# 7201 Project Report – Report on Dataset Analytics (Coursework)

## Structured abstract

Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things.This report will focus on those twitter mentioned about one of most popular popstar-Justin Bieber from 2014.10 to 2014.12,focus on general twitter trend of JB in this period and find out what people concerned about him.

# 1.Introduction

## Why Data analytics important

Data analysis refers to the analysis of large amounts of data collected with appropriate statistical analysis methods, summarizing, understanding, and digesting them, in order to maximize the development of data functions and play the role of data. Data analysis is the process of studying and summarizing data in detail in order to extract useful information and form conclusions.

The purpose of data analysis is to collect and refine the information hidden in a large number of seemingly messy data, so as to find out the internal laws of the research object. In practical applications, data analysis can help people make judgments in order to take appropriate actions. Data analysis is an organized and purposeful process of collecting and analyzing data to make it information. This process is the supporting process of the quality management system.

Throughout the life cycle of the product, including the process from market research to after-sales service and final disposal, it is necessary to properly use the data analysis process to improve the effectiveness. For example, before starting a new design, the designer must pass extensive design surveys and analyze the obtained data to determine the design direction. Therefore, data analysis has an extremely important position in industrial design.

## What are advantages and disadvantages of distributed system?

### Fault Tolerance and Redundancy

Distributed systems are also inherently more fault tolerant than single machines. A business running a cluster of 8 machines across two data centers means its apps would work even if one data center goes offline. This translates into more reliability as in case of a single machine everything goes down with it. Distributed systems stay put even if one or more nodes/sites stop working (performance demand on the remaining nodes would go up).

### Low Latency

Since users can have a node in multiple geographical locations, distributed systems allow the traffic to hit a node that’s closest, resulting in low latency and better performance.

### Cost Effectiveness

Distributed systems are much more cost effective compared to very large centralized systems. Their initial cost is higher than standalone systems, but only up to a certain point after which they are more about economies of scale. A distributed system made up of many mini computers can be more cost effective than a mainframe machine.

# 2.Dataset Analytics

## 2.1. Get the data:

The dataset to be used in this assessment is a collection of tweets available through the free Twitter API which covers a 1% random sample of the entire Twitter stream over 6 months (07/2014-12/2014).First of all,to figure out the structure of twitter data.I run example codes and then run script below:



It can be seen that the various attributes of the Twitter data set. For example, one of the attributes “user” is a structured data, it is divided into user.id, user.url, user.location and so on. My data analysis topic is about Justin Bieber's Twitter trends.

I tend to focus on certain attributes:

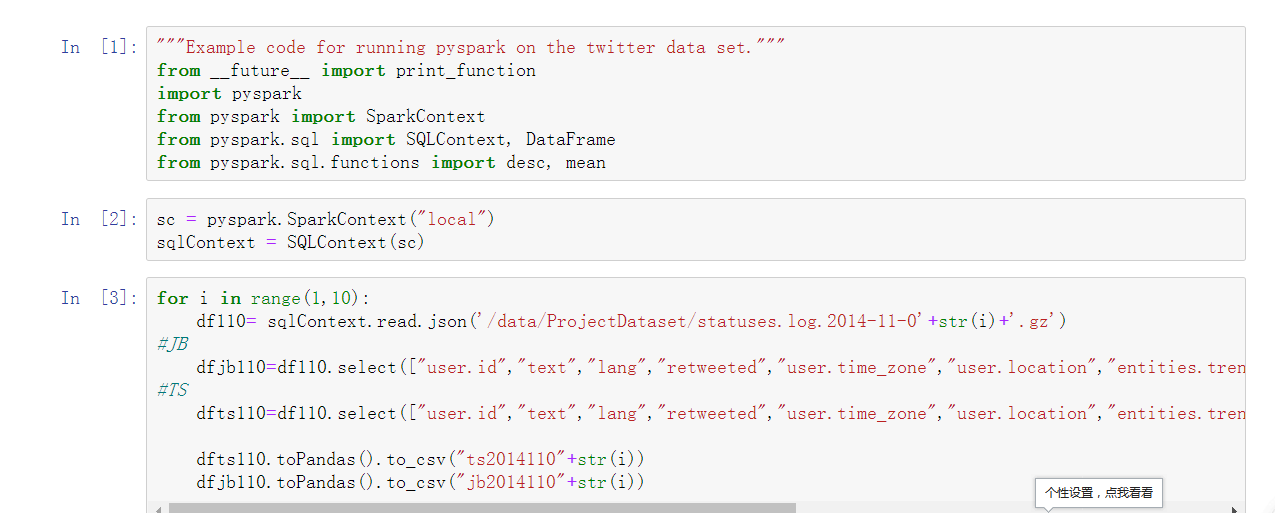
text(the text of twitter)

User.id

User.location

Lang(language of every twitter)

So I need to filter out the dataset include hashtag “JustinBieber”.But when I try this step,I found the volume of dataset is too big,and it can be easily prompted when running long processsing.Therefore I use a loop function to export and download the data set in a time series to a local folder.



## 2.2. Data pre-processing

Before analytics work.The datasets I download have a few problems.For example,the missing data,and also because of the encoding problem, many texts cannot be recognized, resulting in the problem of garbled characters.So the pre-processing includes 3 steps:

1.remove missing data volume

2.keep the features we will use in analytics.For example,I found the ’trend’ feature is useless after downloaded.

3.string data type converted to date data type.

Firstly, I use pandas.tocsv() and readcsv() to collect the data downloaded and write it into a new file named jb.count.Then add time series in the end column.For example,change ”20141101” into “2014-11-01”.Some codes are shown below:



It generates a csv file named ”jbcount”,which contains information about every users’ id,language they use and locations and url and date of every twitter.

## 2.3. Data analytics

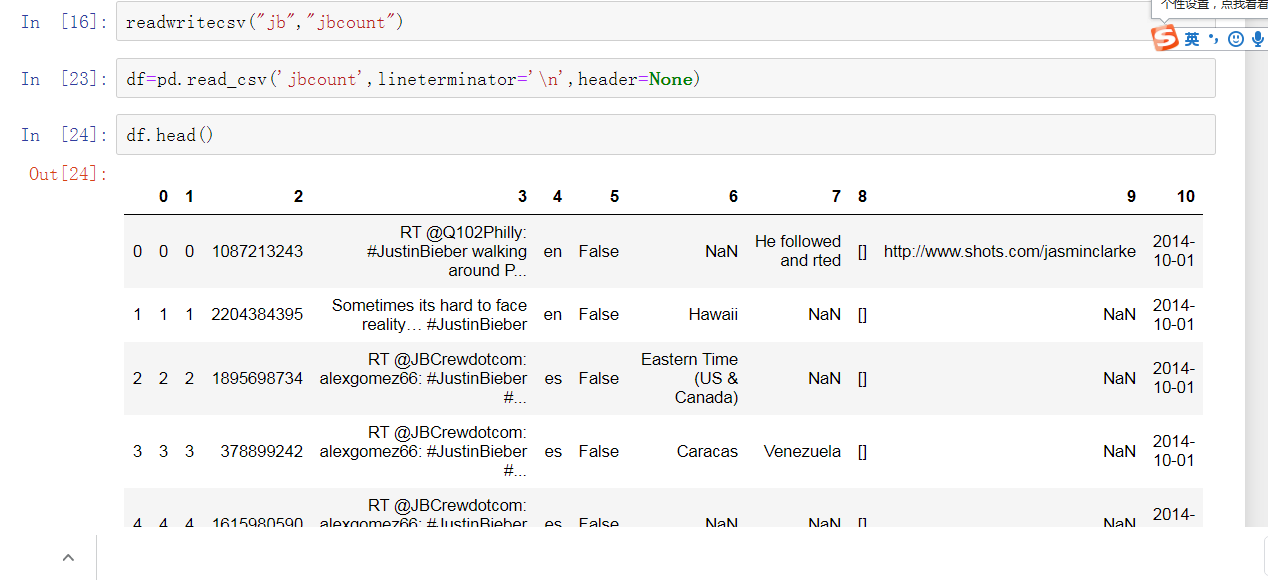
### Tools

In this section I will use python.pandas and python.nltk and tableau to do Justin Bieber’s twitter trend analytics.

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the [Python](https://www.python.org/) programming language.And NLTK is a leading platform for building Python programs to work with human language data. I use tableau to visualize data to make it confess.In this section I will use another dataset named stopword from nltk which contains plenty of English stopwords to do word frequency statistics.

### 2.Data analytics steps

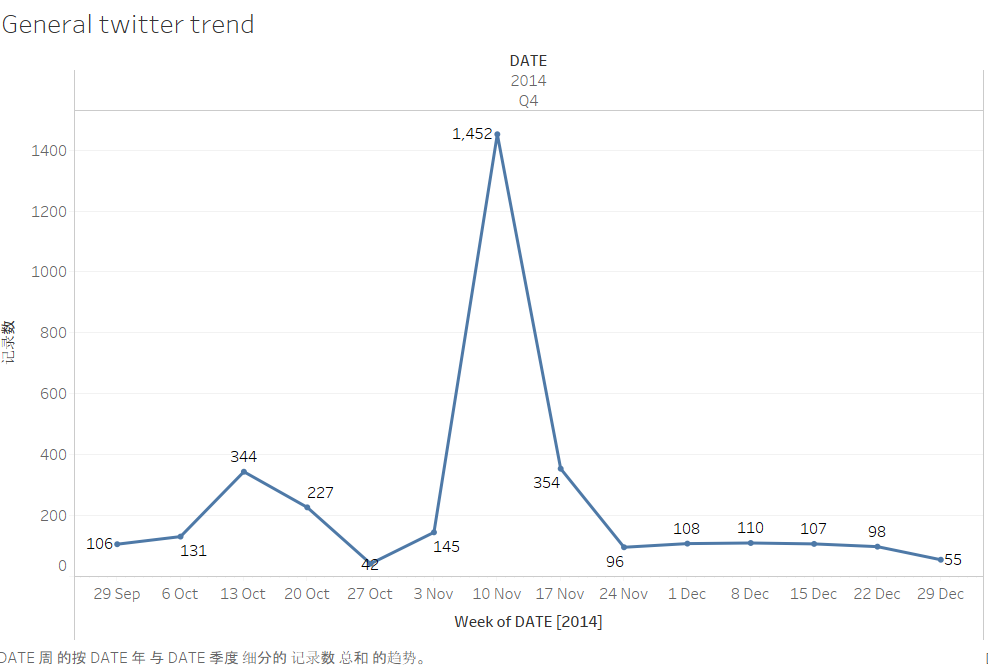
1. After collect the data”jb.count.csv”.I put them into a dataframe called df.run”df.head()”command,and then it shows the structre of the dataset.The attributes contains id,text,language user use,timezone,location and url and date.



Run df.count(),it shows that there are 3375 records about Justin in this period



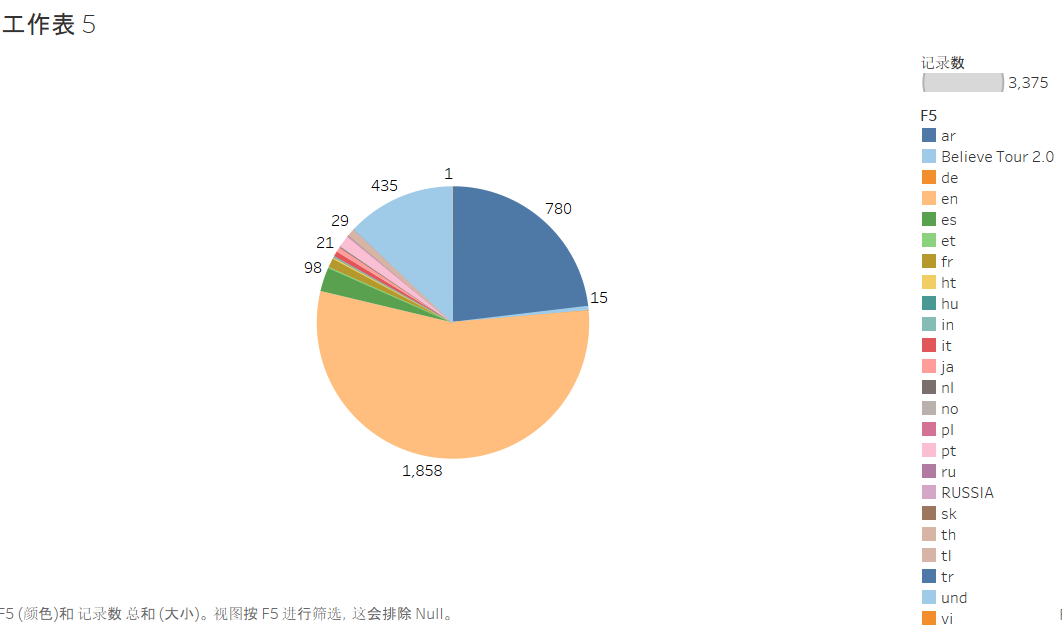
1. To visualize the result and present it better.Import jb.count.csv we generated to tableau.Then I select date as x axis and number of records as y.Import jbdata.csv to tableau and select date as x axis and number of records as y.The unit of date is per week.



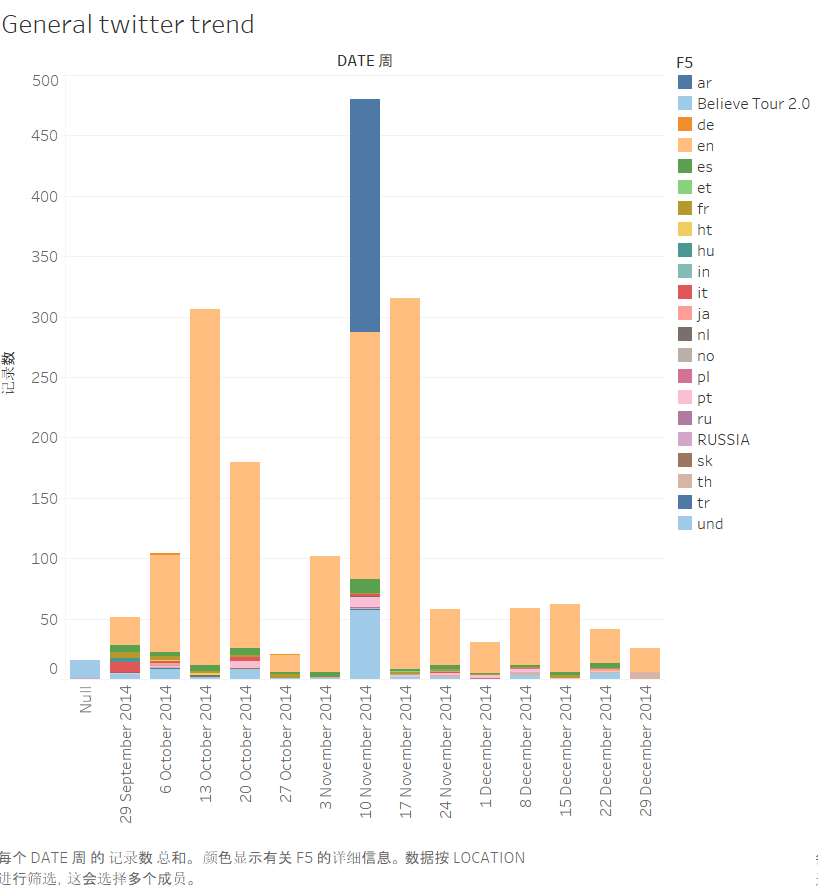
example 1:general twitter trend

The chart shows that Justin Bieber has got highest attention from 10 Nov to 17 Nov.Apart from that,his twitter trend stays stable during this month.

Select language that is not null, add it into column attributes, select color mark “lang”, and size mark "number of records"



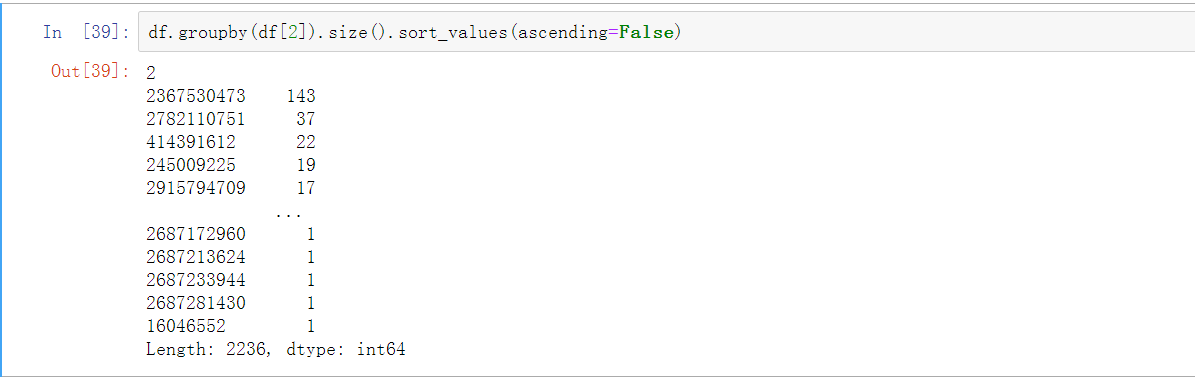
example 2:language user use



example 3:general twitter trend identified by language

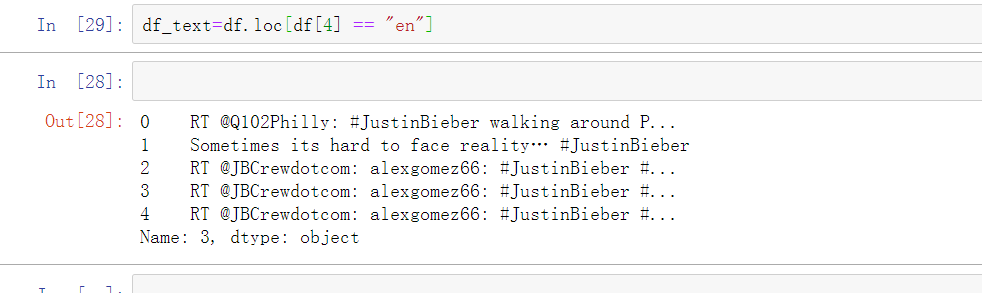
Generally,Most users(1858) in this case use English when mentioning him.Apart from English speaking users,Arabic and UND ranked second and third

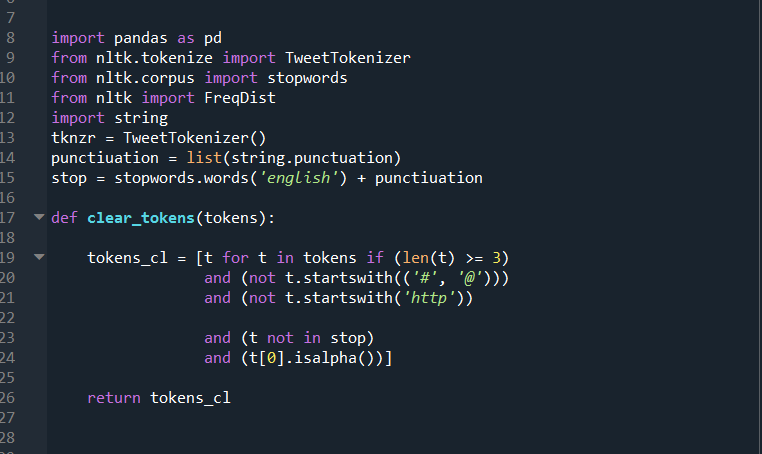
1. Given that the number of user’s id is too large that we can not present it in tableau.I use pandas.sortby() and pandas.groupby() functions to do descending sort.



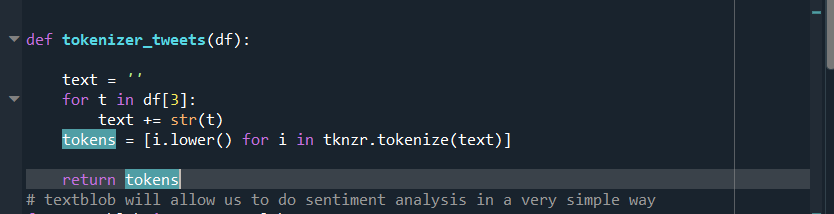
From the result,user”2367530473” tweeted about Justin most(143 tweets).

1. To find out what people usually tweets about him in this period,I choose to do an Word frequency statistics.Firstly,preprocess the text of tweets, which is text (utf-8 encoding).
2. The beginning is https, and the token is recognized as a website link
3. The beginning is RT, then the token is the fixed beginning of the forwarded tweet (retweet)
4. Start with @, it is the username of Twitter
5. The beginning is #, which means hashtag (the tag used to mark the thread topic in twitter)
6. Start with &, it is basically the & symbol in tweet, and &amp; in the text of utf-8

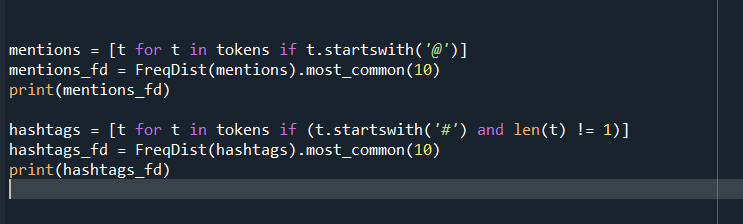


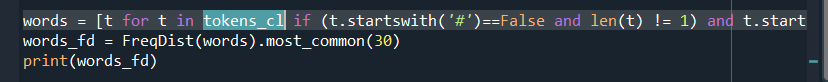
1. After that,we use the TweetTokenizer tokenizer in the NLTK module to clean up the text

Then remove special marks and stopwords

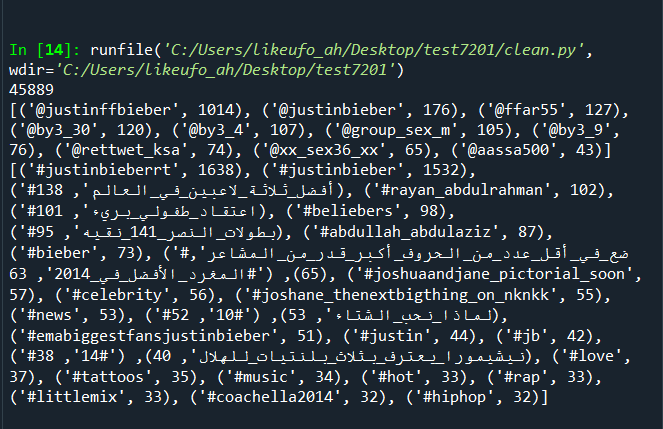


Then we counted the most frequently used tags and mostly mentioned users,and also,most frequent words user use.





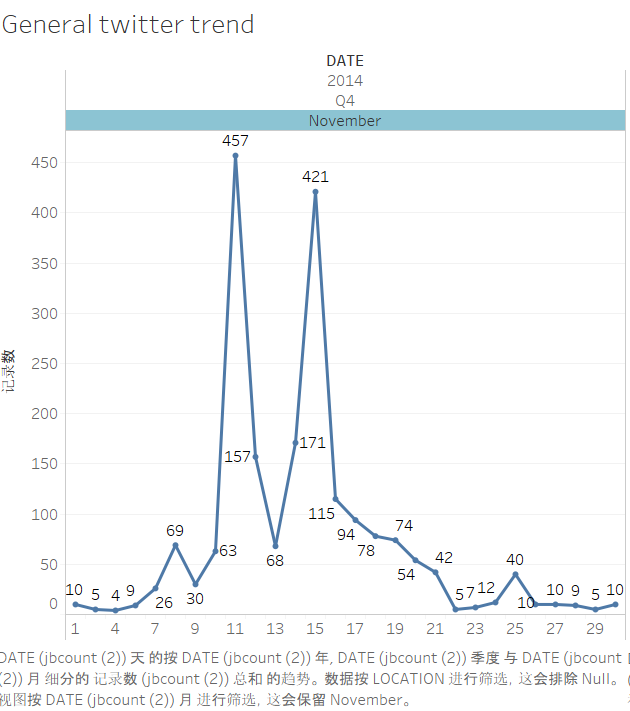
Results show like this:



# 3.Discussion and conclusions of the analysis

## What is Justin Bieber’s general twitter trend in this period?

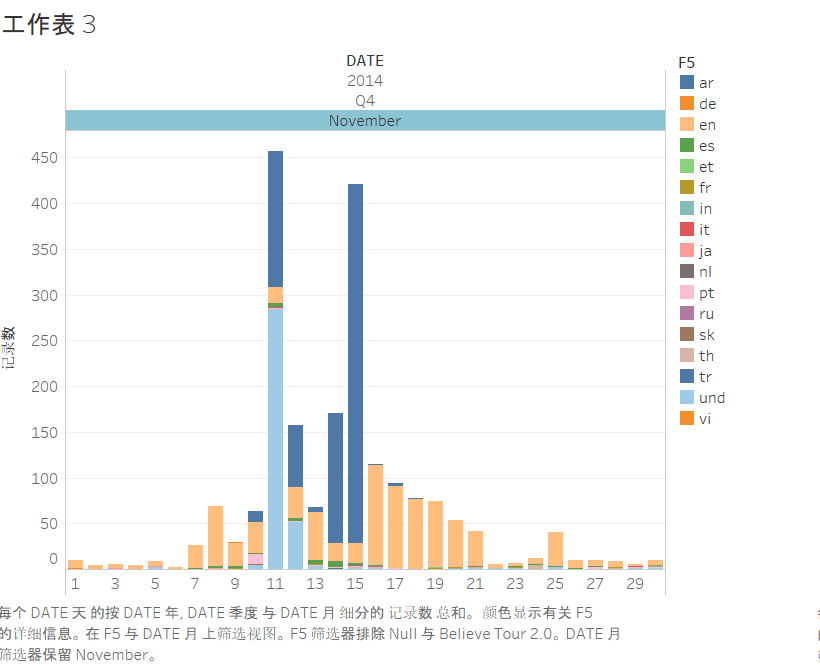
From above,example one shows Justin Bieber's general twitter trend.Generally,JB’s twitter trend stays stable in Oct and Dec.Twitter trend peaked on 11 Nov (457 tweets)and 15 Nov(421 tweets). This may attributes to the new song of Justin Bieber and Cody Simpson”Home to mama” released on 11 Nov.



## The information about users who tweet about him

Among all users tweeted about him,English Twitter accounts for biggest propotion.(1858 records).

But example 3 shows that from 11 Nov to 17 Nov,Justin Bieber got the highest attention in Arabic-speaking areas.To be more specific,in 15 Nov,most tweets with hashtag “JustinBieber” are posted in “ar”(393 records).



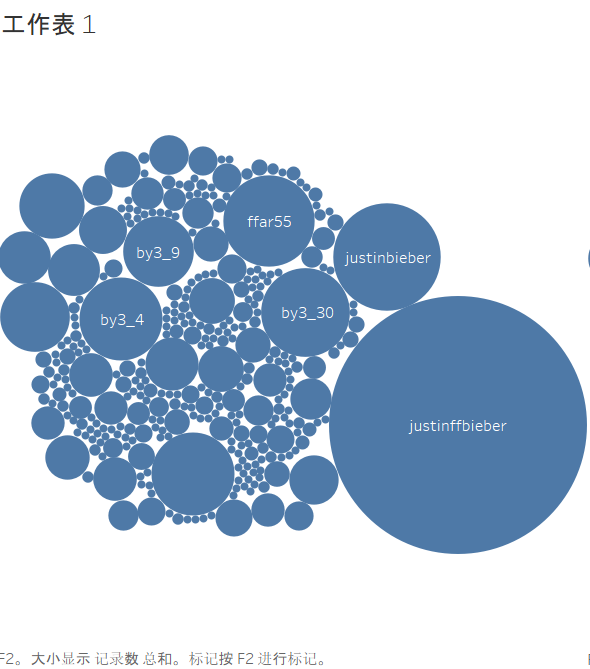
example 4:number of people tweet about JB in Nov.

Also,From step 3,it shows users’ activity in this hashtag. User”2367530473” tweeted about Justin most(143 tweets) in three months.which means this user is very concerned about the activities of Justin Bieber.

## What do people focus on Bieber?

In step 4,We list the most frequent hashtag and user metioned and words.

As we can see the most mentioned users are shown below.Apart from Justin Bieber himself,there are some other twitter accounts like justinffbieber,ffar55 and so on.

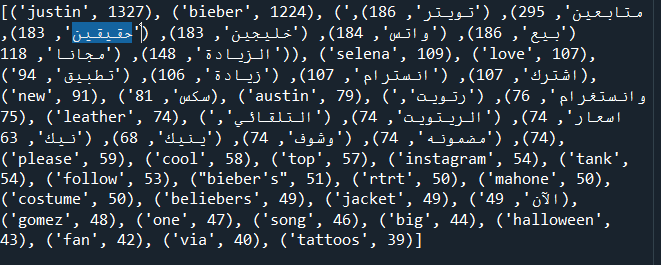


From example6, we print top 30 most frequent hashtag mentioned in all tweets.In spite of those hashtags about JB himself,there are also some hashtags including other celebritieslike”littlemix”,or hashtags like ”hiphop” and “music”.



example 6:most frequent hashtag mentioned

Also,we print top 30 most frequent words appear in these tweets.It is interesting to find that “Selena Gomez” has been mentioned so many times, other key word mentioned most including”Austin Mahone” and “love” and “cool” and “instagram” and so on.



Overall,Justin Bieber's Twitter popularity from October 2014 to December 2014 was not very high, 3347(records in total). According to the above statistics, among all the tweets discussing him, in addition to English, the number of people who use Arabic is the largest.

On the other hand,Among all the tweets about him, there are not many records about his song. This may be because Justin Bieber did not engage in musical activities in the second half of 2014. Compared to his songs, people mainly focused on his life during this time.

Another interesting finding is that many people who follow Justin Bieber also mentioned Selena Gomez.This means that a lot of people who follow Bieber are concerned about Selena Gomez.

This may be instructive for Twitter's accurate recommendation.For example,If one follows Justin Bieber, Twitter can recommend Selena Gomez and Austin Mahon to the user,or can also recommend some entertainment news from Bieber and Selena Gomez, or their endorsement ads.