

# Twitter Big Data Analysis

**Project 2 Group 3**  
**Clair Houlihan, Mindy Jen,**  
**Jefferson Santiago, Melissa White,**  
**Yasmin Yazdi**





# Main Theme - An overview of Group 3 Analysis

## → Questions:

### ◆ **Clair:**

- What are some of the other hashtags that are being used with #BTS?
- What are the average amount of likes that a common BTS post receives?

### ◆ **Yasmin:**

- What are the common words used with the BTS hashtag?
- How does it vary with time?

### ◆ **Jeff:**

- How is Kpop being compared/linked to American pop?
- Which other Kpop artists are famous in the americas?
- Which artist is the most trending on Twitter?
- Which states/country tweets the most about Kpop?

### ◆ **Melissa:**

- Analysis on tweets with BTS/BTS-related hashtags and mentions in different languages
- How many languages are tweeting about BTS?
- What is the distribution of languages tweeting about BTS?

### ◆ **Mindy:**

- How is #BTS distributed globally?
- What countries would you suggest BTS to hold a series of global concert tours in 2022?
- What languages would you recommend BTS to speak in their global concert tours?





# Filter Stream Data : Twitter API - Rule/Filed Objects

List of times and their corresponding time at another timezones

Time Hours (PDT)	CST	GMT+8 (Asia)	GMT 0 (Europe)	
12:00:00 AM	2:00:00 AM	4:00:00 PM	8:00:00 AM	
1:00:00 AM	3:00:00 AM	5:00:00 PM	9:00:00 AM	
2:00:00 AM	4:00:00 AM	6:00:00 PM	10:00:00 AM	
3:00:00 AM	5:00:00 AM	7:00:00 PM	11:00:00 AM	Melissa (Tuesday, Asia)
4:00:00 AM	6:00:00 AM	8:00:00 PM	12:00:00 PM	
5:00:00 AM	7:00:00 AM	9:00:00 PM	1:00:00 PM	
6:00:00 AM	8:00:00 AM	10:00:00 PM	2:00:00 PM	Jefferson (weekend)
7:00:00 AM	9:00:00 AM	11:00:00 PM	3:00:00 PM	
8:00:00 AM	9:00:00 AM	10:00:00 AM	11:00:00 AM	Mindy (Saturday)
9:00:00 AM	11:00:00 AM	1:00:00 AM	5:00:00 PM	
10:00:00 AM	12:00:00 PM	2:00:00 AM	6:00:00 PM	
11:00:00 AM	1:00:00 PM	3:00:00 AM	7:00:00 PM	Melissa (Monday, Europe)
12:00:00 PM	2:00:00 PM	4:00:00 AM	8:00:00 PM	Clair (Saturday)
1:00:00 PM	3:00:00 PM	5:00:00 AM	9:00:00 PM	Clair (Saturday)
2:00:00 PM	4:00:00 PM	6:00:00 AM	10:00:00 PM	Yasmin (Sunday)
3:00:00 PM	5:00:00 PM	7:00:00 AM	11:00:00 PM	Yasmin (Sunday)
4:00:00 PM	6:00:00 PM	8:00:00 AM	12:00:00 AM	
5:00:00 PM	7:00:00 PM	9:00:00 AM	1:00:00 AM	Melissa (Monday, CST)
6:00:00 PM	8:00:00 PM	10:00:00 AM	2:00:00 AM	
7:00:00 PM	9:00:00 PM	11:00:00 AM	3:00:00 AM	
8:00:00 PM	10:00:00 PM	12:00:00 PM	4:00:00 AM	Jefferson
9:00:00 PM	11:00:00 PM	1:00:00 PM	5:00:00 AM	
10:00:00 PM	12:00:00 AM	2:00:00 PM	6:00:00 AM	
11:00:00 PM	1:00:00 AM	3:00:00 PM	7:00:00 AM	

```
curl -X POST 'https://api.twitter.com/2/tweets/search/stream/rules' \
```

```
-H "Content-type: application/json" \
```

```
-H "Authorization: Bearer $TWITTER_BEARER_TOKEN" -d \
```

```
{
```

```
  "add": [
```

```
    {
```

```
      "value": "@bts OR #bts",
```

```
      "tag": ""
```

```
    },
```

```
  ]
```

```
  "value": "@kpop OR #kpop OR @k_pop OR #k_pop",
```

```
  "tag": ""
```

```
}
```

```
}
```

```
}
```

```
curl
```

```
'https://api.twitter.com/2/tweets/search/stream?tweet.fields=lang,geo,public_metrics,
created_at&expansions=geo.place_id&place.fields=full_name' -H "Authorization:
Bearer $TWITTER_BEARER_TOKEN"
```



# Filter Stream Data Sampling for Various Types of Analyses

## ★ Clair

- Requires the field value 'public\_metrics' required for one of the analysis

## ★ Yasmin

- Requires the field value 'text' required for two of the analysis
- Requires the field value 'created\_at' for one of the analysis

## ★ Jeff

- Requires the field value 'public\_metrics' required for two of the analysis
- Requires the field value 'geo' for retrieving location of tweet

## ★ Melissa

- Requires the field value 'lang' to retrieve languages for analysis

## ★ Mindy

- Require one of the field\_value options in tweet\_fields be 'geo'
- Retrieve place\_object by calling expansions where location '**full\_name**' is specified as field\_value

# Clair

What are some of the other hashtags that are being used with #BTS?

Some of the top ones are:

**#SUGA** **#JIMIN** **#JIN** **#JHOPE** **#JUNGKOOK** **#TAEHYUNG**, are all members of BTS. **#ATEEZ** is another kpop group. **#VoteBTS** relates to a poll of the most popular group in music, **#ARMY** is associated with the culture of BTS. **#hobi** and **#V** are nicknames for members of the group.

value	count
#SUGA	45
#JIMIN	27
#JIN	6
#JHOPE	6
#JUNGKOOK	5
#TAEHYUNG	5
#BTSJIN	5
#kpop	4
#ATEEZ	4
#jimin	3
#V	3
#VoteBTS	3
#army	3
#BestBand	3
#kpoptwt	3
#jungkook	3
#taehyung	3
#ARMY	3
#valentinesday2021	2
#hobi	2

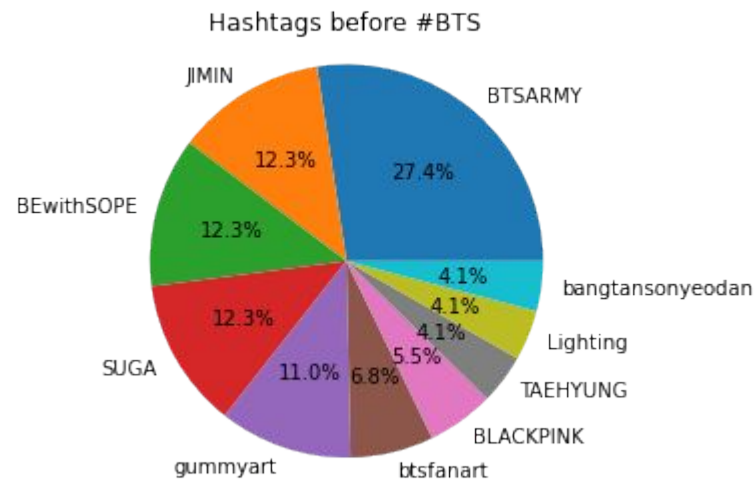
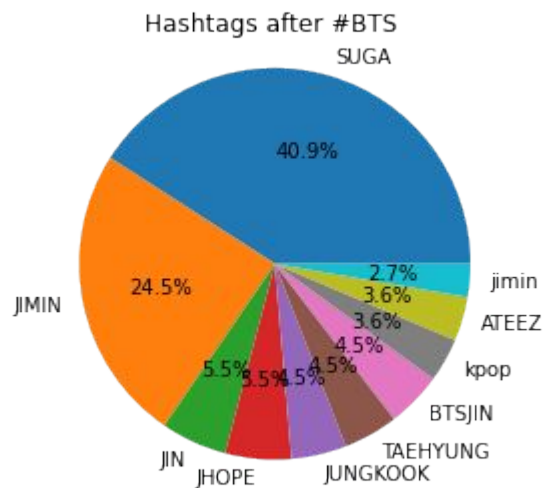
# Clair

What are some of the other hashtags that are being used with #BTS?

#BEwithSOPE is related to J-Hope, #gummyart appears to be an artist, #BLACKPINK is another kpop group, #Lighting appears to be related to BTS themed lighting, #DAY6 is a rock band from South Korea, #nct is another group from South Korea

value	count
#BTSARMY	20
#JIMIN	9
#BEwithSOPE	9
#SUGA	9
#gummyart	8
#btsfanart	5
#BLACKPINK	4
#TAEHYUNG	3
#Lighting	3
#bangtansonyeodan	3
#suga	3
#DAY6	3
#nct	3
#ARMY	3
#jhopeEnBBVA	2
#txt	2
#army	2
#DOYOUNG	2
#trophywife	2
#Army	2

# Clair





# Clair

```
avg: 612662.1  
max: 851746  
min: 426544
```

What are the average amount of likes that a common BTS post receives?

The average here is calculated with 10 of the most recent tweets (from the last few days) where the tweets were in both english and korean. The maximum and minimum amount of likes is posted for comparison.



## What are the average amount of likes that a common BTS post receives?

Here are what the tweets look like for reference:

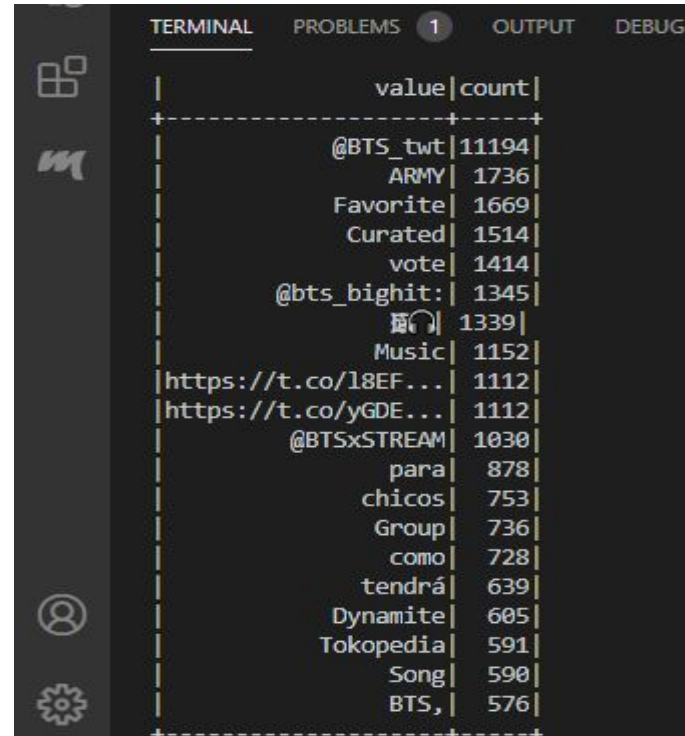
[illegible]

# Yasmin

What are the common words used with the BTS hashtag?

Combined data showed Army, @BTS\_twt, and Favorite as top words

Army refers to the fans of BTS



A screenshot of a terminal window with a dark background. The terminal title bar shows 'TERMINAL', 'PROBLEMS 1', 'OUTPUT', and 'DEBUG'. On the left side of the terminal, there are icons for a file explorer, a code editor, a user profile, and a settings gear. The main content of the terminal is a table with two columns: 'value' and 'count'. The table is enclosed in a dashed border. The data rows are as follows:

value	count
@BTS_twt	11194
ARMY	1736
Favorite	1669
Curated	1514
vote	1414
@bts_bighit:	1345
🎧	1339
Music	1152
<a href="https://t.co/l8EF...">https://t.co/l8EF...</a>	1112
<a href="https://t.co/yGDE...">https://t.co/yGDE...</a>	1112
@BTSxSTREAM	1030
para	878
chicos	753
Group	736
como	728
tendrá	639
Dynamite	605
Tokopedia	591
Song	590
BTS,	576

# Yasmin

How does this vary over time?

At 7-8pm (CST) and 7-10am

Less english words used at night

Most popular (BTS\_twt) shows up at 10pm the most

value	count
create	194
event	194
@kpopidol_en:	129
will	110
support	104
want	99
whose	97
winn...	97
[VOTE]	97
April	97
♥we	97
fundraising	97
15-21	97
more	29
this	27
your	26
[FAN:DRIVE]	19
Request	18
65th	16
name	16

value	count
@BTS_twt	211
back	206
@jimintoday_:	167
방탄소년단	162
'Film	160
'시그널'	158
out'	157
꼭이다.	156
[공식]	156
https://t.co/qNMK...	156
아웃'은	156
number와의	156
방탄소년단과	156
bac...	155
新曲「Film	144
@bts_bighit:	122
ARMY	110
@BTS_jp_official:	108

## List of Popular Kpop Artist other than BTS based on tweets

ATEEZofficial `{"value": "@ATEEZofficial", "count": 221}`

- [@ATEEZofficial](#)

BLACKPINK `{"value": "@BLACKPINK", "count": 62}`

- [@BLACKPINKOFFICIAL](#)

Keeper `{"value": "@reminniescence", "count": 29}`

- [keeper \(@reminniescence\)](#)

Exo\*

- [EXO \(@weareoneEXO\)](#)

BTS Tweets: `{"value": "@BTS", "count": 15385}`

- [방탄소년단 \(@BTS twt\)](#)

# Which artist is the most trending on Twitter? (Other than BTS)

- Blackpink: #LISA, #JENNIE
- NCT: #Taeyong
- Exo: #EXO

```
1 {"text": "RT @kpopidol_en: [VOTE]\nWhose support event you want to create on #CHOEAEODOL\nApril 15-21\n\n♥ We will create a fundraising event for the winn...", "sum(retweet_count)": 2343191}
2 {"text": "RT @kpopspotifydata: Fastest Songs By Korean Artists To Reach 600 MILLION Streams On Spotify :\n\n#1 @BTS_twt - Dynamite - 144 Days\n#2 #BTS -...", "sum(retweet_count)": 174692}
3 {"text": "RT @kpopidol_en: Announcing the 120th #charityfairystar\n\n#Jin #BTS \n\nCelebrating #10300DaysWithJin \n\n$500k($410) Charity donated in the name of...", "sum(retweet_count)": 56222}
4 {"text": "RT @kpopidol_en: 📢 Announcing the 65th #CharityAngel 📢\n\n#EXO #엑소 \n\n$500k($410) Charity donated in the name of EXO-\n\nVote for your Bias\n\n💎 Do...", "sum(retweet_count)": 51854}
5 {"text": "RT @kennethcolereal: \"BTS helped save my mental health\" - @MsLeaSalonga https://t.co/uQBGlpwRF2 #BTS #kpop", "sum(retweet_count)": 30034}
6 {"text": "RT @kpopidol_en: ♥️♥️ DONATION CERTIFICATION♥️♥️\n\n107th #CharityFairystar : #DO #EXO\n\nIdols who receives > 55,555,555 votes during anniversaries wow...", "sum(retweet_count)": 19719}
7 {"text": "RT @dreamwastaken: stop making fun of fandoms.\n\n#kpop is OUTSTANDING.\n\n#muffintwt is AMAZING.\n\n#sleepytw is INCREDIBLE.\n\n#dttwt\n\n#hermitcr...", "sum(retweet_count)": 16860}
8 {"text": "RT @kpopggsuperior: .@BLACKPINK's #LISA is now the most subscribed Kpop Female Solo Artist/Individual on YouTube (5.65M) surpassing her bes...", "sum(retweet_count)": 10718}
9 {"text": "RT @kpopiness: At least he touches grass unlike kpop stans", "sum(retweet_count)": 8444}
10 {"text": "RT @kpop_sbs: [물#사이드캠] TREASURE - MY TREASURE\n\n#인기가요 #TREASURE #트레저 #MYTREASURE @ygtreasuremaker\n\nhttps://t.co/hZDJcnoUZt https://t.co/jza...", "sum(retweet_count)": 7531}
11 {"text": "RT @kpophappenings_: when that kpop boy gave a plushie to a fan who was in a wheelchair today", "sum(retweet_count)": 7292}
12 {"text": "RT @ericnamofficial: Fam - please take a moment to sign this petition. We need these acts of hate to end NOW. https://t.co/ddLwMHbH1G #kpop...", "sum(retweet_count)": 7118}
13 {"text": "RT @kpophighindia: \"I don't want my face to be on every Billboard, but I want our kind, Desis to be known in the mainstream\"", "sum(retweet_count)": 6226}
14 {"text": "RT @mtvasia: february is right round the corner, so i wanna know which feb comeback are you most HYPED about? 🙋\n\n#kpop #THESHOW #comeback...", "sum(retweet_count)": 5169}
15 {"text": "RT @dispatchsns: Which style do you prefer?\n\nhttps://t.co/gvztKhTyfw\n\n#NCT #엔시티 #TAEYONG #태용 #tiktokstage #틱톡스테이지 #kpop #fyp", "sum(retweet_count)": 4801}
16 {"text": "RT @kpoplover727: Baekhyun said recently that Chanyeol's fine\n\n", "sum(retweet_count)": 3566}
17 {"text": "RT @radiokpopway: 📢 THE REQUESTS PLAYLIST📢\n\n(Top 30 most requested)\n\nRequest your favorite #kpop songs!\n\nRequest time: Feb.12, 11pm to Feb.13...", "sum(retweet_count)": 2858}
18 {"text": "RT @pressreels: @Weeekly's cuteness and freshness tickles us! Look forward to their honesty #kpop\n\nNOW: https://t.co/EWsUTd15T2 https://t.co/...", "sum(retweet_count)": 2714}
19 {"text": "RT @kpopggsuperior: .@BLACKPINK's #LISA & #JENNIE are currently tied as the Most Subscribed Kpop Female Solo/Individual on YouTube (5.64M)...", "sum(retweet_count)": 2649}
20 {"text": "RT @kpopggsuperior: \"Celebrity\" by #IU becomes the first song by a Female Artist to achieve 300th Perfect All-Kill in history 🌟🔥 https://t.co/...", "sum(retweet_count)": 2416}
```

## Which US states tweets the most about Kpop?

- California : 42 tweets
- Texas: 70 tweets
- Illinois: 31 tweets
- New York: 32 tweets

```
{ "Place": ["Rio de Janeiro, Brazil"], "count": 197 }
{ "Place": ["İstanbul, Türkiye"], "count": 89 }
{ "Place": ["Myanmar"], "count": 64 }
{ "Place": ["Sao Paulo, Brazil"], "count": 62 }
{ "Place": ["Jeddah, Kingdom of Saudi Arabia"], "count": 49 }
{ "Place": ["Riyadh, Kingdom of Saudi Arabia"], "count": 47 }
{ "Place": ["Ciudad Autónoma de Buenos Aires, Argentina"], "count": 46 }
{ "Place": ["Los Angeles, CA"], "count": 42 }
{ "Place": ["Houston, TX"], "count": 38 }
{ "Place": ["Bogotá, D.C., Colombia"], "count": 35 }
{ "Place": ["Manhattan, NY"], "count": 32 }
{ "Place": ["Saladoblanco, Colombia"], "count": 32 }
{ "Place": ["Chicago, IL"], "count": 31 }
{ "Place": ["Brasília, Brazil"], "count": 29 }
{ "Place": ["Dallas, TX"], "count": 27 }
{ "Place": ["Toronto, Ontario"], "count": 26 }
{ "Place": ["New Delhi, India"], "count": 25 }
{ "Place": ["Dubai, United Arab Emirates"], "count": 25 }
```



## Melissa: Twitter Data Analysis on Languages

---



Analysis on tweets with BTS and BTS-related hashtags and mentions in different languages:

- How many languages are tweeting about BTS?
- What is the distribution of languages?



## Languages

```
"created_at": "2021-02-17T01:36:18.000Z",  
"id": "1361851939273596930",  
"text": "@bts_bighit HALA OMG CUTEETEEE",  
"lang": "et"
```

```
"lang": "es",  
"id": "1361849831011024896",  
"created_at": "2021-02-17T01:27:56.000Z",  
"text": "@bts_bighit Estoy llorando mares"
```

```
"lang": "pt",  
"id": "1361852301540020224",  
"text": "@bts_bighit Amei o Bangtani! Kkk💜💛💛💛",  
"created_at": "2021-02-17T01:37:45.000Z"
```

```
"lang": "in",  
"text": "@bts_bighit @ARMYTEAMIID ada yg tau cara download video design ruang semua member di sini?",  
"id": "1361850443488395266",  
"created_at": "2021-02-17T01:30:22.000Z"
```



# lang



Note: if no language classification can be made the provided result is 'und' (for undefined).

Example: `recommend #paris lang:en`

The list below represents the currently supported languages and their corresponding BCP 47 language identifier:

Amharic: <code>am</code>	German: <code>de</code>	Malayalam: <code>ml</code>	Slovak: <code>sk</code>
Arabic: <code>ar</code>	Greek: <code>el</code>	Maldivian: <code>dv</code>	Slovenian: <code>sl</code>
Armenian: <code>hy</code>	Gujarati: <code>gu</code>	Marathi: <code>mr</code>	Sorani Kurdish: <code>ckb</code>
Basque: <code>eu</code>	Haitian Creole: <code>ht</code>	Nepali: <code>ne</code>	Spanish: <code>es</code>
Bengali: <code>bn</code>	Hebrew: <code>iw</code>	Norwegian: <code>no</code>	Swedish: <code>sv</code>
Bosnian: <code>bs</code>	Hindi: <code>hi</code>	Oriya: <code>or</code>	Tagalog: <code>tl</code>

Bulgarian: <code>bg</code>	Latinized Hindi: <code>hi-Latn</code>	Panjabi: <code>pa</code>	Tamil: <code>ta</code>
Burmese: <code>my</code>	Hungarian: <code>hu</code>	Pashto: <code>ps</code>	Telugu: <code>te</code>
Croatian: <code>hr</code>	Icelandic: <code>is</code>	Persian: <code>fa</code>	Thai: <code>th</code>
Catalan: <code>ca</code>	Indonesian: <code>in</code>	Polish: <code>pl</code>	Tibetan: <code>bo</code>
Czech: <code>cs</code>	Italian: <code>it</code>	Portuguese: <code>pt</code>	Traditional Chinese: <code>zh-TW</code>
Danish: <code>da</code>	Japanese: <code>ja</code>	Romanian: <code>ro</code>	Turkish: <code>tr</code>
Dutch: <code>nl</code>	Kannada: <code>kn</code>	Russian: <code>ru</code>	Ukrainian: <code>uk</code>

English: <code>en</code>	Khmer: <code>km</code>	Serbian: <code>sr</code>	Urdu: <code>ur</code>
Estonian: <code>et</code>	Korean: <code>ko</code>	Simplified Chinese: <code>zh-CN</code>	Uyghur: <code>ug</code>
Finnish: <code>fi</code>	Lao: <code>lo</code>	Sindhi: <code>sd</code>	Vietnamese: <code>vi</code>
French: <code>fr</code>	Latvian: <code>lv</code>	Sinhala: <code>si</code>	Welsh: <code>cy</code>
Georgian: <code>ka</code>	Lithuanian: <code>lt</code>		

## IETF language tag

An IETF BCP 47 language tag is a code to identify human languages. For example, the tag `en` stands for English; `es-LA` for Latin American Spanish; `rm-sursilv` for Sursilvan; `gsw-u-sd` for

## Language Statistics



- Our sample data consisted of 714, 836 tweets
- Out of those tweets, people tweeted in 64 different languages

```
total tweets about BTS
```

```
+-----+  
|total_tweets|  
+-----+  
|714836      |  
+-----+
```

```
number of languages
```

```
+-----+  
|count(DISTINCT lang)|  
+-----+  
|64                  |  
+-----+
```

# Language Statistics



- Top results out of total sample:
  - English (33.3%), Japanese (14.7%), "Undefined"\* (8.6%), Korean (8.6%), Spanish (7.8%)
- Least popular languages (<0.0000013%):
  - Maldivian (dv), Lao (lao), Armenian (hy), Georgian (ka)

language count / total tweets = ratio		
lang	count	lang_to_total_ratio
en	237903	0.33280780486713035
ja	105124	0.14706030474122736
und	61767	0.08640723186856845
ko	58544	0.08189850539144643
es	55966	0.07829208377865693
th	42058	0.05883587284356132
pt	33260	0.04652815470961171
in	28573	0.03997140602879542
ar	21328	0.029836214180595268
tr	17710	0.02477491340671147
fr	10083	0.014105333251263227
tl	8658	0.01211186901610999
hi	4831	0.006758193487737048
it	3541	0.004953583759072011
ru	3264	0.004566082290203627
de	2138	0.0029908958138649983
nl	1982	0.0027726639397008543
fa	1756	0.002456507506616902
pl	1752	0.0024509118175357705

sl	128	1.7906205059622067E-4
mr	104	1.454879161094293E-4
is	104	1.454879161094293E-4
te	103	1.4408899383914632E-4
sr	76	1.0631809254150603E-4
gu	56	7.833964713584655E-5
ml	52	7.274395805471465E-5
bg	47	6.574934670329977E-5
pa	41	5.7355813081601934E-5
kn	40	5.595689081131896E-5
ps	28	3.9169823567923274E-5
am	23	3.21752122165084E-5
si	14	1.9584911783961637E-5
ckb	13	1.8185989513678664E-5
km	13	1.8185989513678664E-5
or	12	1.6787067243395687E-5
dv	9	1.2590300432546767E-5
lo	5	6.99461135141487E-6
hy	4	5.595689081131896E-6
ka	3	4.196766810848922E-6

\*If no language classification can be made, the provided result is "und" for undefined



# Language Statistics and Timezones

"Americas" ~7-8pm CST

```
language count / total tweets = ratio
```

lang	count	lang_to_total_ratio
ko	7874	0.2830744895024446
ja	7691	0.2764955421340236
en	5241	0.1884167385677308
in	1802	0.06478285878631004

**Korean: 28.3%**  
 Japanese: 27.6%  
 English: 18.8%

"Asia" ~7-8pm

lang	count	lang_to_total_ratio
ko	11463	0.349407138720395
und	6919	0.21090011278080897
en	6356	0.19373914103697382
ja	2285	0.0696497698661871
in	2034	0.06199896363580943
es	1409	0.042948151309171824

**Korean: 34.9%**  
 English: 19.4%  
 Japanese: 6.96%  
 Indonesian: 6.2%

"Europe" ~7-8pm


lang	count	lang_to_total_ratio
ja	8101	0.31402876303446137
ko	7674	0.2974764507500872
en	4228	0.1638950265534752
es	1385	0.053688413381400936
in	1298	0.05031592820870644
und	1178	0.04566422452223127
pt	836	0.03240686901577703
fr	457	0.017715238205992946

**Japanese: 31.4%**  
 Korean: 29.7%  
 ...  
**Portuguese: 3.2%**  
**French: 1.77%**

### Q2: What languages would you recommend BTS to speak in their global concert tours?







```
File Edit Selection View Go Run Terminal Help
Runner.scala - tweetsapp [WSL: Ubuntu-20.04] - Visual Studio Code
EXPLORER
OPEN EDITORS
TWEETSAPP [WSL: UBUNT...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
cjen@DESKTOP-CE98V3V: ~/210104-usf-bigdata/week6/tweetsapp$ spark-submit ./target/scala-2.11/tweetsapp-assembly-0.1.0-SNAPSHOT.jar tweetsSparkSql twitterFilteredStream_210213_geoOnly.json
ana_tweetFilteredStream_210213_geoOnly.json
```

# Mindy: How is #BTS distributed globally?

```
File Edit Selection View Go Run Terminal Help
Runner.scala - tweetsapp [WSL: Ubuntu-20.04] - Visual Studio Code
EXPLORER
OPEN EDITORS
TWEETSAPP [WSL: UBUNT...
ana2_BTSkpop_geoOnly...
ana2_BTS_geoOnly_2...
ana2_BTS_geoOnly_2...
ana2_BTSkpop_geoOnly_21...
build.sbt
fruits.csv
persons.json
twitterFilteredStream...
twitterFilteredStream...
OUTLINE
Ubuntu-20.04
def main(args: Array[String]): Unit = {
    if(args.length != 3) {
        println("Options:")
        println("tweetsSparkSql [input file] [output folder]")
        println("tweetMoreSparkSql [input folder/input file] [output folder]")
        println("json2csv [input folder/input file] [output folder]")
        System.exit(1)
    }
    MapPartitionsRDD[71] at rdd at tweetsSparkSql.scala:51 []
    MapPartitionsRDD[70] at rdd at tweetsSparkSql.scala:51 []
    ShuffledRowRDD[69] at rdd at tweetsSparkSql.scala:51 []
    + (200) MapPartitionsRDD[68] at rdd at tweetsSparkSql.scala:51 []
    | MapPartitionsRDD[64] at rdd at tweetsSparkSql.scala:51 []
    | ShuffledRowRDD[63] at rdd at tweetsSparkSql.scala:51 []
    + (1) MapPartitionsRDD[62] at rdd at tweetsSparkSql.scala:51 []
    | MapPartitionsRDD[61] at rdd at tweetsSparkSql.scala:51 []
    | FileScanRDD[60] at rdd at tweetsSparkSql.scala:51 []
    21/02/20 11:39:38 WARN WindowExec: No Partition Defined for Window operation! Moving all data to a single partition, this can cause serious performance degradation.
    +-----+-----+-----+-----+-----+-----+
    |full_name|counts|avg_Retweet|avg_Replies|avg_Like|avg_Quotes|index|
    +-----+-----+-----+-----+-----+-----+
    |[Botswana]|30|0.2|0.0|0.2|0.0|1|
    |[Montevideo, Uruguay]|20|0.0|0.0|0.0|0.0|2|
    |[San José, Costa Rica]|20|0.0|0.0|0.0|0.0|3|
    |[Georgia]|19|0.0|0.0|0.0|0.0|4|
    |[Kalideres, Indonesia]|12|0.0|0.0|0.0|0.0|5|
    |[Santa Maria, Ilocos Region]|7|0.0|0.0|0.0|0.0|6|
    |[Bengaluru South, India]|7|0.0|0.0|0.0|0.0|7|
    |[Cemahabang, Indonesia]|7|0.0|0.0|0.0|0.0|8|
    |[Ногинский район, Россия]|7|0.0|0.0|0.0|0.0|9|
    |[Mungyeong-si, Republic of Korea]|7|0.0|0.0|0.0|0.0|10|
    |[Myanmar]|7|0.0|0.0|0.0|0.0|11|
    |[Kasaoka-shi, Okayama]|7|0.0|0.0|0.0|0.0|12|
    |[Rivers, Nigeria]|7|0.0|0.0|0.0|0.0|13|
    |[Шекино, Тульская область]|7|0.0|0.0|0.0|0.0|14|
    |[Rajangan, Indonesia]|7|0.0|0.0|0.0|0.0|15|
    |[광천교역점 버스정류장 BTS Bus Stop]|7|0.0|0.0|0.0|0.0|16|
    |[Carapicuíba, Brasil]|7|0.0|0.0|0.0|0.0|17|
    |[Нытвенский район, Россия]|7|0.0|0.0|0.0|0.0|18|
    |[Cilandak, Indonesia]|7|0.0|0.0|0.0|0.0|19|
    OUTLINE
    Ubuntu-20.04
```

## More in Appendix

21/02/13 morning Twitter Filtered Streams

Top 15

Twitter User Counts

Countries

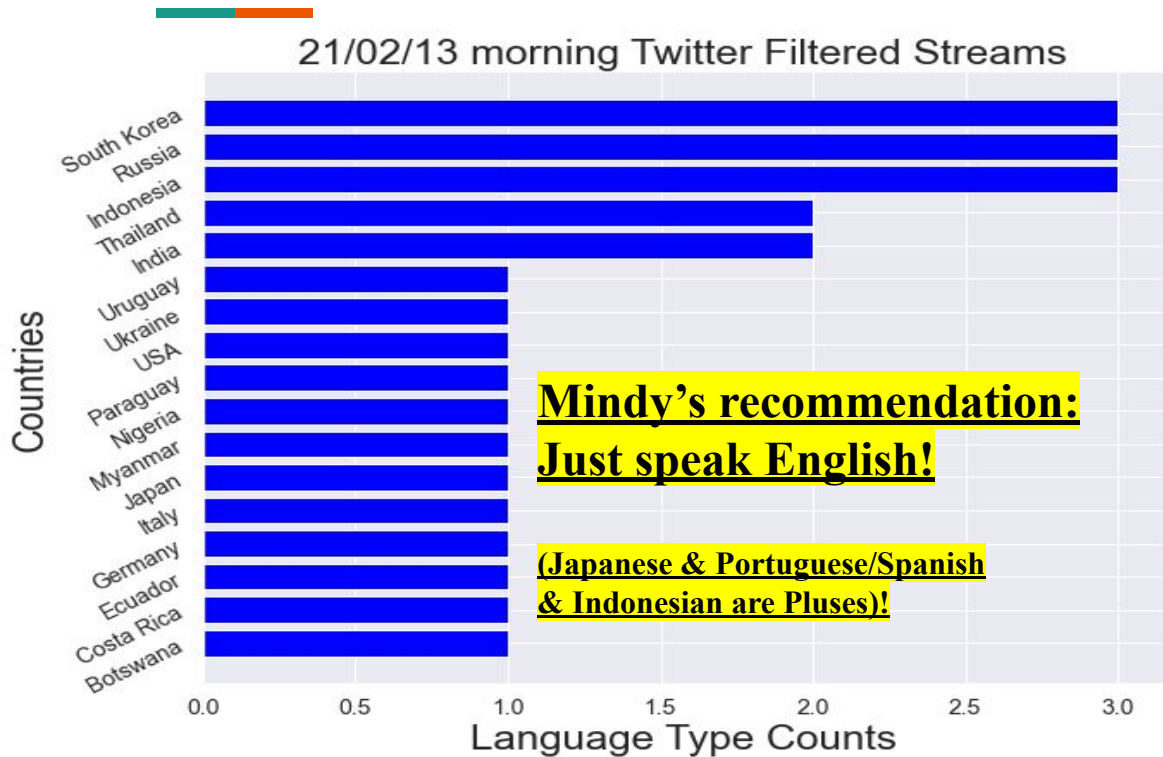
Country	Twitter User Counts (approx.)
Botswana	30
Russia	28
South Korea	27
USA	19
Indonesia	19
Japan	15
India	13
Thailand	13
Italy	13
Myanmar	7
Nigeria	7
Germany	5
Uruguay	5
Ecuador	5
Paraguay	5

2/12 all day, 2/13 afternoon, 2/14 afternoon  
2/15 morning, 2/15 afternoon, 2/15 night-1, night-2  
2/16 morning

**Mindy's recommendation: Italy (1), NYC (2), Brazil (3), Indonesia (4), Japan (5)**



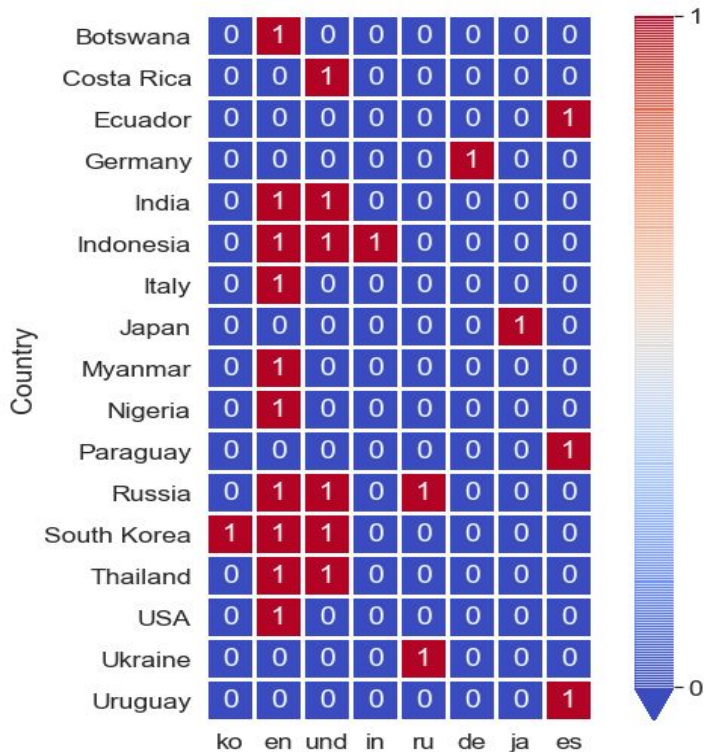
# Q2: What languages would you recommend BTS to speak in their global concert tours?



**Mindy's recommendation:**  
**Just speak English!**

**(Japanese & Portuguese/Spanish & Indonesian are Pluses)!**

More in Appendix





# Conclusion

## ❖ Clair

- People who tweet about #BTS commonly add members from the group to their tweet.
- People also tweet about other musical acts from south korea with #BTS
- The BTS tweets in english tend to be higher on average then their tweets in korean

## ❖ Yasmin

- Most popular words are Army, Favorite, and @BTS\_twt. This varies depending on the time of day but popular words show up at 10pm the most

## ❖ Jeff

- Ateez, Blackpink, Keeper, exo are kpop artists that are also popular around the world also
- Blackpink is trending in twitter more than BTS
- Texas tweets more about kpop than other states followed by California, New York and Illinois

## ❖ Melissa

- Languages: There is a large amount of language diversity among BTS tweets since 64 languages out of 70 supported languages tweeted about BTS. Some languages tweet more than others at certain times

## ❖ Mindy

- I would suggest BTS to hold concert tours in countries as follows: Italy, NYC, Brazil, Indonesia and Japan
- I would advise BTS to primarily speak English in their concert tours. However, it's better to also speak a bit of Japanese in Japan, Indonesian in Indonesia, Spanish & Portuguese in both Italy and Brazil in order to gain their popularities and visibilities in these countries.



Thank You / Q&A / GitHub

Github Link

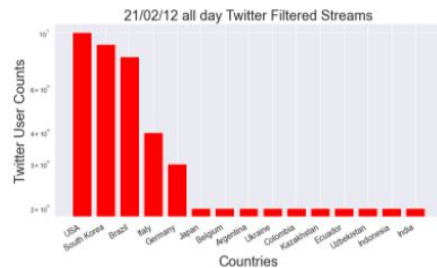
**<https://github.com/Jeffy892/project-2>**

**We're ready to move on to Project 3...**

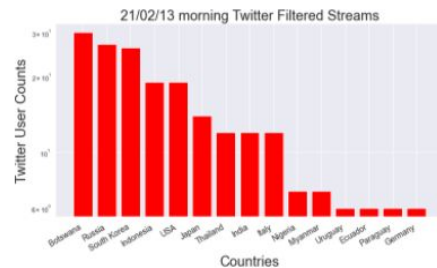
# Appendix

---

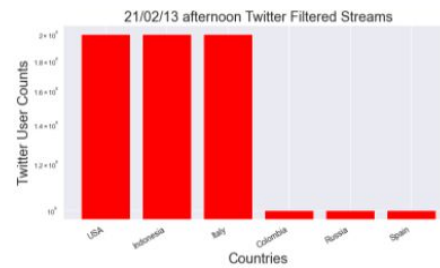
### 02/12 all day



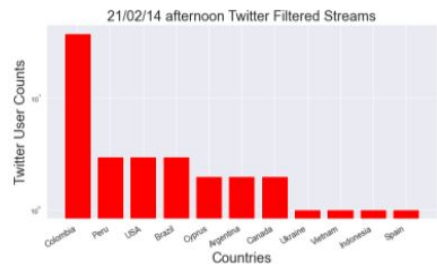
### 02/13 morning



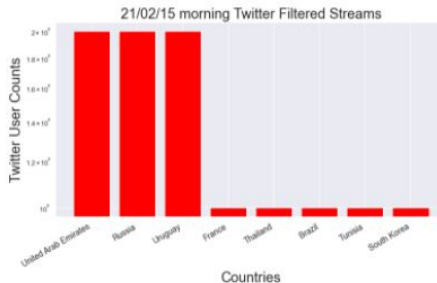
### 02/13 afternoon



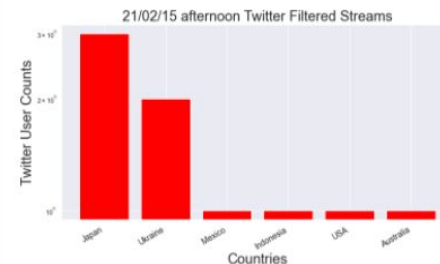
### 02/14 afternoon



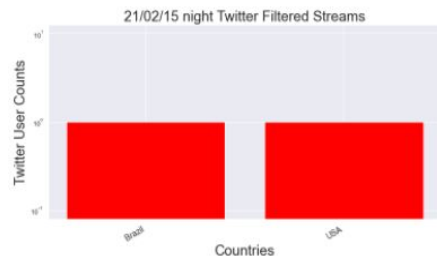
### 02/15 morning



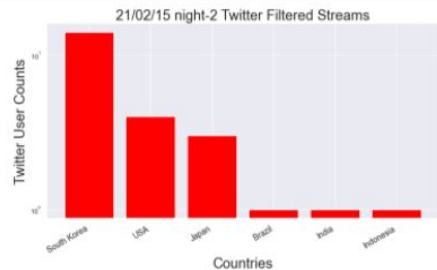
### 02/15 afternoon



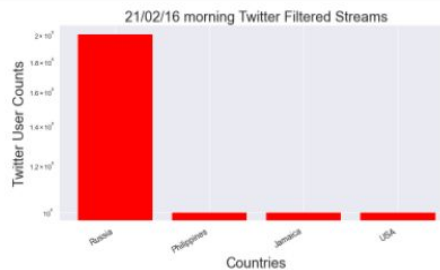
### 02/15 evening



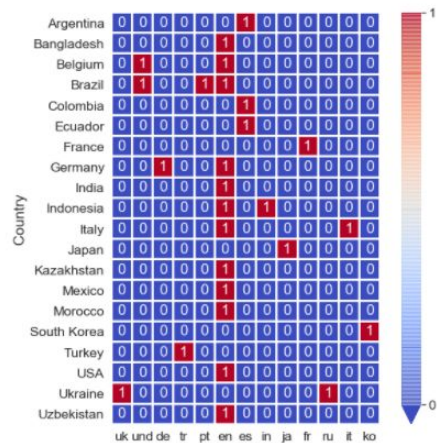
### 02/15 night



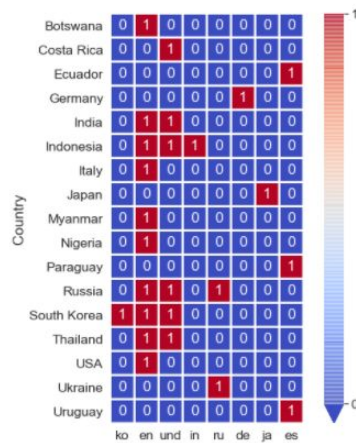
### 02/16 morning



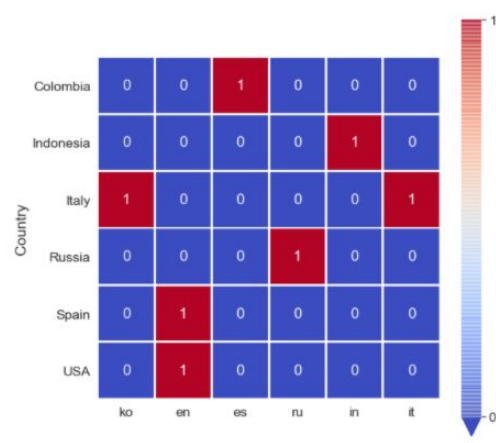
02/12 all day



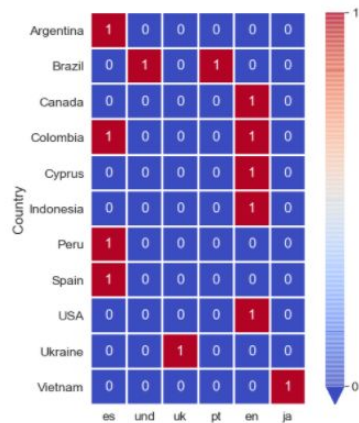
02/13 morning



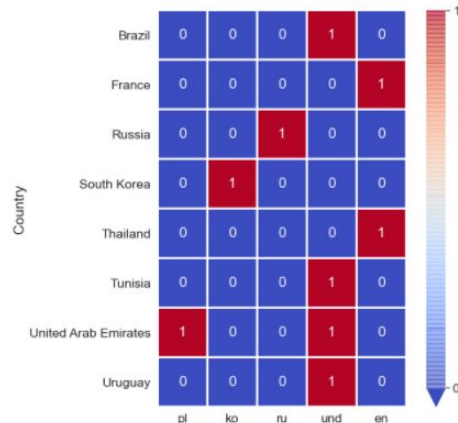
02/13 afternoon



02/14 afternoon



02/15 morning



02/15 afternoon

