the Formula Renault Eurocup with the French… https://t.co/JG0RXGcwjH
       1   RT @TheHinduSports: #FormulaOne approved a plan to hold three sprint races instead of six in 2022 at an F1 Commission meeting on February 1…
       1   RT @TheCheckerFlag: Victor Martins returns to ART Grand Prix in FIA F3 for 2022, having won the Formula Renault Eurocup with the French tea…
       1 RT @TheCheckerFlag: Caio Collet will stay with MP Motorsport for 2022 after two podiums in his maiden FIA F3 campaign.\n\nBoth he and the tea…
```

Figure 4: Retweets collected based on frequency for data exploration. Keyword: '2022'



```
   Retweets
   11233                     RT @YOUNGJAExArs: Merry Ars-mas! YOUNGJAE Special FAN CON Behind\nhttps://t.co/1PFRqukLvg\n\n#영재 #YOUNGJAE https://
   10435           RT @wh0regalor3: If Faye doesn’t tell Fez &amp; Ashtray about her bf, I want her lips deflated IMMEDIATELY #Euphoria https://t
    4320 RT @CloneX7100: 🌹VALENTINE'S DAY GIVEAWAY🌹\n\n1x WL of Each, courtesy of the team🧡\n\nHow to enter:\n🔷Follow\n@CloneX7100\n@Wws1294
    3320     RT @AnimeMetaverse_: 🌸Anime Metaverse X Clementine's Nightmare👻\n\n🎁2x Clementine's WL Spot\n\nHow to enter:\n1️⃣❤️+ RT + Tag 3 f
    3065                            RT @JAYBUMAOM: Make another IG yes or no? Feels kinna good not havin one but feel like i might need it for pr
    2505                                                                                                    RT @GameOfThrones: We don't get to choose
    2424                                                                                      RT @jakegalvez: Unpublished snaps from my cam https://t
    1786           RT @mostemax: If a picture paints a 1000 words,then this picture just wrote war and peace.#TeamLH @LewisHamilton  was robbed ,
    1101                                                                 RT @seungdduk: i just think han jisung in his celine tank top... https://t
    1096       RT @hermawan_devina: RESEP GRILLED CHICKEN\n\nCukup 1 Wajan! Resep GRILLED CHICKEN WITH MUSHROOM SAUCE [Menu Valentine]\n\n--- a t
```

Figure 5: Retweets collected based on frequency for data exploration. All tweets in database.



```
Retweets                                                                                                                                  Tweet
     320   RT @RobLMyers: It’s incredible that some ‘fans’ are saying that the Horner/Wheatley/Masi #AbuDhabiGP radio isn’t anything new, implying it…
     156  RT @Becks_F1: There is so much evidence that this wasn’t a mistake, the FIA knew exactly what they were doing. I don’t underhand how some p…
     134 RT @FIAWEC: Great news from @GDrive_Racing!\n\nFormer @F1 driver, @kvyatofficial will make his endurance debut alongside @ReneBinder12 and @J…
     132                                    RT @fia: Meet the new faces of @F1 2022. PART 5/10\n\n#F1 @AlphaTauriF1 #AT03 #AT02 https://t.co/Lm5kYUD1oT
     125  RT @formulamaja: Here’s your reminder Michael Masi broke the rules so F1 could crown a new champion and Lewis Hamilton was robbed of his 8t…
      69   RT @tiff_tv: Well that's cleared everything up then and put fans minds at rest so we can start looking forward to the new season. Oh, wait…
      69                                 RT @fia: Meet the new faces of @F1 2022. PART 2/10\n\n#F1 @redbullracing  #RB18 #RB16B https://t.co/gNUHvJmOgj
      50  RT @WhingerSpice: The website https://t.co/hjQv927N1X is currently an open letter to the @FIA, but I plan to expand it and include a news/a…
      40 RT @f1shaderoom: “Sky Sports understands Masi is likely to be moved on, potentially for a new safety role within the FIA.”\n\nEven aside from…
      39  RT @Mr_Lawrence4: So the @fia @Ben_Sulayem AbuDhabiGP findings are out today. I won't hold my breath with hearing anything new.\n#WeStandWit…
```

Figure 6: Retweets collected based on frequency for data exploration. Keyword: ‘new’

References

Lange, Author: David. “Number of TV viewers Formula One (F1) Racing from 2008-2020” *Statista.com*, 19 Mar. 2021, https://www.statista.com/statistics/480129/cable-or-broadcast-tv-networks-formula-one-f1-racing-watched-within-the-last-12-months-usa/.