

Group 4 Project 2 Presentation

Trevor Buck, Liam Hood, Jordan Juel, Alan Liang

Sample Stream Data

- Sample stream was collected and used for most of the analysis
 - 17 hours and 40 minutes total
- 40 minutes at UTC 23:00 (6 pm Eastern) November 13th
- 4 hours starting at UTC 21:00 (3 pm Eastern) November 14th
- 13 hours starting at UTC 6:30 (1:30 am Eastern) November 15th
- Stopped when I used my computer

Which hashtags have the longest/shortest average tweet length?

Grab data for tweet text and hashtags

Count the words for each tweet

Split hashtag arrays into individual entries

Count and filter out hashtags with less than 100 occurrences

Average the word counts for each hashtag

hashtag	avgLength
WeLoveYouJae	34.29
WeAreHereForYouJae	34.22
stop	24.94
DigitalSoldiers	20.85
coronavirus	19.88
UlyssesPH	19.79
COVID19	19.16
Kalbajar	19.01
shopmycloset	18.82
BLM	18.47
PS5	18.47
Dominion	18.24
Georgia	18.13
JB	18.03
EndSARS	17.72
NFL	17.62
Antifa	17.27
SoumitraChatterjee	17.16
Trump	17.16
Smartmatic	16.94

Which hashtags are largely associated with positive/negative tweets?

Get data for tweets like previous query

Slip the tweet text into words and remove any punctuation

Used words from a research paper to check for positive or negative words

Generate score based on how many positive and/or negative are present

Split hashtag arrays into individual entries

Average the scores for each hashtag

hashtag	rating
WeAreHeForYouJae	14.0
WeAreHereForYouJae	11.13
BaileyMay	11.0
Baeleys	11.0
WeLoveYouJae	10.85
SHSU	10.0
MahalKitaBaileyMay	9.0
AMAsI	8.0
Aedu	8.0
WithYouJae	7.65
dfescorts	7.0
escortsdf	7.0
Sikh	7.0
mexicoescorts	7.0
houseoftides	7.0
CovidFree2022World	7.0
LoveYou	7.0
Mookkuthiamman	7.0
ocr	6.0
fortyhall	6.0

hashtag	rating
TyphoidTrump	-38.0
SoreLoserInChief	-18.0
Chinazi	-9.0
IStandWithGinaCarano	-8.0
MAGAt	-8.0
Fliplife	-7.0
TrumpCancelled	-6.0
Cleansing	-6.0
plasticBarbie	-6.0
IVotedForBlackpin...	-6.0
Darkness	-6.0
TrumpIsA_DANGER_T...	-6.0
TrumpGenocide	-6.0
Darkages	-6.0
TimeWasted	-6.0
STREAMHYLTMV	-6.0
NurownEurope	-6.0
Trumpflakes	-6.0
F██BlackA██holes	-6.0
Narcissists	-6.0

Which hashtags are most commonly used with other hashtags?

- Select “tweet_id” and the list of “hashtags” from dataset
- Made 2 separate Dataframes
 - First with id and exploded hashtags
 - Second with id and the number of hashtags associated with the id
- Joined on Id
- Filtered by which tweets have more than one hashtag
- Counted which hashtags were the most popular

ID	Hashtags
1	a, b, c, d
2	c, d, e
3	f, g

ID	Tag
1	a
1	b
1	c
1	d
2	c

ID	Num
1	4
2	3
3	2

Continued...

ID	Tag	Num
1	a	4
1	b	4
1	c	4
1	d	4
2	c	3

Filtered by Num >= 2

Counted the
number of tags

Tag	Count
a	1
b	1
c	2
d	2

Actual Graph sorted by count:

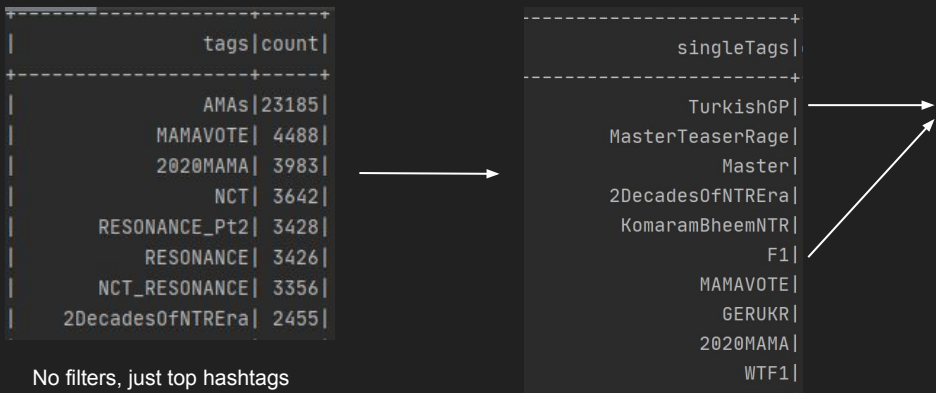
Runner		
	tags	count
↑		
↓		
↺	MAMAVOTE	4474
↻	2020MAMA	3797
↱	NCT	3621
🖨	RESONANCE_Pt2	3428
🗑	RESONANCE	3426
	NCT_RESONANCE	3356
	1111ShopeeStrayKids	2363
	BTS	2035
	Master	1721
	StayWithShopee	1621
	ShopeexChanyeoLEX0	1275
	MasterTeaser	1242

Fun fact: 'MAMAVOTE' appeared as a hashtag 4488 times.
Meaning that 99.7% of the time it was used with another hashtag.

Which hashtag is used the most by popular accounts?

And how does that compare with regular accounts?

- Define popular accounts
 - Organized tweets by sum(likes, retweets, replies) and associated IDs
 - Took the top 200 IDs as my 'popular accounts'
- Inner joined original DataFrame with popular accounts
- Found the most popular hashtag



The diagram illustrates the process of finding the most popular hashtag. It starts with a DataFrame containing a list of hashtags and their counts. An arrow points from this DataFrame to another DataFrame where the hashtags are listed individually. A second arrow points from the 'TurkishGP' entry in the second DataFrame to a text box explaining its significance.

tags count
AMAs 23185
MAMAVOTE 4488
2020MAMA 3983
NCT 3642
RESONANCE_Pt2 3428
RESONANCE 3426
NCT_RESONANCE 3356
2DecadesOfNTERa 2455

No filters, just top hashtags

singleTags
TurkishGP
MasterTeaserRage
Master
2DecadesOfNTERa
KomaramBheemNTR
F1
MAMAVOTE
GERUKR
2020MAMA
WTF1

Turkish GrandPrix is a Formula One racing event that lots of famous people go to.

Top User Engagement

- Looked at the top fifteen most followed accounts
- Used recent search to get the tweets
 - Searched using from username
 - Extracted public metrics
 - Called sum of metrics engagement
- Found total engagement and average engagement
- Compared to normalizing by the number of followers
- Looked at how engagement was distributed for each user

User	Followers (Millions)	Total Engagement (Per Thousand)	Average Engagement (per Thousand)	Like	Quote	Retweet	Reply
Barack Obama	126.45	6.52	1.63	85.8	2.5	7.6	4.1
Justin Bieber	113.17	11.01	0.79	77.6	3.5	15.4	3.6
KATY PERRY	108.94	0.73	0.18	78.4	4	13	4.7
Cristiano Ronaldo	89.39	5.82	1.46	94.9	0.4	3.8	1
Donald J. Trump	88.93	228.85	2.34	68.8	3	18	10.1
Taylor Swift	87.52	9.53	4.76	77.3	5.6	12.6	4.5
Lady Gaga	82.74	1.06	0.21	84	1.6	10.2	4.3
Ellen DeGeneres	79.51	1.95	0.08	91	1.2	4.6	3.2
Ariana Grande	79.05	37.4	1.78	82.4	2.5	9.5	5.5
YouTube	72.41	0.01	0	84.5	1.9	7.8	5.7

User	Followers (Millions)	Total Engagement (Per Thousand)	Average Engagement (per Thousand)	Like	Quote	Retweet	Reply
Kim Kardashian West	67.59	4.82	0.09	94.5	0.7	2.7	2.1
Justin Timberlake	64.23	0.03	0.01	77.5	0.5	15.2	6.8
Narendra Modi	63.67	45.29	0.74	88.2	0.7	9	2.1
Selena Gomez	63.61	2.4	0.24	82.8	1.7	11.4	4.1
CNN Breaking News	59.61	6.53	0.08	73.1	5.2	12.3	9.4

User	Total Engagement	Total Engagement (per Thous)	User	Shift	Average Engagement	Average Engagement (per Thous)
Donald J. Trump	20352	228.85	Taylor Swift	+4	417	4.76
Ariana Grande	2956.7	37.4	Donald J. Trump	-1	207.7	2.34
Narendra Modi	2883.9	45.29	Barack Obama	+3	206.2	1.63
Justin Bieber	1245.8	11.01	Ariana Grande	-2	140.8	1.78
Taylor Swift	834	9.53	Cristiano Ronaldo	+2	130.1	1.46
Barack Obama	824.9	6.52	Justin Bieber	-2	89	0.79
Cristiano Ronaldo	520.6	5.82	Narendra Modi	-4	47.3	0.74
CNN Breaking News	389.1	6.53	KATY PERRY	+5	19.8	0.18
Kim Kardashian West	325.5	4.82	Lady Gaga	+3	17.6	0.21
Ellen DeGeneres	155.2	1.95	Selena Gomez	+1	15.3	0.24
Selena Gomez	152.8	2.4	Ellen DeGeneres	-1	6.5	0.08
Lady Gaga	87.9	1.06	Kim Kardashian West	-3	6.4	0.09
KATY PERRY	79.1	0.73	CNN Breaking News	-5	4.5	0.08

Languages associated with Topics

- First looked at a sampled stream of data
 - Would have worked better with a larger data set
 - It stopped working for me when ever I used the internet on my computer
- Then used recent search at UTC 08:00:00 on November 17, 2020
 - Searched by context
 - Looked for last 5000 tweets about each of the contexts
 - Searched until ID of oldest tweet in the last search (50 searches of 100 tweets)
 - Loop through all contexts
- Multiple context IDs can occur in multiple tweet allowing some topics to occur more than the 5000 times tweets they were searched in
- Used Trump, Biden, Shinzo Abe, Yoshihide Suga , and Angela Merkel
 - Wanted widely spaced countries speaking different languages

Topic	Occurrences	Language	Percent
Donald Trump (Donald Trump)	14228	en	81.43
Donald Trump (Donald Trump)	14228	und	6.87
Donald Trump (Donald Trump)	14228	ja	3.36
Donald Trump (Donald Trump)	14228	fr	1.9
Donald Trump (Donald Trump)	14228	es	1.84

Topic	Occurrences	Language	Percent
Joe Biden (Joe Biden)	11834	en	83.83
Joe Biden (Joe Biden)	11834	und	5.05
Joe Biden (Joe Biden)	11834	es	3.14
Joe Biden (Joe Biden)	11834	fr	1.94
Joe Biden (Joe Biden)	11834	de	1.47
Joe Biden (Joe Biden)	11834	ja	1.35

Topic	Occurrences	Language	Percent
Abe Shinzo (Abe Shinzo)	10424	ja	95.53
Abe Shinzo (Abe Shinzo)	10424	en	2.34
Abe Shinzo (Abe Shinzo)	10424	und	1.73
Yoshihide Suga (Yoshihide Suga)	10398	ja	92.75
Yoshihide Suga (Yoshihide Suga)	10398	und	3.77
Yoshihide Suga (Yoshihide Suga)	10398	en	3.17

Topic	Occurrences	Language	Percent
Angela Merkel (Angela Merkel)	9970	de	53.56
Angela Merkel (Angela Merkel)	9970	en	21.83
Angela Merkel (Angela Merkel)	9970	es	9.39
Angela Merkel (Angela Merkel)	9970	fr	4.11
Angela Merkel (Angela Merkel)	9970	und	3.71
Angela Merkel (Angela Merkel)	9970	ja	2.17
Angela Merkel (Angela Merkel)	9970	it	1.34
Angela Merkel (Angela Merkel)	9970	tr	1.2

Topic	Occurrences	Language	Percent
COVID-19 (COVID-19)	3603	ja	32.69
COVID-19 (COVID-19)	3603	en	31.22
COVID-19 (COVID-19)	3603	de	22.09
COVID-19 (COVID-19)	3603	es	5.16
COVID-19 (COVID-19)	3603	und	4.36
COVID-19 (COVID-19)	3603	fr	1.47
COVID-19 (COVID-19)	3603	it	1.19
Tokyo 2021 Summer Olympics (Tokyo 2021 Summer Olympics)	1306	ja	90.81
Tokyo 2021 Summer Olympics (Tokyo 2021 Summer Olympics)	1306	en	6.89
Tokyo 2021 Summer Olympics (Tokyo 2021 Summer Olympics)	1306	und	1.61

Topic	Occurrences	Language	Percent
TV/Movies Related (TV/Movies Related)	778	ja	51.67
TV/Movies Related (TV/Movies Related)	778	en	40.87
TV/Movies Related (TV/Movies Related)	778	und	4.63
Barack Obama (Barack Obama)	714	en	47.62
Barack Obama (Barack Obama)	714	de	29.69
Barack Obama (Barack Obama)	714	ja	11.76
Barack Obama (Barack Obama)	714	und	3.36
Barack Obama (Barack Obama)	714	fr	3.36
Barack Obama (Barack Obama)	714	nl	2.24

Topic	Language	Interactions	Topic	Language	Interactions
Donald Trump (Donald Trump)	en	8.92E+07	Bernie Sanders (Bernie Sanders)	en	1.36E+06
Joe Biden (Joe Biden)	en	6.50E+07			
George Floyd (George Floyd)	en	8.42E+06	Tokyo 2021 Summer Olympics (Tokyo 2021 Summer Olympics)	ja	1.35E+06
Donald Trump (Donald Trump)	und	7.17E+06	Tom Fitton (Tom Fitton)	en	1.29E+06
Abe Shinzo (Abe Shinzo)	ja	6.95E+06	Barack Obama (Barack Obama)	en	9.95E+05
COVID-19 (COVID-19)	en	5.74E+06	Services (Services)	en	9.00E+05
Yoshihide Suga (Yoshihide Suga)	ja	4.87E+06	Lara Trump (Lara Trump)	en	8.28E+05
Kamala Harris (Kamala Harris)	en	3.24E+06	Sylvester Stallone (Sylvester Stallone)	en	8.28E+05
Tracy Beanz (Tracy Beanz)	en	3.02E+06	Ron Perlman (Ron Perlman)	en	8.14E+05
Dan Scavino (Dan Scavino)	en	1.89E+06	The White House (The White House)	en	7.66E+05
United States Congress (United States Congress)	en	1.75E+06	Online Site (Online Site)	en	7.04E+05
Michael Flynn (Michael Flynn)	en	1.64E+06	Home & family (Home & family)	en	6.85E+05

What is the most popular recurring twitter hashtag that is mainly used on {day}? (Wednesday, Thursday, Friday, Saturday, and Sunday)?

Goal: Look for the most popular recurring hashtags for (Wednesday, Thursday, Friday, Saturday, Sunday). *Hashtag has to start with the name of the day*

Examples: #WednesdayMotivation, #FridayFeeling, #ThursdayThoughts

Data used: Collected sample Twitter real-time data stream for analysis from 10am - 3pm eastern standard time from Wednesday to Sunday (11/18 - 11/22)

Total: 25 hours

According to Sysomos, Twitter's peak times are between 11a and 3pm est

Limited the amount of data used to 850,000 tweets per day for consistency

How many times did #{hashtag} get mentioned in a tweet on {day}?
(WednesdayWisdom, TBT, FridayFeeling, Caturday, SundayFunday)

Goal: See how many times the hashtag was used on a particular day

According to Twitter Business, #WednesdayWisdom, #TBT, #FridayFeeling, #Caturday, and #SundayFunday are top recurring hashtags

Data used: Sample Twitter real-time data stream from 10am - 3pm eastern standard time from Wednesday to Sunday (same as previous question)

Limited the amount of data used to 850,000 tweets per day

Stored data in parquet files

Overview of Process

- Filter data in dataframe so we are only dealing with data.text
- Split tweets (data.text) into individual words and count the number of occurrences for each word
- Perform group by on word column
- Filter words so only hashtags will appear in the table
- Filter words so only certain hashtags will appear in the table

=== Wednesday's top hashtags ===

hashtag	count
#wednesdaythought	45
#Wednesday	39
#WednesdayMotivation	25
#WednesdayWisdom	19
#wednesdaythought...	9
#wednesdaywisdom	4
#WednesdayThoughts	4
#Wednesdayvibe	3
#WednesdayInspira...	3
#wednesdayfics	2
#WednesdayThought...	2
#wednesdaythought...	2
#WednesdayMorning...	2
#wednesdays	2
#WednesdayThought	2
#wednesdaymotivation	2
#WednesdayRockNight	1
#wednesdaythought...	1
#WednesdayMotivat...	1
#wednesdaymorning	1

=== Thursday's top hashtags ===

hashtag	count
#TBT	52
#ThrowbackThursday	47
#thursdaymorning	24
#thursdayvibes	22
#ThursdayThoughts	20
#ThursdayMotivation	17
#Thursday	6
#ThursdayMotivati...	3
#ThursdayMorning	2
#ThursdayNightFoo...	2
#ThursdayThoughts	2
#ThursdayVibez	2
#ThursdayVibes	2
#thursdayswithjc	2
#ThursdayThanks	1
#ThursdayFuckFrom...	1
#ThursdayTip	1
#thursdayvibes	1
Morty	1
#ThursdayThoughts...	1
#ThursdayMotivati...	1

=== Friday's top hashtags ===

hashtag	count
#FridayLivestream	576
#FridayLivestream...	233
#FridayLivestream...	49
#FridayFeeling	41
#fridaymorning	40
#FridayLivestream...	38
#FridayMotivation	24
#FridayThoughts	22
#FridayVibes	19
#Friday	16
#friday	14
#FridayLivestream...	11
#FridayHappiness	10
#FridayLivestream...	7
#FridaysForFuture	7
#FridayLivestream...	7
#FridayLivestream	6
#FridayLivestream...	6
#FridayLivestream...	5
#FridayLivestream	5

Out of 850,000 tweets

=== Saturday's top hashtags ===

hashtag	count
#SaturdayVibes	29
#SaturdayMorning	25
#SaturdayThoughts	23
#Caturday	23
#SaturdayHappiness	23
#SaturdayMotivation	22
#saturday	9
#SaturdaysAreGoodFor	7
#SaturdayHappines...	6
#SaturdaySunshine	5
#Saturday	4
#saturdayvibes	3
#SaturdayHappines...	3
#SaturdayMood	3
#SaturdayThoughts	3
By	3
#SaturdayThoughts...	3
#Caturday,	3
#SaturdayThoughts...	2
#Caturday	2
https://...	2
#SaturdayThoughts...	2



=== Sunday's top hashtags ===

hashtag	count
#SundayGrandWord	81
#sundayvibes	77
#SundayThoughts	34
#SundayMorning	30
#SundayMotivation	14
#SundayNightMovie	14
#SundayFunday	12
#Sunday	10
#sunday	9
#sundayfunday	6
#SundayFunday...	4
#SundayHappiness	4
#SundayFeels	3
#SundaySpirit	3
#sundaythoughts	3
#SundayInspiration	3
#SundayVibes	3
#SundayGame	2
#sundayworship	2
#sundaynightfootball	2



Sources

Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."

Proceedings of the ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,

Github link

https://github.com/LiamAHood/Project_2_Group_4.git