TWITTER DATA COLLECTION REPORT

Comm337 Project 2

Abstract

A project that uses 3 days' worth of tweets (14.4K tweets or more) for data visualization, sentiment analysis, time trend analysis as well as the insights of the Twitter data.

Prepared by:

Haihua Zhu	20806155
Mo Chen	52909158
Junyuan Wang	36063162

Haihuazhu1004@gmail.com
Mo.angela.chen@gmail.com
noic.junyuanwang@hotmail.com

Procedures:

For this project, we collected 3 days' worth of tweets (100 tweets/hour x 24 hours x 3 days x 2 keywords = 14.4K using the model we have built from the last project. The aim of this project is to apply the text mining methods to analyze the topics and popular words that related to the two brand Microsoft and Amazon. In the following report, we will provide detailed procedures and insightful information based on the sentiment analysis and the word cloud we have made.

Before stepping into the requirements, we designed a series of helper functions to set up. This includes functions like removing punctuation and digits, remove stop-words, remove meaningless words("https","co","rt","http",'amp','de','theyr'), loading data from relative path, Stemming and so on. We do this because it will make the upcoming steps a lot easier.

A. [Preliminary Analysis]

Find the ten most popular words with and without stop words

- We did two versions of analysis on the most popular words, one is with stop words, the other is without stop words.
- For the version with stop words, to get the frequent words, we simply used freq = nltk.FreqDist(words) to achieve our goal. For the version without stop words, we used list method to sort out punctuation and digits, and then removed stop words and meaningless words such as ("https","co","rt","http" "https","co","rt", "theyre",'de').
- Applied Stemming and Lemmatization methods to find families of derivationally related words with similar meanings. Then tokenize text file with nltk and find the top 10 most frequent words.

Find the ten most popular hashtags (#hashtag)

- To find the the ten most popular hashtags, we used t.startswith('#') to pick up all the hashtags and then applied the same method as before when we found ten most popular words.

Find the ten most frequently mentioned usernames

- First, we set cba=[t for t in word3 if t.startswith('@')] to find all the mentioned users name, the used freq = nltk.FreqDist(cba) to find the ten most frequently mentioned usernames and printed them out.

Find the most influential tweet

- We defined that the influence score is the sum of retweet count, reply count, favorite, and quote count.
- 4 lists were created, and in each list, we applied list.append() method to find all the 4 key components retweet count, reply count, favorite, and quote count. During this process we found that if one tweet was already a retweet, then the

retweet_count would be 0, and we were supposed to check the retweeted_status as well .

B. [Word Cloud]

- First, install wordcloud by typing
  !pip install wordcloud
- Sorted out all the meaningless words and stop words in English, asked python to read the whole text used text = all_tweets and generated a word cloud image by excuting wordcloud = WordCloud().generate(text). In this process we displayed words in a lower case and saved the image in both PDF and PNG files.

C. [Sentiment Analysis]

- Installed textblob
- !pip install textblob
- Analyzing the sentiment score of each tweets from the perspective of polarity and subjectivity.
- For polarity and subjectivity, created two bar charts

```
plt.hist(pol_list, bins=10) #, normed=1, alpha=0.75)
plt.xlabel('polarity score')
plt.ylabel('sentence count')
plt.grid(True)
plt.savefig('polarity.pdf')
plt.show()
```

INSIGHT:

Data-driven insights

For Microsoft(MS), the most popular and meaningful keywords(from word-cloud, keywords hashtags data) beside microsoft itself are:

1. Microsoft product related: Azure, Adobe, Office, xbox, xboxone, xboxonex

2. Function of MS Product: Cloud, service, game

3. General term related to MS: tech, business, buy, AI

4. MS People: CEO

5. Rivals: Amazon

6. Emotional words: Awful

7. Event related: Adobesummit

From time trend we see on the second day and third day, the “awful” word appeared on the word cloud and the polarity score dropped a little due to this event. But beyond that there is no much change over the 3-days period.

Generally, the subjectivity histogram seems behave the same, around 1300 tweets are around 0, all the other tweets are more or less subjective, which is normal compared to other tweet data. The polarity histogram is normal too, majority falls under the range 0.00-0.25, which means major tweets about MS are a little bit positive but not so much. Other finding is, beyond the majority, most tweets about MS is positive.

For Amazon, the most popular and meaningful keywords (from word-cloud, keywords hashtags data) beside Amazon itself are:

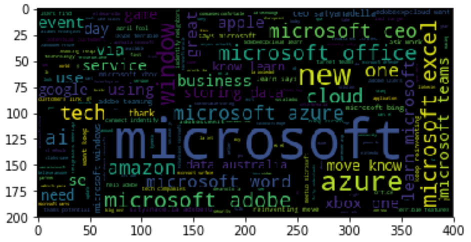
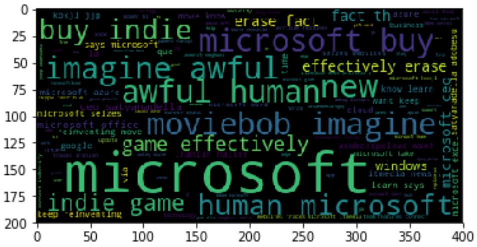
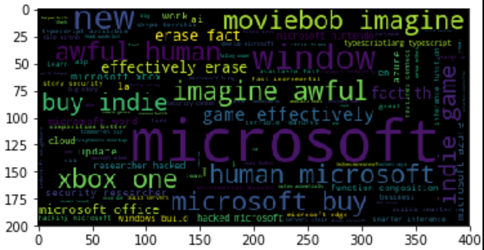
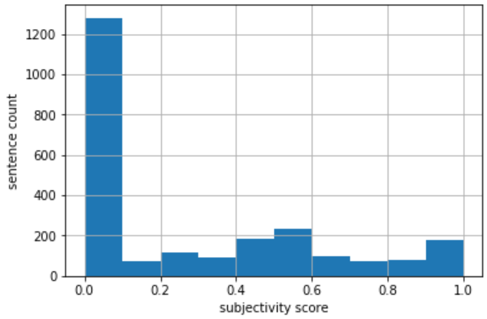
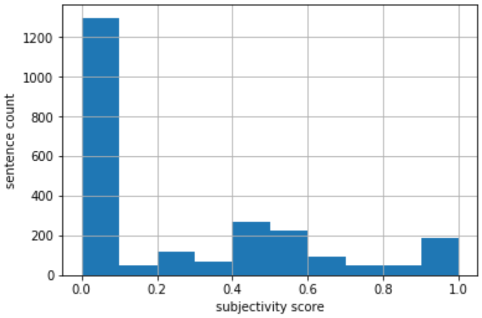
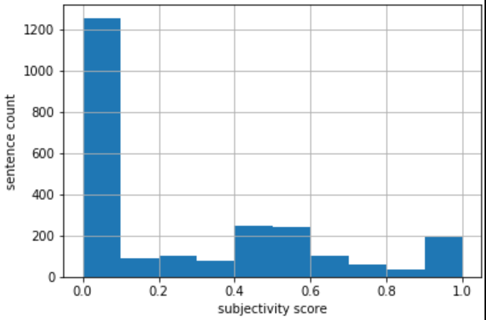
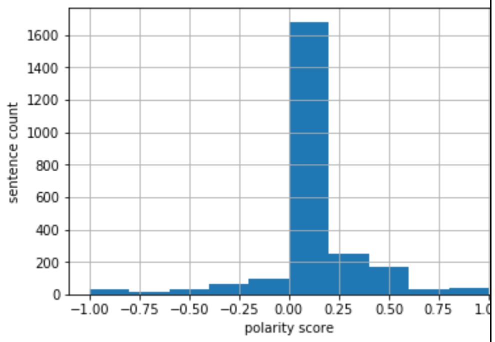
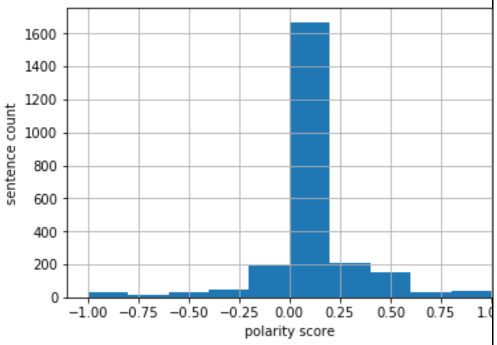
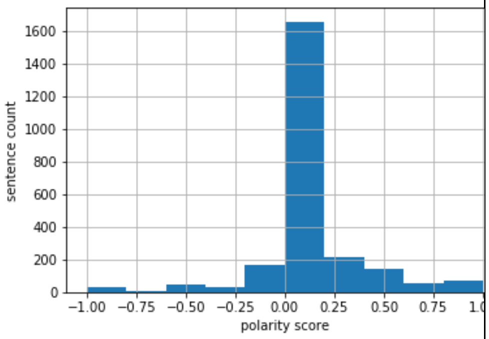
1. Amazon product related: drink, chips
2. Function of Amazon Product: Pay
3. General term related to Amazon: PSA (Public social announcement)
4. Amazon People: @divblita- the person who promote his product on amazon and twitter
5. Amazon partners: Yucho

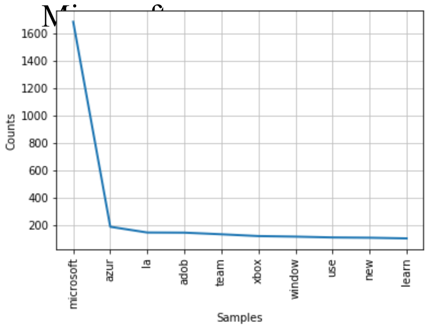
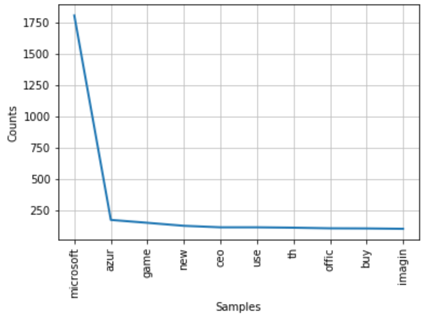
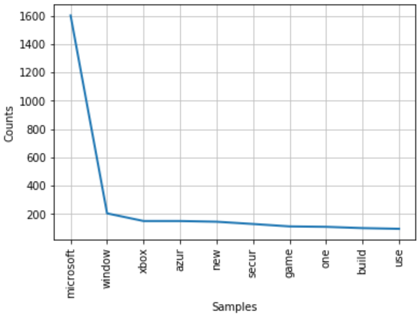
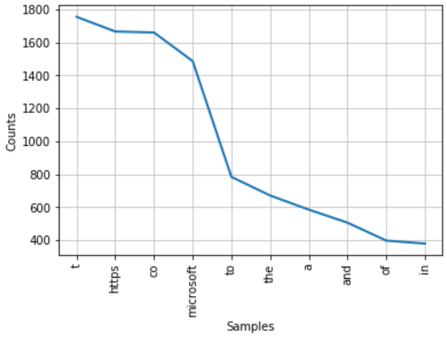
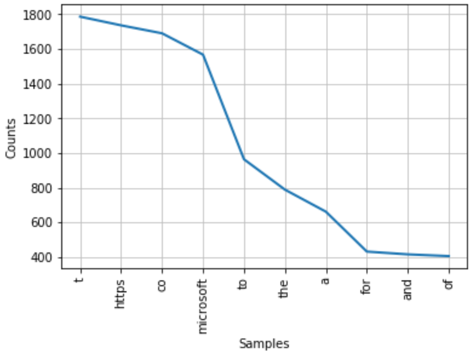
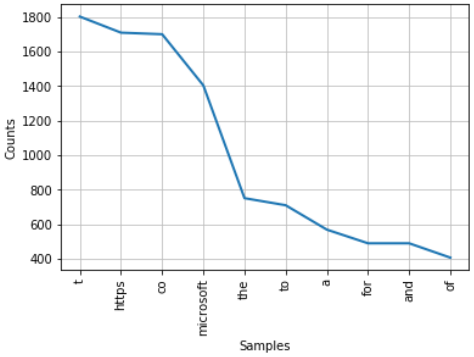
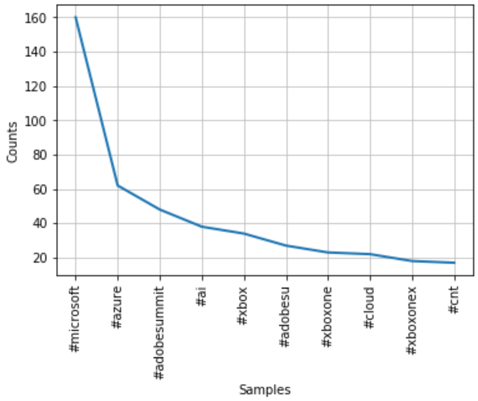
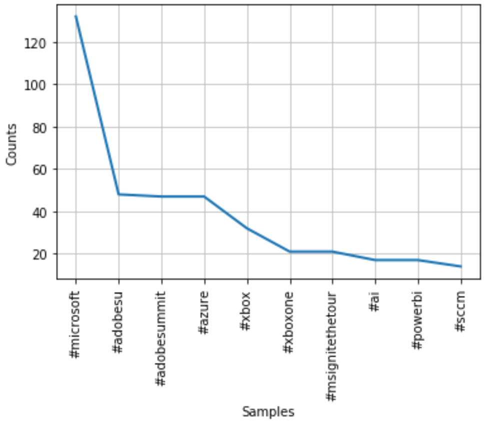
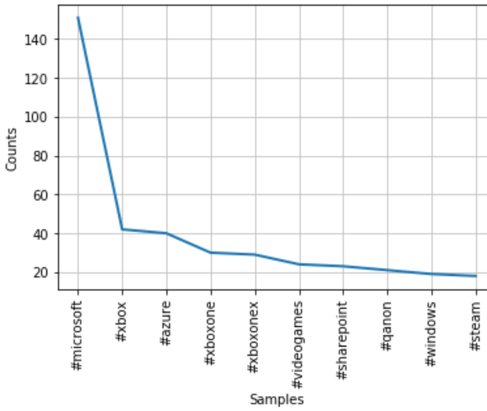
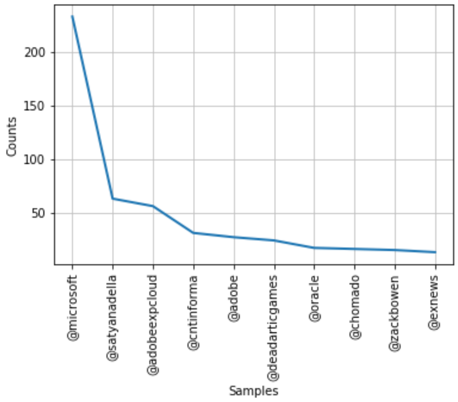
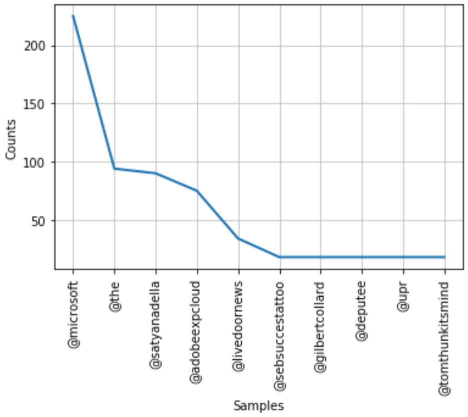
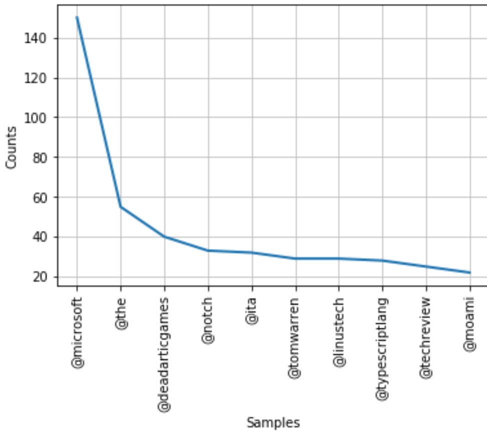
Unlike microsoft tweeting info, amazon’s data involved a lot of advertisement and promotion instead of brand related data. We think this makes sense as lots of small vendors online depends on amazon to sell their goods, versus microsoft does not have this distribution of merchandise like amazon. That is why there are some tweets that has a strong subjectivity compared to microsoft tweets data. Generally, the subjectivity histogram seems behave the same, around 1500 tweets are around 0, all the other tweets are more or less subjective, which is normal compared to other tweet data. The polarity histogram is normal too, majority falls under the range 0.00-0.25, which means major tweets about amazon are a little bit positive but not so much. Other finding is, beyond the majority, most tweets about amazon is positive. We think this is because there are more ads on twitter that mentioned amazon, they are more likely presenting themselves as positive. Even though we cannot conclude that amazon is well received in public eyes, we could say the market power is strong as lots of vendors rely heavily on amazon.

Project insights:

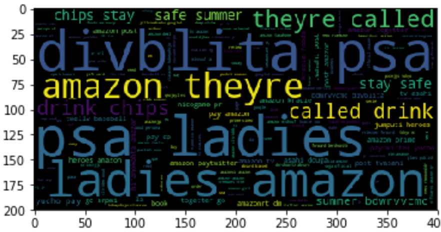
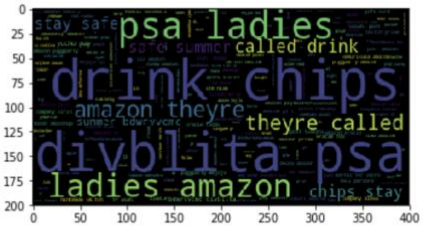
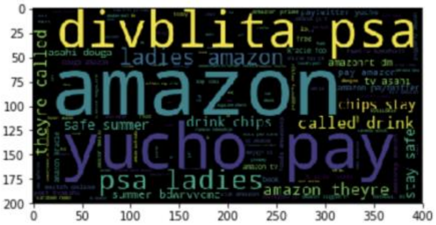
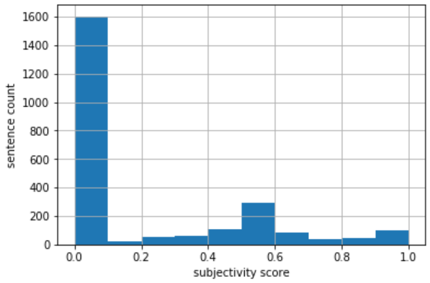
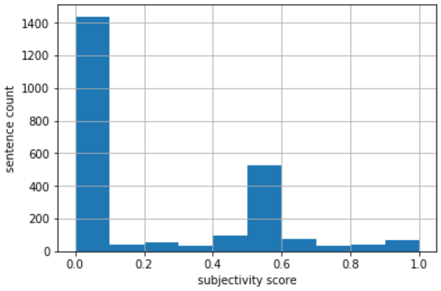
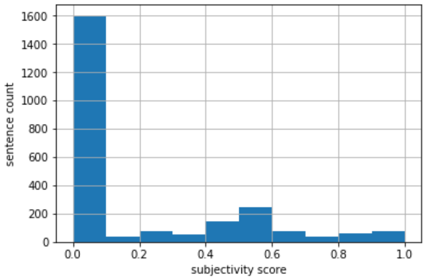
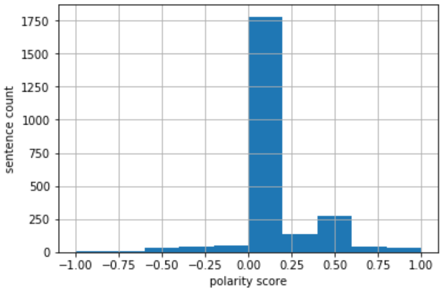
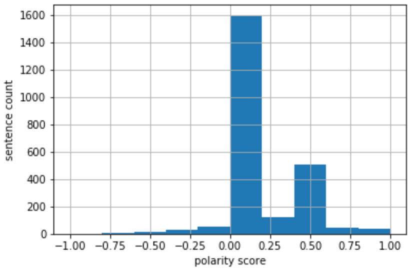
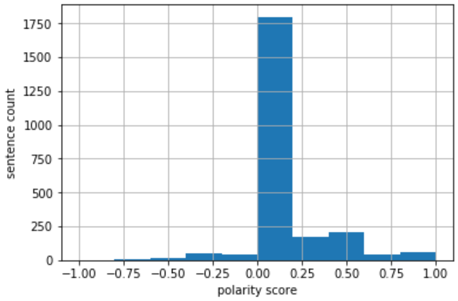
We learned a bit of the taste of managing large code project, with systematic approaches. We wrote functions as a set up and as a helper for the main part of the code to fulfill the requirements of this project. Building the helper function is so useful when it is a complex project that we can use them to break down steps and procedures. We feel building those helper functions really let us have an understanding of the purpose of functions. We feel confident about coding after this project, and we feel proud and surprised that, with resources, we can build something meaningful and insightful. Last but not least, we want to continue to explore data as a career and an interest.

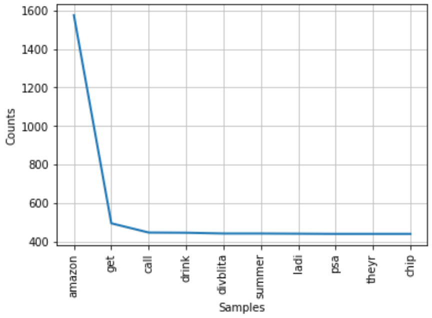
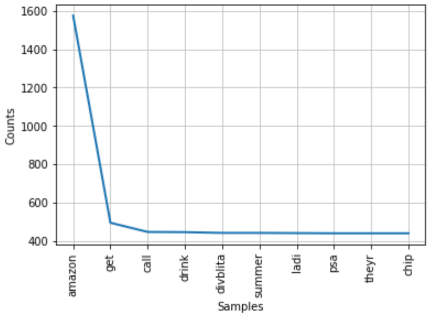
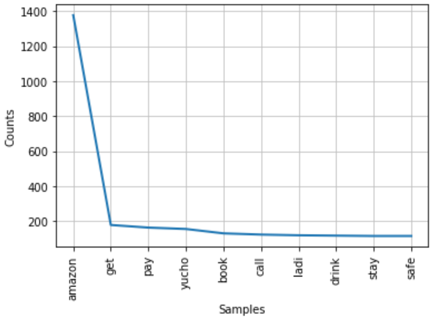
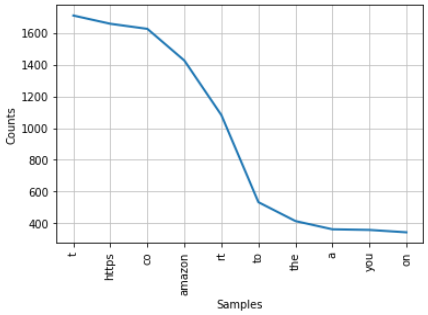
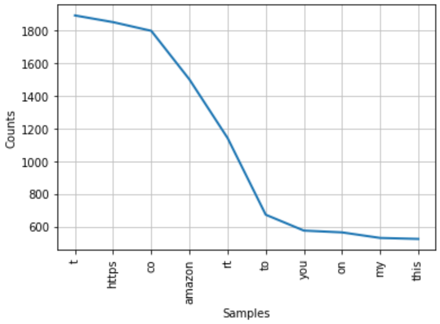
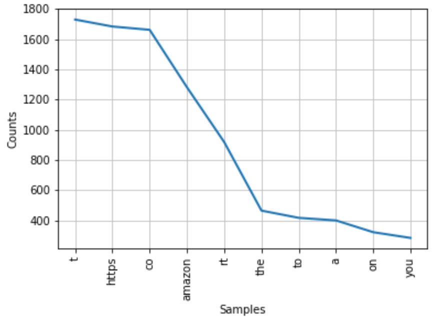
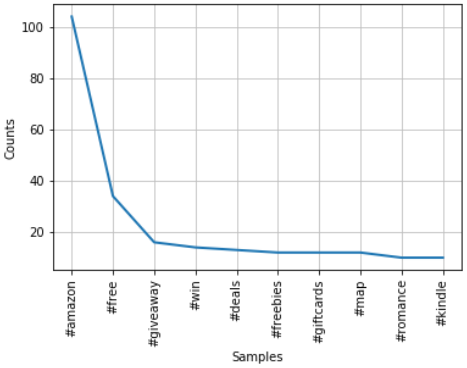
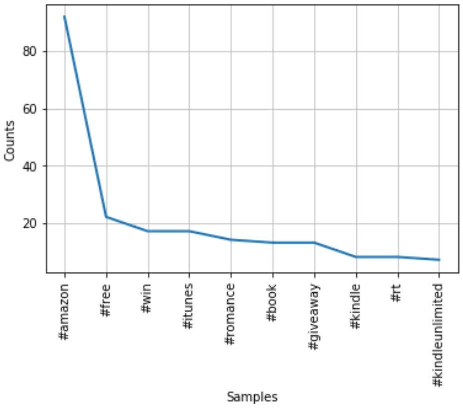
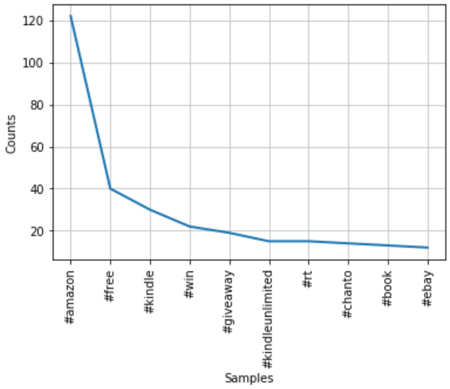
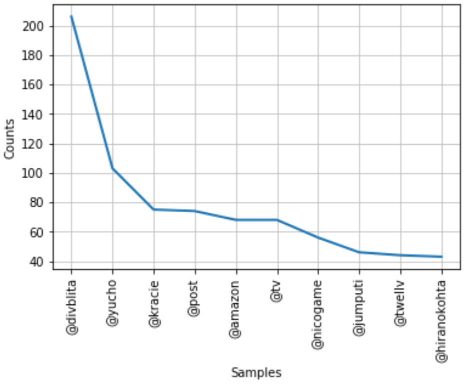
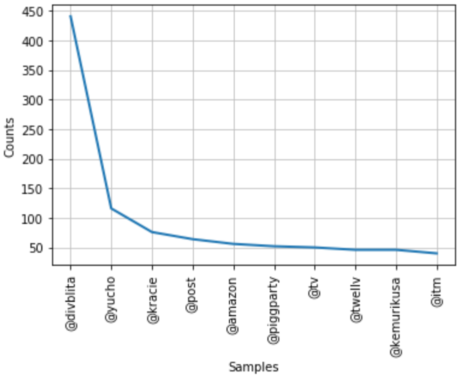
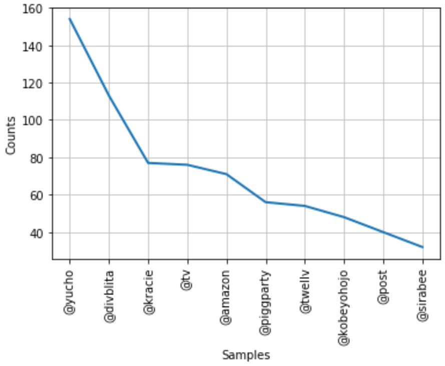
Microsoft

Day 1	Day 2	Day 3
The most frequently tweeting person is LabVIEWopenJS	The most frequently tweeting person is LabVIEWopenJS	The most frequently tweeting person is Rudi A.R.
the most influential tweet is this: RT @link0230: 新型ウイルスに感染しました。 皆さんも注意してください。 （一応 Microsoft に報告しました。） https://t.co/S4bYXJgCvH	**the most influential tweet is this:** RT @i_winnn: แป้นพิมพ์ลัด สำหรับ Microsoft Word Ctrl+U = ขีดเส้นใต้ข้อความ Ctrl+F = เปิดกล่องค้นหาในบานหน้าต่าง Ctrl+O = เปิด Ctrl+S = บั…	**the most influential tweet is this:** RT @ColIegeStudent: using microsoft word *moves an image 1 mm to the left* all text and images shift. 4 new pages appear. in the distance…
		
		
 The average subjectivity is: 0.2621723377606191 The average polarity is: 0.07589889426843767	 The average subjectivity is: 0.2561957499098123 The average polarity is: 0.057487888746091995	 The average subjectivity is: 0.2603240073776537 The average polarity is: 0.07960540940059162

Day 1	Day 2	Day 3
 Top 10 most frequent words witho ut stop words		
 Top 10 most frequent words with stop words		
 Top 10 most popular hashtag		
		

Amazon

Day 1	Day 2	Day 3
The most frequently tweeting person is 日テレ公式@宣伝部	The most frequently tweeting person is 日テレ公式@宣伝部	The most frequently tweeting person is 日テレ公式@宣伝部
the most influential tweet is this: RT @bts_bighit: #BTS MAP OF THE SOUL : PERSONA Pre-order Notice ▶Big Hit Shop(US ONLY): https://t.co/eThX79aiG7 ▶Amazon: https://t.co/Rwd…	**the most influential tweet is this:** RT @divblita: PSA to all my ladies! You can get these on amazon. They're called drink chips. Stay safe this summer 😝 https://t.co/bDwrvvCmC3	**the most influential tweet is this:** RT @MrBeastYT: I'm going to dm someone who retweets this tweet a $1,000 Amazon Gift card in 72 hours. If you're picked but don't follow me,…
 divblita psa amazon theyre psa ladies ladies amazon drink chips called drink theyre called chips stay safe summer stay safe 0 25 50 75 100 125 150 175 200 0 50 100 150 200 250 300 350 400	 psa ladies drink chips divblita psa ladies amazon amazon theyre theyre called called drink stay safe chips stay 0 25 50 75 100 125 150 175 200 0 50 100 150 200 250 300 350 400	 divblita psa amazon yucho pay psa ladies ladies amazon drink chips called drink safe summer amazon theyre 0 25 50 75 100 125 150 175 200 0 50 100 150 200 250 300 350 400
		
 **The average subjectivity is: 0.1865579644506727 The average polarity is: 0.09100097305038715**	 **The average subjectivity is: 0.21174672243266002 The average polarity is: 0.1450761642810795**	 **The average subjectivity is: 0.18035051988135312 The average polarity is: 0.09582735171536601**

Day 1	Day 2	Day 3
 Top 10 most frequent words without stop words		
 Top 10 most frequent words with stop words		
 Top 10 most popular hashtag		
 Top 10 most popluar metion@	 450 400 350 300 250 200 150 100 50 Counts @divblita @yucho @kracie @post @amazon @piggparty @tv @twellv @kemurikusa @itm Samples	 160 140 120 100 80 60 40 Counts @yucho @divblita @kracie @tv @amazon @piggparty @twellv @kobeyohojo @post @sirabee Samples