MiniProject3 DataBase

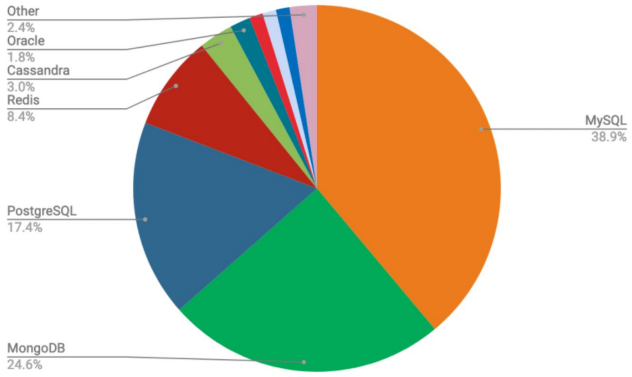
YueLiu

1 Review major used database systems in cloud

(e.g, SQL/mySQL, Mongo) 34-58

1.1 Basic technologies

A research by Scalegrid showed that MySQL and MongoDB are the most popular databases.We will discuss it in the later implementation.



(source:https://dev.to/scalegrid/2019-database-trends--sql-vs-nosql-top-databases-single-vs-multiple-database-use-1nma)

The following table compares the properties of Firebase, Redis and Cassandra. Redis supports the most programming languages with in memory capabilities; Firebase supports Android systems and has a similar product Google Analytics; Cassandra is earliest to start with map reduce ability.

Name	**Firebase**	**Redis**	**Cassandra**
License	Commercial	BSD 3-clause	Apache License 2.0
Written in	NA	ANSI, C	JAVA
Stable Release	NA	5.0.7(2019-11)	3.11.5(2019-10)
Developed by	Firebase Inc; Google	Redis Labs	Apache Software Foundation

Type	NoSQL DBMS; Document Store	NoSQL DBMS; Key-Value Store	NoSQL DBMS(SQL like); Wide Column Store

1.2 Who uses them?

According to Elmasri, R., & Navathe, S. (2017), there are 3 types of users: database administrators, database designers and end users. DBA is responsible for authorizing access; designers are to identify the data and select the proper method; the end users are people who use them in jobs.

In the following examples, we start from the user side, because the data is not structured very well. It requires optimization of the API from time to time.

1.3 When do you use them?

To visualize the result, we need to do more operations in the collected data, which can be realized by the database.

There is a popular post in the StackOverflow forum. It summarizes the advantages of a database. If your projects have the following requirement, you can use the database.

1. You can query data in a database (ask it questions).
2. You can look up data from a database relatively rapidly.
3. You can relate data from two different tables together using JOINs.
4. You can create meaningful reports from data in a database.
5. Your data has a built-in structure to it.
6. Information of a given type is always stored only once.
7. Databases are ACID..
8. Databases are fault-tolerant.
9. Databases can handle very large data sets.
10. Databases are concurrent; multiple users can use them at the same time without corrupting the data.
11. Databases scale well.

In one word, do not use the database if it costs too much.

2 What is the difference between them?

To do further analysis of tweets, we cannot rely on spreadsheet anymore, because what we post is more than numbers. If we collect the live tweets, there should be infinity target, and we can only use database to store them.

	SQL	**NoSQL**
pros	Great storage solution (servers); The best solution for structured data and transactional needs; Can be accessed by many users at the same time.	Very fast; Not requiring fixed table schemas; Scales horizontally.
cons	Not the best solution when dealing with data growing exponentially (ex. social media); Need for deep expertise of programming skills: steep learning curve(C++).	Not transactional/ACID; Can get messy.

Later, we will use a combination of SQL and NoSQL. MySQL is an open-source relational database management system (RDBMS). MySQL is written in C and C++. Its SQL parser is written in yacc, but it uses a home-brewed lexical analyzer. The latest version is released in 14 October 2019. MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema, which is written in C++, Go, JavaScript, Python. The latest version released in 11 October 2019.

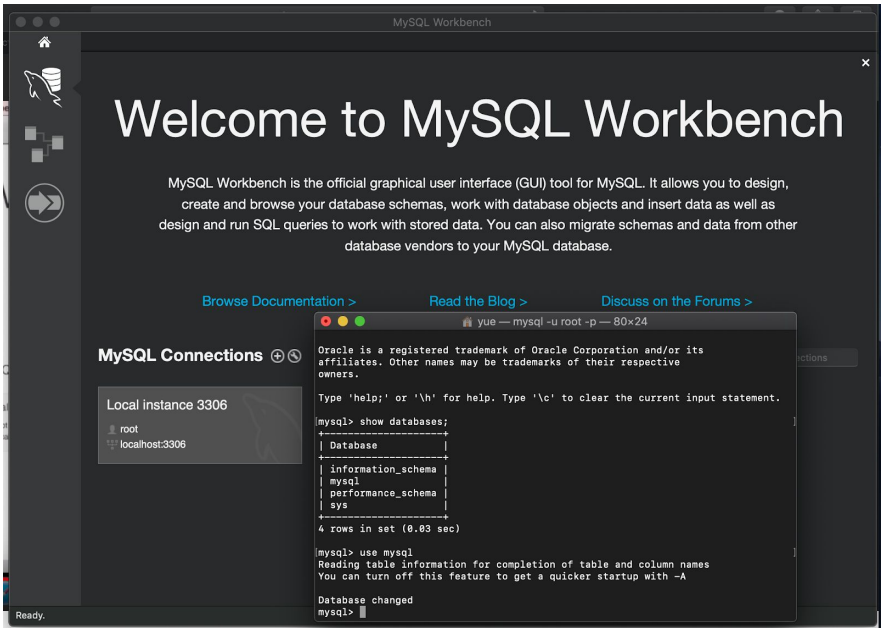
3 Implementation

3.1 Twitter API+MySQL: track and store published tweets

Unlike the SQL lite, which is proper for personal user, MySQL is good for a group of users.To avoid the ambiguous in syntax, I use MySQL as the starting point in DBS. This is what you will see after successfully install the MySQL community server.

```
(base) YUEs-Air:~ yue$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 9
Server version: 8.0.18 MySQL Community Server - GPL

Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

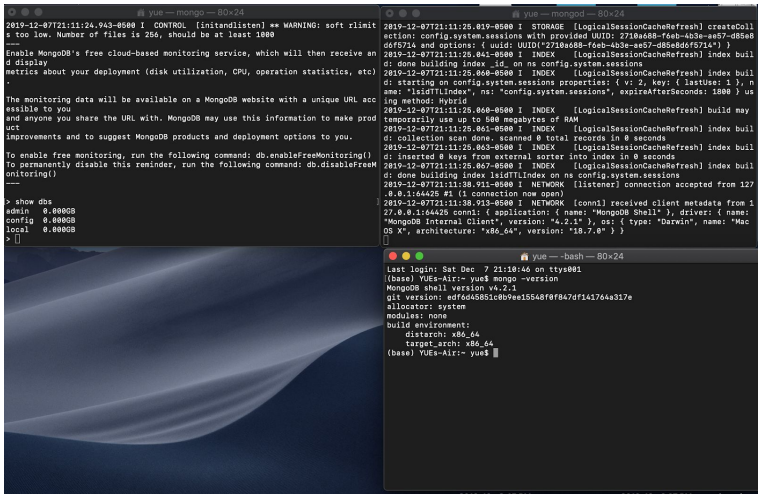
There are two ways to use the MySQL server. I tried the command show and use for the default database.



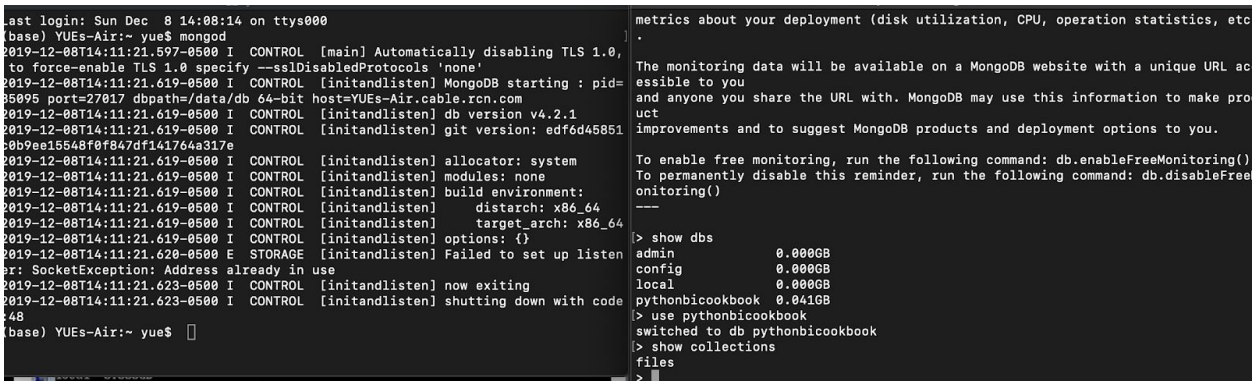
After installing MySQL server, we use the code in mini project 1 to create two new ".csv" files about the vogue and elle in a small scale.

3.2 Twitter API+MongoDB: track and store the live stream tweets

If you can see the database by default and the version, the MongoDB is installed successfully in your laptop.



After running the code 3.2.2, we import a new "‘.json" file into the server, named pythonbicookbook..



Normally, MongoDB is perfect to store large live data, especially for the primary key(user_id) in our data. I shrink the size of the data in the next step.

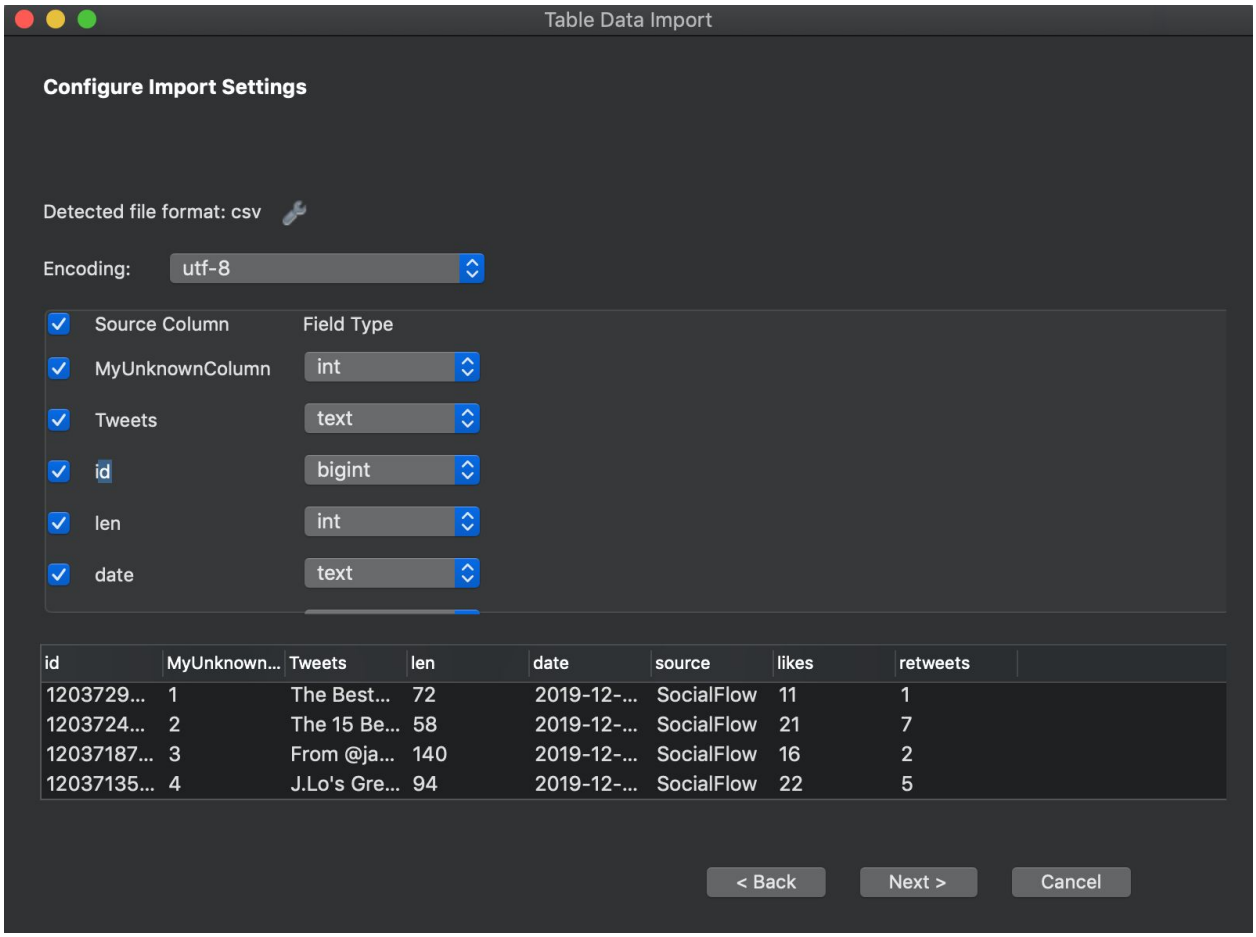
3.3 Convert the sentiments in Google NLP: from MySQL to Mongo

We will only use ‘tweet_vogue.csv’ and ‘tweet_elle.csv’ in the following. In the sentiment analysis, the result of Vogue is mixed, while that of Elle is positive.

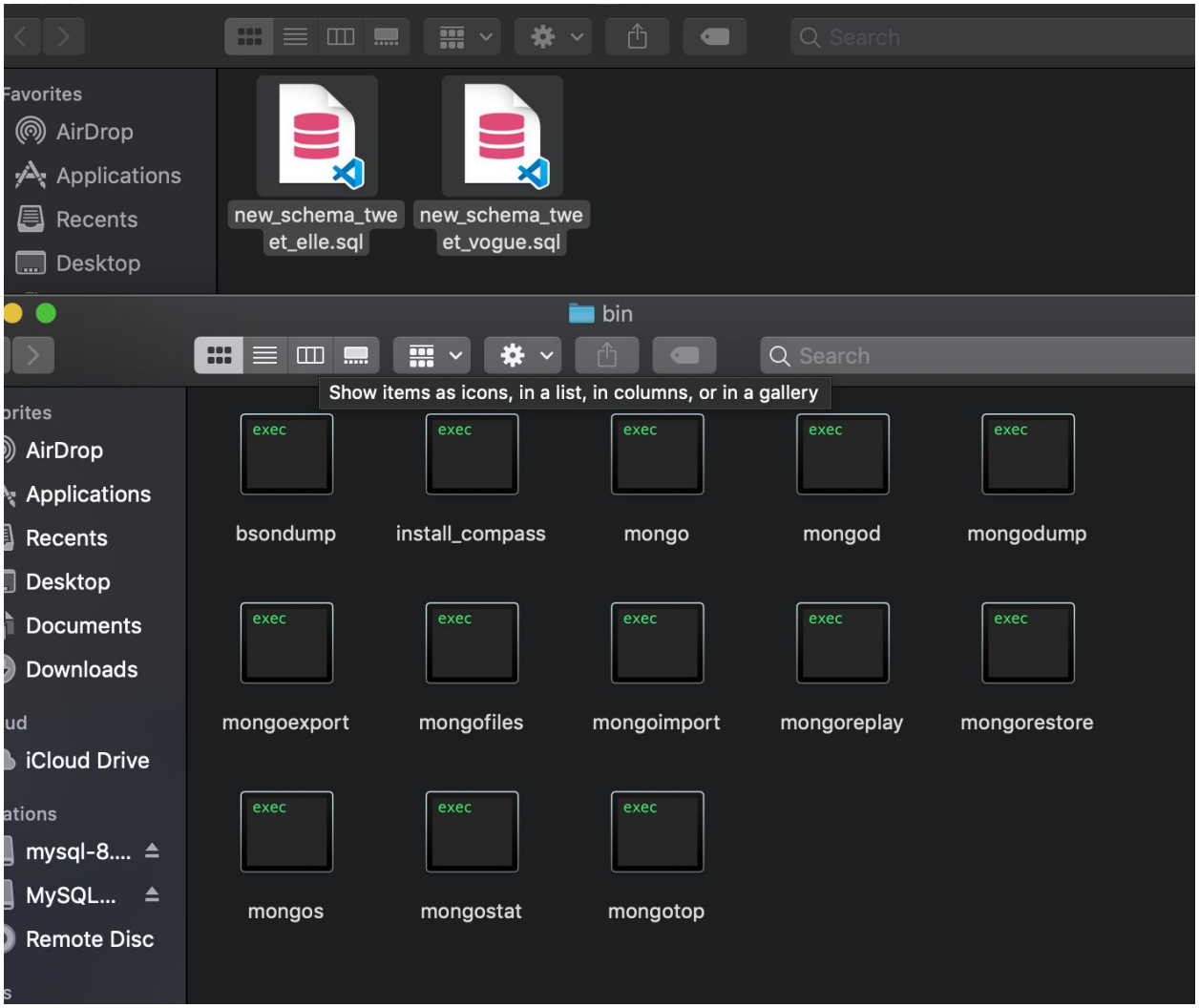
```
PyDev console: starting.
Python 3.7.4 (v3.7.4:e09359112e, Jul  8 2019, 14:54:52)
[Clang 6.0 (clang-600.0.57)] on darwin
runfile('/Users/yue/Desktop/BU Grad Courses/19fallEC601/MiniProject/MiniProject3/Code/3.3GoogleNLP+MySQL+MongoDB/331GoogleNLPsentiment.py', wdir='/Users/yue/Desktop/BU Grad Courses/19fallEC601/MiniProject/MiniProject3/Code/3.3GoogleNLP+MySQL+MongoDB')
Text: ,Tweets.id,len,date,source,likes,retweets
0,A very important update for @TheCrownNetflix season 4. https://t.co/QpIFX48Ckf,1203732285759066117,78,2019-12-08 17:45:15,SocialFlow,61,8
1,The GQ Men of the Year Awards turned out to be an exhibition of creative suiting for nearly every attendee. https://t.co/AyXyJ9HyKk,1203727152811778048,131,2019-12-08 17:24:52,SocialFlow,22,4
2,Hailey Bieber has a knack for making the buttoned-up two piece feel less corporate. https://t.co/4LyClHN2zP,1203721945579184128,107,2019-12-08 17:04:10,SocialFlow,55,2
3,""Van Cleef &amp;amp; Arpels: Time, Nature, Love" is staged at the Palazzo Reale, a former royal palace. https://t.co/mWWAenNPh4",1203716816276742144,123,2019-12-08 16:43:47,SocialFlow,70,14
4,"Solange showed off her teased-out mane, Kelsey Lu revealed a bold new look, and Lady Gaga put her high-shine lip gl… https://t.co/bDGgwpzgzd",1203711360095465473,140,2019-12-08 16:22:06,SocialFlow,47,5
5,The perfect red lipstick! The legendary skin elixir! The world's most famous hair brush! https://t.co/7yAOCmNKZP,1203706019874050048,112,2019-12-08 16:00:53,SocialFlow,70,12
6,Katie Holmes showed how to dress up the humble winter boot for city streets. https://t.co/STH9W4Ctlx,1203700801182601216,100,2019-12-08 15:40:09,SocialFlow,63,11
7,"Los Angeles—it's a city of dreams, even for known sociopaths. https://t.co/hcUl3qyZk7",1203695513452847104,85,2019-12-08 15:19:08,SocialFlow,110,17
8,"""Every time a twenty-something I know returns from a holiday with a certain glow and that definitive diamond on her… https://t.co/0q8K95Rofx",1203690260737265665,140,2019-12-08 14:58:16,SocialFlow,52,13
9,"FLWRDWN's capsule of puffers is filled with natural wildflowers, a rare sustainable solution that's neither cruel t… https://t.co/Y02LxOYTR6",1203685196534550529,140,2019-12-08 14:38:08,SocialFlow,184,52
10,"Go inside the star-studded 2019 British Fashion Awards, hosted by Tracee Ellis Ross. https://t.co/JQILniSAbw",1203679961118237824,108,2019-12-08 14:17:20,SocialFlow,72,11
11,The 15th annual UNICEF Snowflake Ball gathered hundreds of luminaries and philanthropists to raise $4.95 million. https://t.co/QPhwh2tN8q,1203674578922414081,137,2019-12-08 13:55:57,SocialFlow,75,15
12,"This holiday, there are a number of new books that make perfect gifts for those looking to spruce up their space an… https://t.co/uGo6vJUOGw",1203668261174136832,140,2019-12-08 13:34:49,SocialFlow,81,20
13,"Exclusively for the #VogueShop, @publicschoolnyc's designers have created four limited-edition shirts with a messag… https://t.co/ebD5CgpTZD",1203664016469626881,140,2019-12-08 13:13:59,SocialFlow,28,4
14,Is picking out presents for the fashion-inclined in your life proving to be a challenge? Don't fret–the #VogueShop… https://t.co/a2BcrTAleq,1203658665770864641,139,2019-12-08 12:52:43,SocialFlow,37,8
15,There's a secret nobody tells you about getting engaged: you'll probably lose a friend (or two) before the cake is… https://t.co/rO55KcmbLv,1203653438216835072,139,2019-12-08 12:31:57,SocialFlow,163,31
16,".@TroyeSivan spends a rare day off with us answering #73Questions.
https://t.co/GYWvPt4PFe",1203648109952570146,91,2019-12-08 12:10:46,SocialFlow,58,8
17,"Inside the world of @tylerthecreator, hip hop's most unlikely style icon. https://t.co/S5pjwRSk19",1203643058904084480,97,2019-12-08 11:50:42,SocialFlow,78,12
18,Step inside the #VogueShop. Get all the details on our brand-new line of merch here. https://t.co/NvShV7ONmX,1203637576395116545,108,2019-12-08 11:28:55,SocialFlow,19,0
19,"With a brand-new line of merchandise, #VogueShop has something for everyone on your holiday gift list.… https://t.co/2G2bYzUHwP",1203632123229483008,127,2019-12-08 11:07:15,SocialFlow,62,10
20,Kids these days want more than their two front teeth when it comes to holiday gifting. https://t.co/BHO7r6Zdgn,1203626921126445061,110,2019-12-08 10:46:34,SocialFlow,44,5
Sentiment: 0.20000000298023224, 8.5
Mixed
```

```
0         1         2         3         4         5         6         7         8         9         10        11        12        13        14        15        16
Text: ,Tweets.id,len,date,source,likes,retweets
0,Everlane Is Now Making Knee-High Boots (Yay) https://t.co/sW8LpNnKuc,1203734612289175553,68,2019-12-08 17:54:30,SocialFlow,12,1
1,The Best Nintendo Switch Games for Casual Gamers https://t.co/Pet1jr3y6O,1203729489173323781,72,2019-12-08 17:34:09,SocialFlow,11,1
2,The 15 Best Carly Rae Jepsen Songs https://t.co/PShubLPzef,1203724232162349574,58,2019-12-08 17:13:37,SocialFlow,21,7
3,"From @janetmock to @issarae, a wave of female-fronted projects is rising up out of Hollywood, and these 12 women ar… https://t.co/um2brqFdOs",1203718783250182144,140,2019-12-08 16:51:36,SocialFlow,16,2
4,"J.Lo's Green Versace Gown Made a Surprise Appearance on SNL Last Night
https://t.co/1MVXCTz2Cg",1203713501170196485,94,2019-12-08 16:30:37,SocialFlow,22,5
5,"Here's What Gabrielle Union Says About Her Departure from 'America's Got Talent'
https://t.co/thyLm0UFGq",1203708759736012801,104,2019-12-08 16:11:46,SocialFlow,16,4
6,Everything You Need To Know About #StrangerThings Season 4 https://t.co/YPXazK6038,1203703194083299328,82,2019-12-08 15:49:39,SocialFlow,25,3
7,"If You Get One Spa Treatment in Hawaii, Please Make It the Maui Mahulia https://t.co/a325mQasNk",1203697988784312320,95,2019-12-08 15:28:58,SocialFlow,14,3
8,Gordo Will Return in the Lizzie McGuire Reboot #DisneyPlus https://t.co/U8ly0UXXny,1203692943091097600,82,2019-12-08 15:08:55,SocialFlow,52,8
9,The Best Thongs According to https://t.co/JbtHbxOl6t Editors https://t.co/ANV3SSRPuU,1203687784646623232,84,2019-12-08 14:48:25,SocialFlow,16,4
10,9 https://t.co/JbtHbxOl6t Editors Confess What They Bought at Net-A-Porter's Massive Sale https://t.co/OXoeIulpd6,1203682504722518016,113,2019-12-08 14:27:27,SocialFlow,17,4
11,10 Easy Ways to Be More Sustainable at Work and Home https://t.co/ts89eW9rk5,1203677089301516288,76,2019-12-08 14:05:55,SocialFlow,17,6
12,"In our latest Style File,@yunamusic talks about developing her iconic modest style, why making a fashion faux pas i… https://t.co/oihvLRCXcG",1203671674455035904,140,2019-12-08 13:44:24,SocialFlow,14,3
13,Watch Sara Sampaio's 10-Minute Morning Routine https://t.co/2aWbBfzLZw,1203666442316664833,70,2019-12-08 13:23:37,SocialFlow,15,4
14,Gabriela Hearst to Donate 100 Percent of Proceeds to Help Children in Yemen https://t.co/MqGDS69MHv,1203661098093858820,99,2019-12-08 13:02:23,SocialFlow,52,7
15,Guess who's back? A new season of @projectrunway is here with ELLE Editor-in-Chief @ninagarcia. Get to know all 16… https://t.co/2aRT3JRsfn,1203655866676846594,139,2019-12-08 12:41:36,SocialFlow,17,4
16,"14 Best Oil Diffusers to Help You Relax, Finally https://t.co/TMLMrEG9yZ",1203650568318832641,72,2019-12-08 12:20:32,SocialFlow,15,5
17,Model Charlotte McKinney Wakes Up With an 8-Step Skincare Routine https://t.co/8rkjVbKdXH,1203645329247215619,89,2019-12-08 11:59:43,SocialFlow,14,0
18,This Beauty Routine is the Secret to @justineskye's Perfect Skin https://t.co/QnhHqrS4lf,1203639890669457408,88,2019-12-08 11:38:07,SocialFlow,14,0
19,Designer Nafsika Skourti's Clothes Are Made to Transcend Borders https://t.co/MTTtgXT4Io,1203634813435043840,88,2019-12-08 11:17:56,SocialFlow,19,0
20,11 Women With Stressful Jobs on the Reality TV Shows That Help Them Unwind https://t.co/xym55pvZFh,1203629506659045376,98,2019-12-08 10:56:51,SocialFlow,16,4
Sentiment: 0.30000001192092896, 1.5
Positive
```

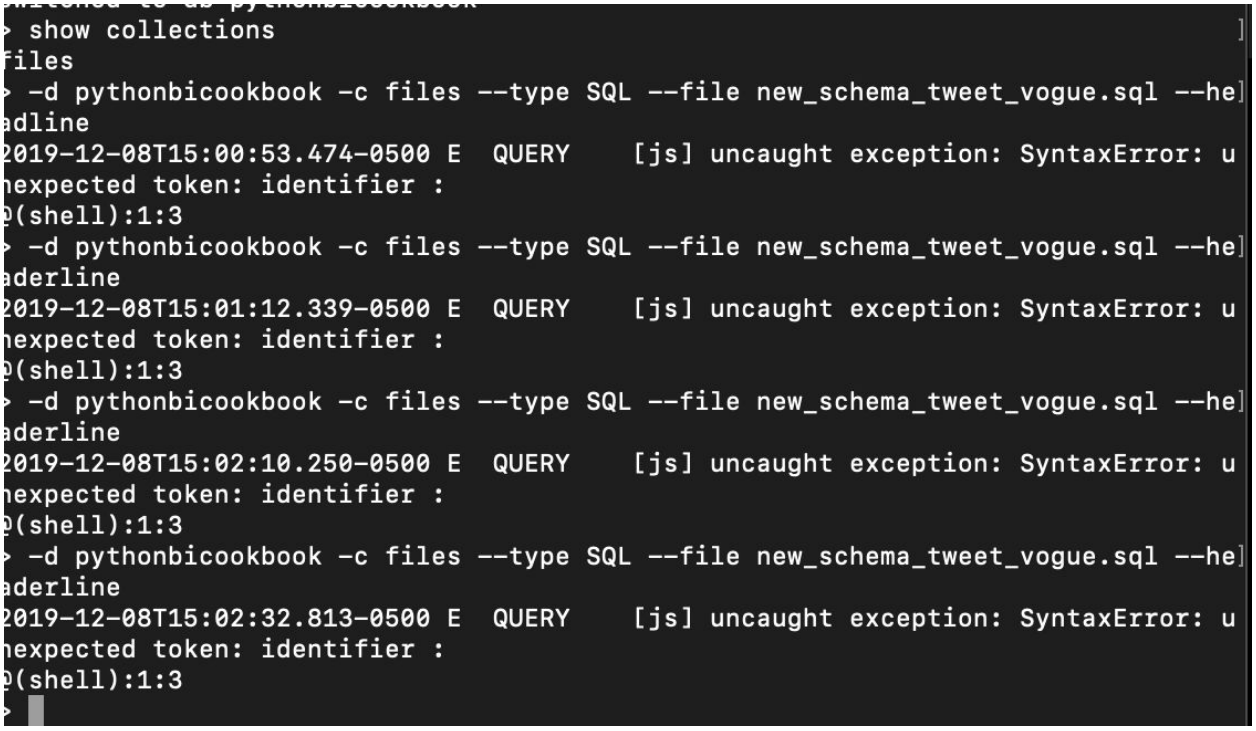
The tweets about elle worths to be stored in the mongoDB, because its size. We need to focus on the format in the future. To make up for this point(‘.txt format’), I use MySQL to change the sequence of user_id into the beginning.



After that, I import the two “.csv” file into mongoDB. All the commands used are shown in the screen shot. The important thing is to put them in the bin folder of mongoDB.

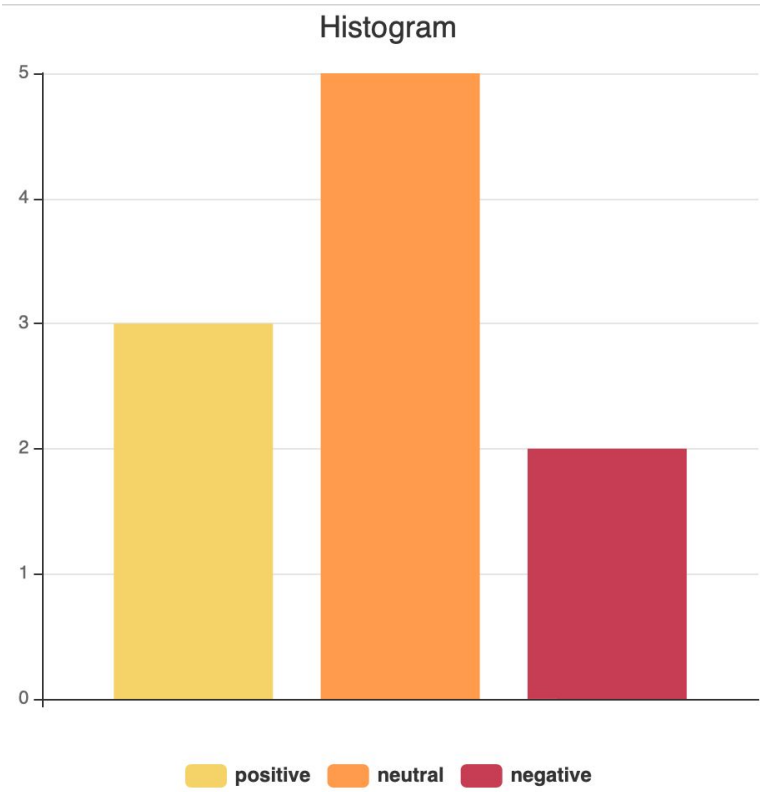


Currently, we cannot import .sql to the mongoDB, but we can do it successfully by using: (1) .json; (2) .csv. It means that the Mongo API is important to be applied in Python. Otherwise, we can only store the original data.



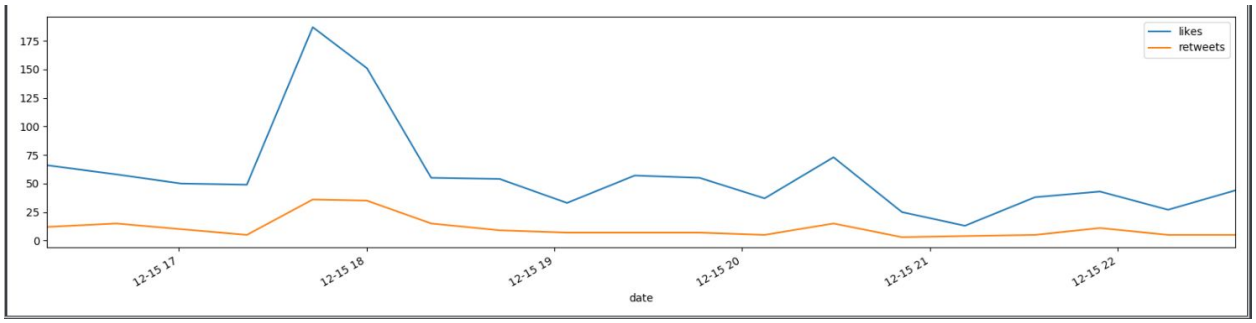
3.4 Statistics Result and Visualization for Tweets

The histogram can only gather the probability of the sentiment. As can be seen in the histogram, people have different attitudes about the tweets from Vogue.



Although it is a common way to calculate the statistical result, I don't think it make full use of the timeline functions in tweepy, which is the core operations of these two projects.

In the larger scale, we can visualize the dataset by using the .py code.

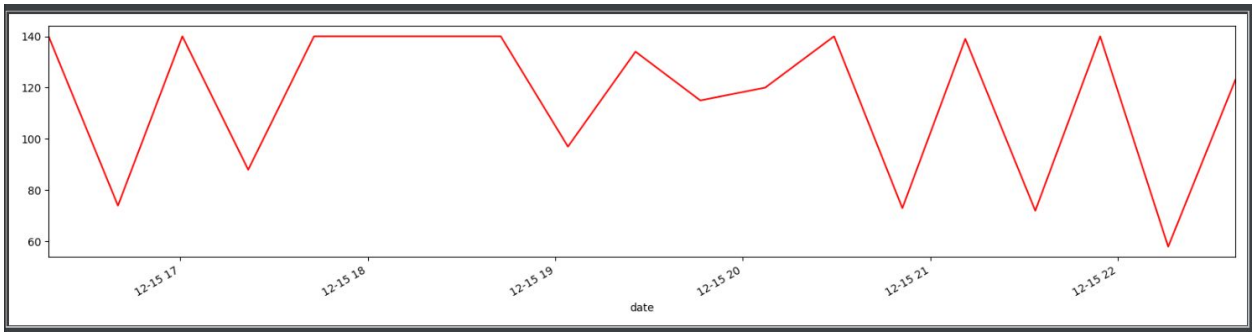


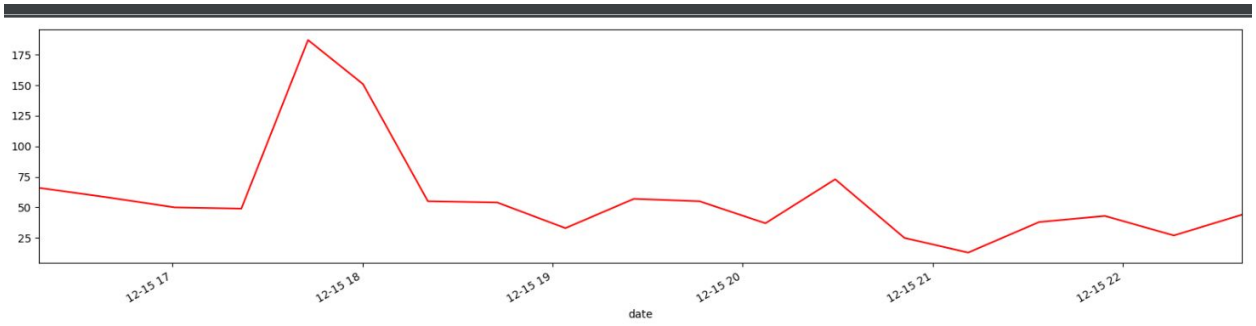
As can be seen from the multilayer plot, there is a large gap between retweets and like among the users, the maximum can be over 150. In Python, getting multilayer is easier than spreadsheet, because we can narrow down the legend by printing the results. Also, it is not easy to get the errors because we can build the plot layer by layer.

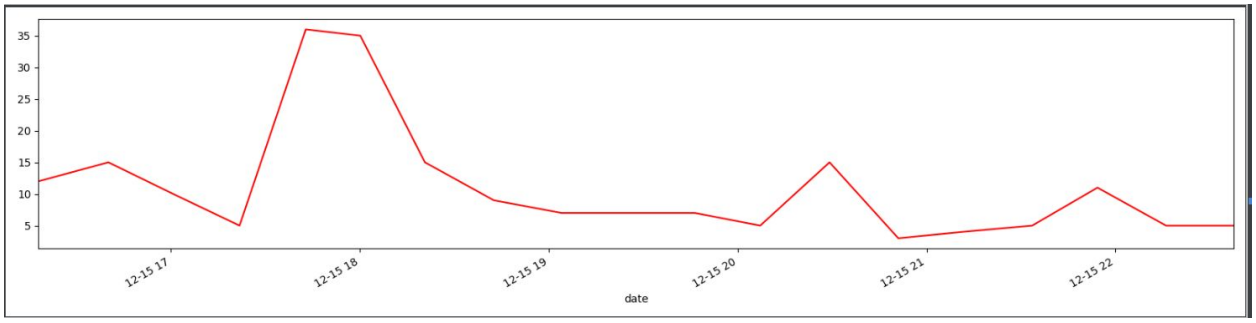
```
127    # Get average length over all tweets:
128    print(np.mean(df['len']))
129
130    # Get the number of likes for the most liked tweet:
131    print(np.max(df['likes']))
132
133    # Get the number of retweets for the most retweeted tweet:
134    print(np.max(df['retweets']))
135
```

Console × 341visualizeTweets ×

```
['__class__', '__delattr__', '__dict__', '__dir__', '__doc__', '__eq__',
5
116.47368421052632
187
36
```







References

[1] System Properties Comparison Cassandra vs. Firebase Realtime Database vs. Redis (n.d.) Retrieved from https://db-engines.com/en/system/Cassandra%3BFirebase+Realtime+Database%3BRedis

[2]Porlier,Marie.Josee.(2019, February 8) 4 Examples Of Database Application Retrieved from https://www.kohezion.com/blog/4-examples-database-application/

[3] Why use a database instead of just saving your data to disk?(n.d.) Retrieved from https://softwareengineering.stackexchange.com/questions/190482/why-use-a-database-instead-of-just-saving-your-data-to-disk

[4] Elmasri, R., & Navathe, S. (2017). *Fundamentals of database systems* (pp.3-29). Pearson.

[5] MySQL.(2019). Retrieved from

[6] MongoDB.(2019). Retrieved from https://en.wikipedia.org/wiki/MongoDB