Data description

We used a publicly available dataset from followthehashtag.com repository for our analysis. It is a database of 170,000 Apple tweets collected over a 24 hour period from 4/28/2016 9:00 a.m to 4/29/2016 9:00 a.m. Retweets were excluded from this search, so the dataset has only original tweets.

Tools used

Spark (python+scala), Tableau

Data Preprocessing

Only the content part in the dataset was used for ngram and sentiment. Firsts, each word was normalized and removed punctuations. Then stop words were also excluded from the dataset. Finally the cleaned texts were grouped to count ngram or mapped with positive and negative words to count sentiment.

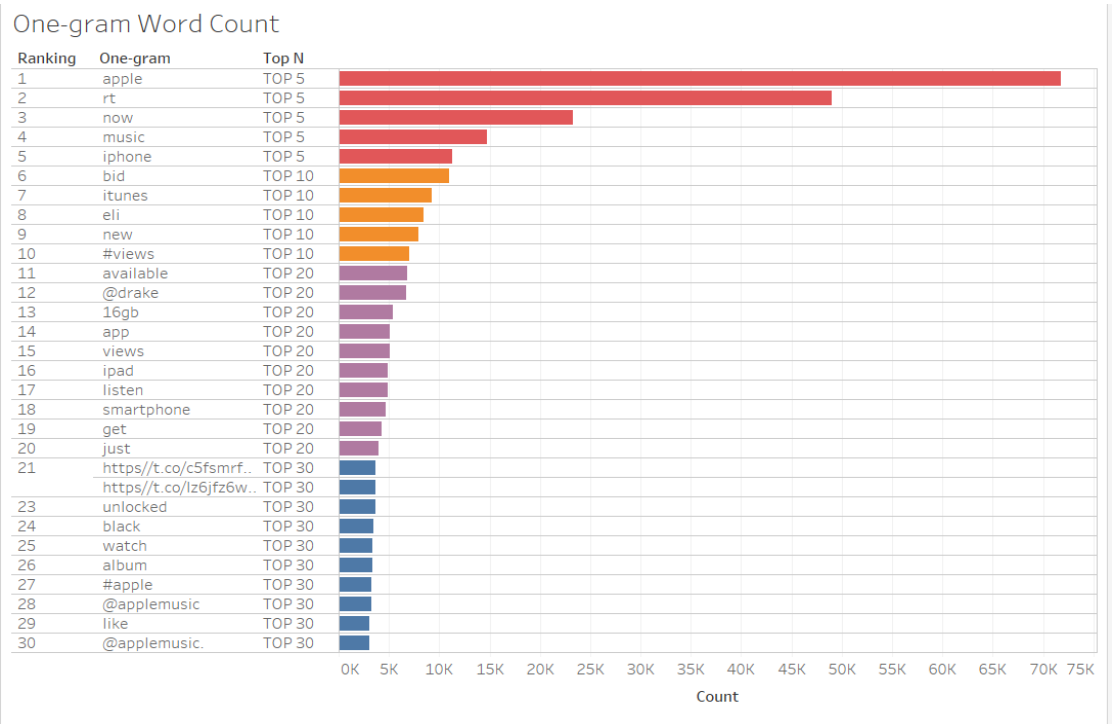
Findings

- Apple iTunes was a highly popular platform for musicians to promote new releases
- New buzzwords could go viral swiftly on twitter
- New market trends in general were tweeted with a more positive sentiment
- New filter criteria should be set for irrelevant terms (ex: Eli Apple)

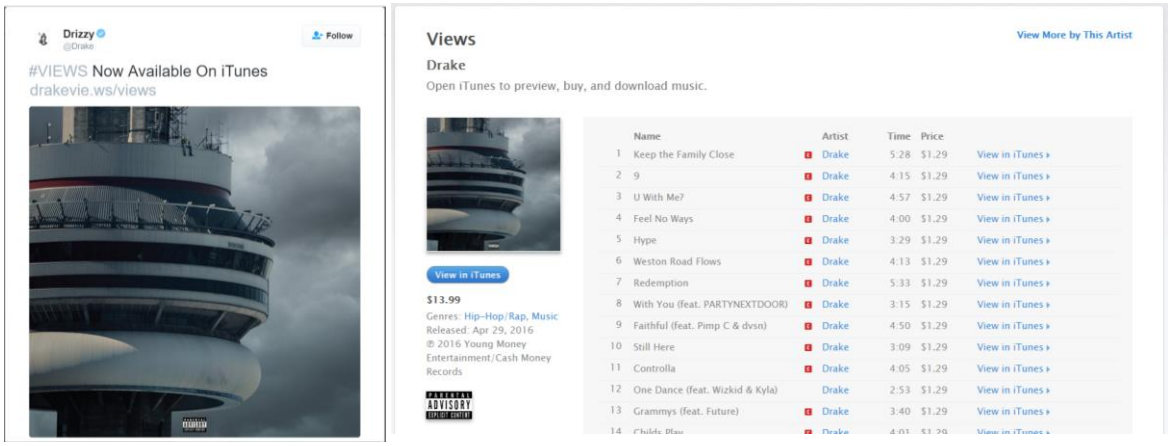
Methodology

- **Ngram for market trend**

Simply using from n-gram analysis could reflect what the tweets about apple were talking about. One-gram could be generated by simply using reduceByKey function on paired RDD. The top 30 results on one-gram are shown below.



As the chart indicated, music and iphone were commonly mentioned in the tweets about Apple. In addition, the data was collected on April 29, 2016, which happened to be the release date of singer Drake's new album Views. Therefore, relevant words, such as "#views", "available", "@drake", "views", and "album" were very popular on that day. The two links on top 30 direct to the twitter and itune webpages of Drake's album too.

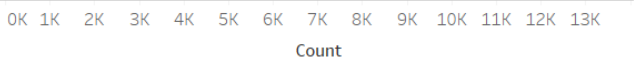


Screen shot of link https//t.co/c5fsmrf8zw and https//t.co/lz6jfz6wlc

Bigrams were also applied to explore more information. To explore bigrams on pyspark, the package nltk was combined with RDD and other python functions. Bigrams were extracted from each tweet after stop words and punctuations were removed. Still, top 30 results from bigrams are illustrated in the following chart.

Bigrams Counts

Ranking	Bigrams	Top N
1	apple music	TOP 5
2	bid now	TOP 5
3	apple iphone	TOP 5
4	eli apple	TOP 5
5	@drake #views	TOP 5
6	available itunes	TOP 10
7	now available	TOP 10
8	apple ipad	TOP 10
9	#views now	TOP 10
10	rt @drake	TOP 10
11	https//t.co/lz6jfz6wlc https//t.co/c5fsmrf8zw	TOP 20
	itunes https//t.co/lz6jfz6wlc	TOP 20
13	apple watch	TOP 20
14	space gray	TOP 20
15	apple ipod	TOP 20
16	carl icahn	TOP 20
17	apple music.	TOP 20
18	https//t.co/c5fsmrf8zw rt	TOP 20
19	iphone 6	TOP 20
20	now itunes	TOP 20
21	rt @applemusic	TOP 30
22	ipod touch	TOP 30
23	big apple	TOP 30
24	bid now\\ufffd	TOP 30
25	iphone 5s	TOP 30
26	smartphone bid	TOP 30
27	unlocked smartphone	TOP 30
28	ipad mini	TOP 30
29	apple macbook	TOP 30
30	latest model	TOP 30



The market trend of the day becomes more obvious after viewing top bigrams. Apple's products, including apple music, apple iphone, apple ipad, and apple watch were commonly tweeted. Moreover, Drake's new album Views is now the top 5 most popular words, and bigrams in top 6 and top 10 are almost relevant to the album release as well.

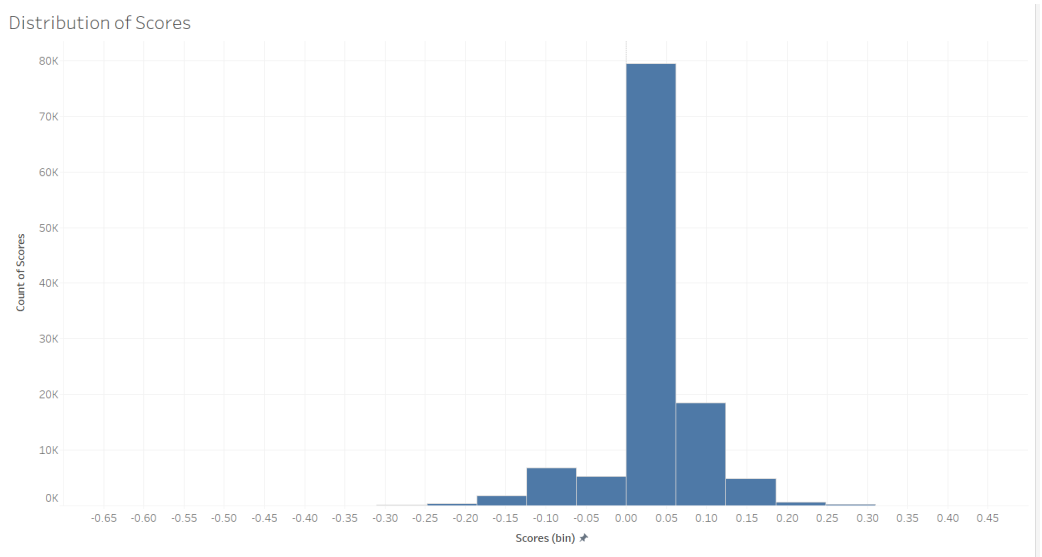
Interestingly, an American football player Eli Apple is the top 4 popular word, which is around 5% of the data. When generating streaming from twitter, tweets about Eli Apple were also accidentally included. Unfortunately, it was too late for us to uncover this fact, so the data consists tweets about Eli Apple. For similar studies in the future, this step could be done firstly to exclude unwanted data.

Based on the result from bigrams, we discovered some keywords related to tweets related to Apple on that day. Therefore, in the following sentiment analysis, we would also look by tweets with different keywords: Apple music, Apple iTunes, Apple iPhone, Apple iPad, Apple Watch, and Drake.

- **Sentiment Analysis**

The sentiment analysis adopted common approaches. For each tweet, the number of positive and negative words were calculated. The differences between the number of positive and negative of the word would then be divided by total words of that tweet to get the sentiment score. A score higher than 0 means positive sentiment, and a score below 0 means negative sentiment. Zero could be would be considered as neutral or no sentiment.

The distribution of shows that most of the tweets have positive tones. The highest score is 0.5 whereas the lowest score is -0.67. On average, the sentiment score is 0.0163.



However, tweets with different keywords had distinct level of sentiment. Users seem to had more positive tones when tweeting about iTunes and Drake, whereas tweets about music, iPhone or iPad were indifferent in terms of sentiment.