

Reto: Análisis de datos para tweets durante la pandemia

Equipo 1:

- Alejandro Mariacca Santin
- Jorge Rodrigo Colín Rubio
- Pablo Ricardo Durán Sánchez
- Karen Paula Mayorga Guerrero
- José Antonio Pacheco Chargoy
- Andrés Tavera Mihailide

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re

data = pd.read_csv("covid19_tweets.csv")
data.head(5)
```

	user_name	user_location	user_description	user_created	user_followers	use
0	~i@t	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud # [] #Cavs ...	2019-03-07 01:45:06	197	
4	DIPR-J&K	Jammu and Kashmir	🖋 Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	

```
# Encontrar columnas del dataframe
df = data[['user_name', 'user_location', 'user_description', 'user_created', 'user_followers']]
print(df)
```

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date
0	Viola	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	950	18775	False	2020-07-25 12:27:21
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	1677	24	True	2020-07-25 12:27:17
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	9525	7254	False	2020-07-25 12:27:14
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[][] #Cavs ...	2019-03-07 01:45:06	197	987	1488	False	2020-07-25 12:27:10
4	DIPR-J&K	Jammu and Kashmir	🔪 Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	168	101	False	2020-07-25 12:27:08
...
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	Animal Scientist Muslim Real Madrid/Chelsea	2013-12-30 18:59:19	412	1609	1062	False	2020-08-29 19:44:21
179104	Jason	Ontario	When your cat has more baking soda than Ninja ...	2011-12-21 04:41:30	150	182	7295	False	2020-08-29 19:44:16
179105	BEEHEMOTH ⌚	🇨🇦 Canada	🔪 The Architects of Free Trade 🔪 Really Did ...	2016-07-13 17:21:59	1623	2160	98000		
179106	Gary DelPonte	New York City	Global UX UI Visual Designer. StoryTeller, Mus...	2009-10-27 17:43:13	1338	1111	0		
179107	TUKY II	Aliwal North, South Africa	TOKELO SEKHOPA TUKY II LAST BORN EISH TU...	2018-04-14 17:30:07	97	1697	566		

```

179105      False  2020-08-29 19:44:15
179106      False  2020-08-29 19:44:14
179107      False  2020-08-29 19:44:08

```

```

                                text \
0      If I smelled the scent of hand sanitizers toda...
1      Hey @Yankees @YankeesPR and @MLB - wouldn't it...
2      @diane3443 @wdunlap @realDonaldTrump Trump nev...
3      @brookbanktv The one gift #COVID19 has give me...
4      25 July • Media Bulletin on Novel #CoronaVirus

```

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	
user_followers	179108.0	109055.528184	841467.000703	0.0	172.0	992.0	5284.00	494425
user_friends	179108.0	2121.701566	9162.553072	0.0	148.0	542.0	1725.25	4973
user_favourites	179108.0	14444.105663	44522.698958	0.0	206.0	1791.0	9388.00	20471

####

Jorge Rodrigo Colín Rubio

Objetivo: Encontrar el máximo de número de personas en los países donde se utilizo el #Covid además del pais con mayor interacción con este hashtag,viendo desde que dispositivo se mando, si un iphone o un android

```
covid= dataframe.loc[dataframe.hashtags == "['COVID19']"]
covid
```

```

-----
NameError                                Traceback (most recent call last)
<ipython-input-1-e9704e085272> in <module>
----> 1 covid= dataframe.loc[dataframe.hashtags == "['COVID19']"]
      2 covid

NameError: name 'dataframe' is not defined

```

SEARCH STACK OVERFLOW

```
condition =(covid.source == "Twitter for Android") | (covid.source == "Twitter for i
```

```
new = covid.loc[condition]
new
```

```
local = new.user_location.dropna().value_counts()
```

```
condicion = local > 41

local

conteo = local.loc[condicion]

plt.boxplot(conteo)
plt.title('Diagrama de caja del número de personas en cada país en época de covid')
plt.xlabel('Países')
plt.ylabel('Personas')
print()
```

Alejandro Mariacca Santin

Encontrar características entre usuarios con más de 100,000 seguidores a partir de los tweets que escriben

A01654102

Objetivo

Este análisis se centrará en la búsqueda e identificación de características similares entre los tweets de los usuarios con más de 100 mil seguidores durante la época de la pandemia por Covid-19.

Visto con las métricas para metas smart, se puede visualizar de la siguiente forma:

- **Specific:** Encontrar las palabras en común que twitteen las personas con más de 100 mil seguidores y además, encontrar el promedio de seguidores que tienen las personas que twitteen sobre el covid.
- **Measurable:** Se medirá de acuerdo con la frecuencia que aparezcan las palabras.
- **Achievable:** La meta es alcanzable porque se disponen de los conocimientos y herramientas para llevarla a cabo.
- **Relevant:** Es relevante, porque a partir de ello se puede dividir la información en grupos y obtener insights que de otra forma no eran visibles.
- **Time bound:** Esta meta debe cumplirse para antes del 24 de marzo de 2023 a las 11 am con tiempo de la Ciudad de México.

Primer paso: Importar librerías y archivo csv

```
# Importar librerías
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from collections import Counter
import re
import nltk.corpus
# Descomentar si no está instalado
nltk.download('stopwords')
from nltk.corpus import stopwords
# Mi python anda loco y no quiere instalar wordcloud :(
#from wordcloud import WordCloud

# Cargar datos
data = pd.read_csv("covid19_tweets.csv")
# Obtener una breve visualización de los datos
data.head(5)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

	user_name	user_location	user_description	user_created	user_followers	use
0	~i@t	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	624	
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	2253	
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275	
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud # [] #Cavs ...	2019-03-07 01:45:06	197	
4	DIPR-J&K	Jammu and Kashmir	🖋 Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009	

```
# Describir información
data.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	
user_followers	74436.0	105951.312913	822289.985493	0.0	166.0	960.0	5148.00	138928
user_friends	74436.0	2154.721170	9365.587474	0.0	153.0	552.0	1780.25	4973
user_favourites	74436.0	15297.472030	46689.714291	0.0	220.0	1927.0	10148.00	20471

Segundo paso: Hacer dataframe de usuarios con más de 100,000 seguidores

```
# Limpiar la columna de texto para evitar caracteres extraños
def clean(s):
    s = s.replace(r'<lb>', "\n")
    s = re.sub(r'<br */*>', "\n", s)
    s = s.replace("&lt;", "<").replace("&gt;", ">").replace("&amp;", "&")
    s = s.replace("&quot;", """)
    s = re.sub(r'\>(*https*://[^\\])*\\*', "", s)
    s = re.sub(r'\\*', '', s)
    s = re.sub(r'_+', ' ', s)
    s = re.sub(r'"', "'", s)
    return str(s)

data['text_clean'] = ''
for i, row in data.iterrows():
    data.at[i, 'text_clean'] = clean(row.text)

# Depurar con stopwords
stop_words = stopwords.words('english')
data['text_relevant'] = data['text_clean'].apply(lambda x: ' '.join([word for word in

# Mostrar resultados
data[['text', 'text_clean', 'text_relevant']])
```

text


	text	text
0	If I smelled the scent of hand sanitizers toda...	If I smelled the scent of han
1	Hey @Yankees @YankeesPR and @MLB - wouldn't it...	Hey @Yankees @YankeesPR and @I
2	@diane3443 @wdunlap @realDonaldTrump Trump nev...	@diane3443 @wdunlap @realDonaldTrump
3	@brookbanktv The one gift #COVID19 has give me...	@brookbanktv The one gift #COVID
4	25 July : Media Bulletin on Novel #CoronaVirus...	25 July : Media Bulletin on Nov

```
# Obtener usuarios con el número deseado de followers
```

```
above_100k_followers = data[data['user_followers'] > 100000]
```

```
# Limpiar columnas que no son relevantes
```

```
above_100k_followers.drop(['user_name', 'user_created'], axis=1)
```

	user_location	user_description	user_followers	user_friends	user_verified
4	Jammu and Kashmir	 Official Twitter handle of Department of Inf...	101009	168	
25	Mumbai, India	Focused on matching blood donors with those in...	1215920	2047	
60	NaN	Official Government of India updates on #COVID...	100214	70	

Tercer paso: Analizar sus tweets y encontrar palabras

En esta sección se analizarán las variables "text_relevant" y "hashtags".

```

--               development organization...
# Obtener los hashtags más comunes
hashtags = Counter(above_100k_followers['hashtags'])

# Pasarlos a un dataframe y organizarlos
hashtagsDF = pd.DataFrame.from_dict(hashtags, orient='index').reset_index()
hashtagsDF.columns = ['Hashtags', 'Count']
hashtagsDF.Count = hashtagsDF.Count.astype(int)
sortedHashtags = hashtagsDF.sort_values('Count', ascending=False).head(20)

# Imprimir
sortedHashtags

```


	Hashtags	Count
3	['COVID19']	1277
8	NaN	806
4	['Covid19']	178
71	['coronavirus']	35
26	['coronavirus', 'COVID19']	32
0	['CoronaVirusUpdates', 'COVID19']	29
33	['covid19']	27
75	['IndiaFightsCorona', 'COVID19']	26
112	['Hyderabad', 'Blood']	18

Obtener las palabras más comunes en los tweets

```
texts = Counter(" ".join(above_100k_followers["text_relevant"]).split()).most_common(1
```

Pasarlos a un dataframe y organizarlos

```
wordsDF = pd.DataFrame(texts)
```

```
wordsDF.columns = ['Words', 'Count']
```

```
wordsDF.Count = wordsDF.Count.astype(int)
```

```
sortedWords = wordsDF.sort_values('Count', ascending=False).head(20)
```

Imprimir

```
sortedWords
```

	Words	Count
0	#COVID19	1909
1	cases	584
2	The	367
3	new	338
4	:	326
5	#Covid19	300
6	&	228

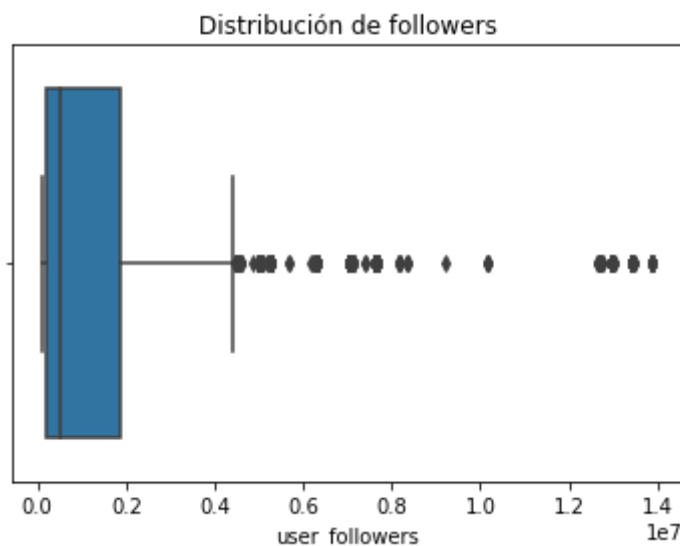
Cuarto paso: Analizar resultados

A continuación se presentarán gráficas así como mi interpretación de las mismas.

Primero comienzo con la distribución de usuarios y su número de seguidores. Como se puede apreciar en la gráfica de abajo, el 75% de los usuarios con más de 100 mil seguidores no pasa de los 2 millones de seguidores. En promedio, tienen poco menos de 1 millón, aunque eso ya se sabía. Lo que es realmente nuevo es ver la cantidad de outliers. Se puede ver que a partir de los 4 millones y medio de seguidores aproximadamente, las cuentas simplemente son atípicas y ya no forman parte de la caja o los bigotes.

```
sns.boxplot(x=above_100k_followers['user_followers']).set_title("Distribución de follc
```

```
Text(0.5, 1.0, 'Distribución de followers')
```



En el siguiente histograma se puede observar el número de veces que fue usado cada hashtag. Al ser en un periodo de COVID, es entendible el uso de la palabra. Al observar la lista de hashtags, es notable la presencia del gobierno e instituciones indias. De este histograma se puede concluir que las personas con más de 100 mil seguidores twitteraon sobre temas relacionados con el covid. Se podría intuir que usaban su poder para crear consciencia o que buscaban aumentar sus seguidores al twittear sobre el tema dle momento. Es interesante también notar que se puede deducir el estado de la pandemia al revisar los hashtags "SecondPeak" y "SecondWave".

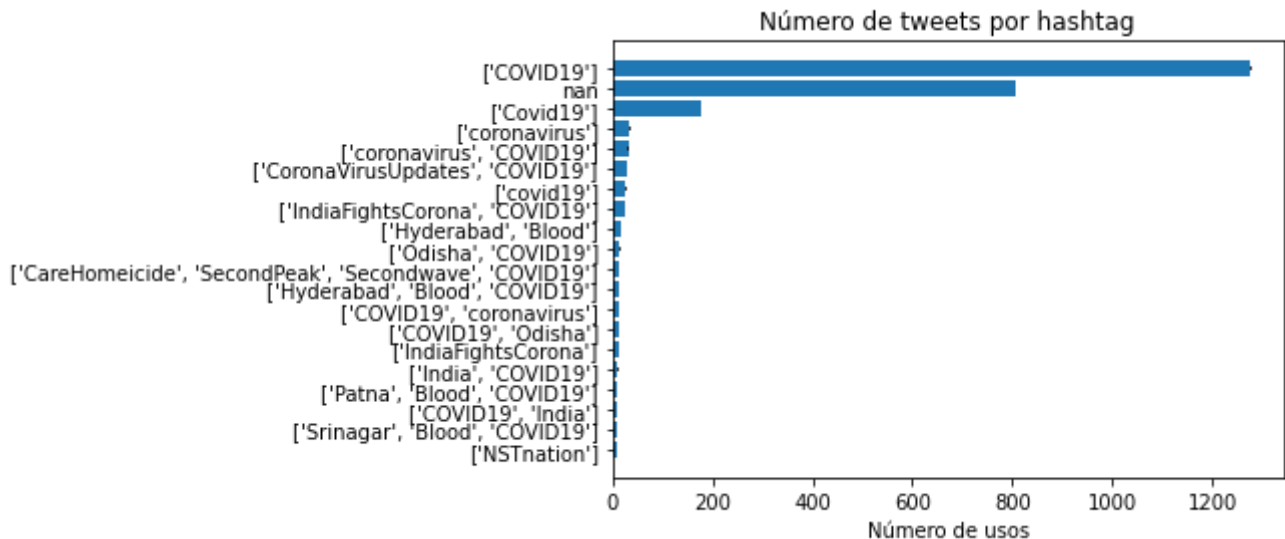
```
# Hacer plot de la gráfica

fig, ax = plt.subplots()

people = sortedHashtags['Hashtags']
y_pos = np.arange(len(people))
performance = sortedHashtags['Count'].to_numpy()
error = np.random.rand(len(people))

ax.barh(y_pos, performance, xerr=error, align='center')
ax.set_yticks(y_pos, labels=people)
ax.invert_yaxis() # labels read top-to-bottom
ax.set_xlabel('Número de usos')
ax.set_title('Número de tweets por hashtag')

plt.show()
```



También se hizo un histograma sobre las 20 palabras más comunes en los tweets de los usuarios. Aquí también se puede ver la congruencia con la gráfica anterior y las suposiciones hechas. Aquí también hace una aparición India. Las palabras menos usadas tiene que ver con información como es el caso de "tested", "reported", "total". Y las más usadas se enfocan en el nombre del virus y los

nuevos casos. A partir de esto, es posible sugerir que los usuarios seleccionados enfocaban su atención en el incremento de casos.

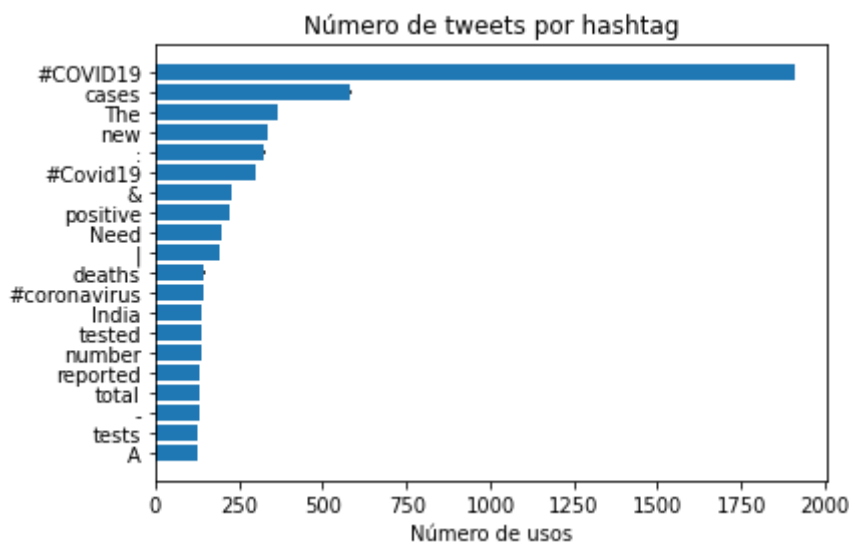
```
# Hacer plot de la gráfica

fig, ax = plt.subplots()

people = sortedWords['Words']
y_pos = np.arange(len(people))
performance = sortedWords['Count'].to_numpy()
error = np.random.rand(len(people))

ax.barh(y_pos, performance, xerr=error, align='center')
ax.set_yticks(y_pos, labels=people)
ax.invert_yaxis() # labels read top-to-bottom
ax.set_xlabel('Número de usos')
ax.set_title('Número de tweets por hashtag')

plt.show()
```



Quinto paso: Conclusiones

A pesar de no ser un análisis extensivo, se puede notar que los usuarios con más de 100 mil seguidores comparten algunas características. Entre ellas, se puede identificar que a pesar de ser cuentas con muchos seguidores, no llegan a ser colosales. También se debe mencionar su uso similar de hashtags y palabras para comunicar sus ideas a través de tweets. El análisis de caja y bigotes se usó para encontrar la distribución de usuarios. Un heatmap podría ayudar a confirmar algunas de las suposiciones al comparar usuarios y mostrar las zonas de convergencia, sin embargo, no fue empleado por falta de tiempo. Incluso una nube de palabras o un análisis de sentimientos hubiera arrojado datos mucho más decisivos sobre el comportamiento de estos usuarios. Es

totalmente posible incluir un análisis de K-means, pues este hubiera revelado datos que son visibles a simple vista. K-means hubiera aportado una nueva perspectiva para saber qué tanto se parecen los usuarios entre sí y qué tan poco se parecen. Al final, es posible decir que sí hay características entre los usuarios seleccionados.

Double-click (or enter) to edit

Andrés Tavera

La base de datos con la que trabajaremos muestra los distintos tweets que se hicieron en el año 2019, durante la pandemia, y muestran algunos datos del usuario.

Objetivo

Se busca saber las carecteristicas en común que tienen los usuarios con más de 500,000 seguidores durante la pandemia

```
#Para eso debemos de filtrar la base de datos y trabajar solo con los que presenten má  
df_nfamosos=df_famosos= df[df["user_followers"]<500000]  
df_famosos= df[df["user_followers"]>500000]  
df_famosos
```



	user_name	user_location	user_description	user_created	user_follower
25	Blood Donors India	Mumbai, India	Focused on matching blood donors with those in...	2008-12-23 07:55:39	121592
77	IMF	Washington, DC	Breaking news and alerts from the Internationa...	2009-03-18 16:13:51	169552
78	Oxfam International	NaN	Oxfam is a world-wide development organization...	2007-12-31 18:27:35	84367
87	New Straits Times	Malaysia	News, views and up-to-date reports from Malays...	2009-07-09 09:04:01	71634
120	Livemint	India	Breaking news and analyses of Indian and world...	2008-11-27 09:07:38	190288
...
178823	Cognizant	Global (HQ: Teaneck, NJ)	Cognizant (Nasdaq: CTSI) is dedicated	2009-08-28 20:31:48	76145

```
df_famosos.describe().transpose()
```

	count	mean	std	min	25%	50%	75%
user_followers	5196.0	3.379612e+06	3.651779e+06	500070.0	761451.00	1878954.0	5007492
user_friends	5196.0	3.190647e+03	2.822635e+04	0.0	126.00	335.0	1070
user_favourites	5196.0	4.185503e+03	1.070050e+04	0.0	143.75	874.0	2960

Después de filtrar la base de datos, debemos de hacernos ciertas preguntas. Las que se responderan son las siguientes

- ¿Cuántos tweets hace en promedio una persona famosa?
- ¿Están verificados?
- ¿Qué relación existe entre los seguidores y a los likes que dan?

```
tweets_per_username = df_famosos.groupby("user_name")
ntweets_per_username = df_nfamosos.groupby("user_name")
```

#Promedio de tweets por persona

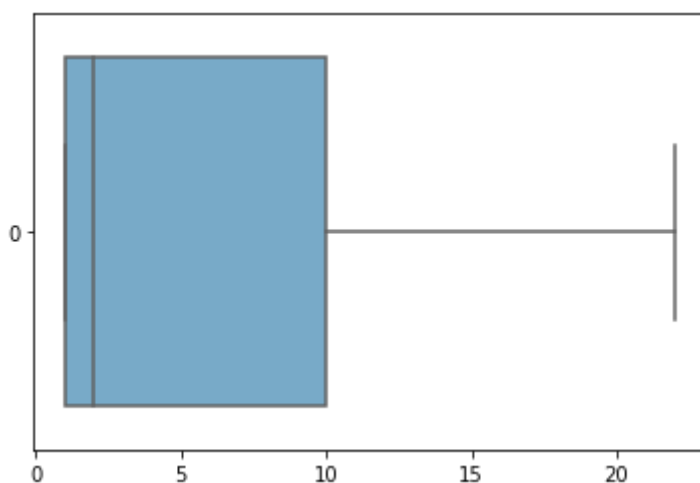
```
df_famososbxplt=tweets_per_username["text"].count()
```

```
df_nfamososbxplt=ntweets_per_username["text"].count()
```

```
df_famososbxplt.mean()
```

```
14.761363636363637
```

```
sns.boxplot(data=df_famososbxplt, orient="h", palette='Blues', showfliers=False)
plt.show()
```



```
df_famososbxplt.median()
```

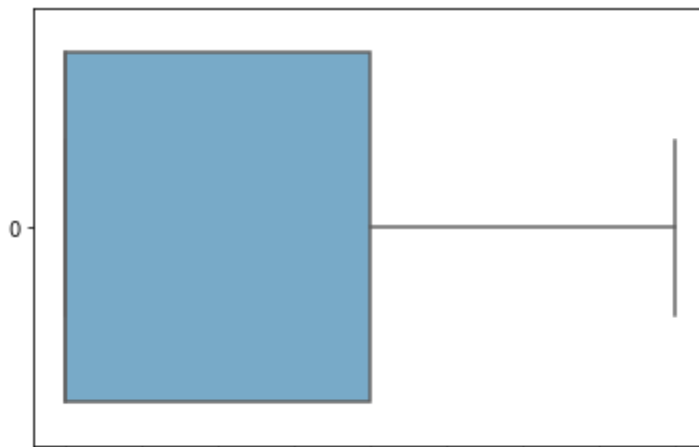
```
2.0
```

```
print(df_famososbxplt.max())
print(df_famososbxplt.min())
```

```
print(df_nfamososbxplt.median())
print(df_nfamososbxplt.max())
print(df_nfamososbxplt.min())
print(df_nfamososbxplt.mean())
```

```
1.0
679
1
1.891725495741464
```

```
sns.boxplot(data=df_nfamososbxplt, orient="h", palette='Blues', showfliers=False)
plt.show()
```



Aquí podemos apreciar como los usuarios no famosos varian, hay la mediana es de un tweet y su maxima alcanza los 679 twwets. Aquí podemos encontrar tanto gente muy activa como muy poco activa. Aún así la media por usuario es de 1 tweet, es decir, son muy poco activos en su mayoría

```
df_famosos.groupby("user_verified").count()
```

	user_name	user_location	user_description	user_created	user_fol
user_verified					
False	59	55	59	59	
True	5137	4813	5135	5137	

Podemos ver que en efecto estos usuarios, en su mayoría tienen la verificación. Esto se sospechaba desde un inicio, pero con lo que descubrimos previamente sobre la cantidad de tweets, teníamos que invesitgar.

```
df_nfamosos.groupby("user_verified").count()
```

	user_name	user_location	user_description	user_created	user_fol
user_verified					
False	155954	121091	145679	155954	
True	17958	16378	17949	17958	

En cuanto a los no famosos, podemos observar que en su mayoría no estan verificados. Caolo 18,000 de los casi 170,000 tweets fuero hechos por alguien verificado.

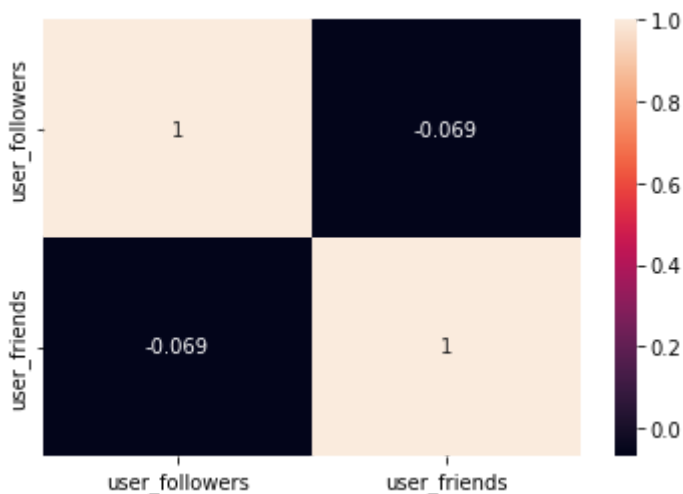
```
print(df_famosos.iloc[:,[4,5]].corr())
```

```

                user_followers  user_friends
user_followers      1.000000    -0.068686
user_friends       -0.068686     1.000000

```

```
sns.heatmap(df_famosos.iloc[:,[4,5]].corr(), annot=True)
plt.show()
```



Podemos ver que no hay ninguna similitud entre estos dos. Es decir no hay nada que parezca indicar que uno afecta a la otra.

Hasta el momento no hemos visto nada que concluya que características tiene alguien famoso.

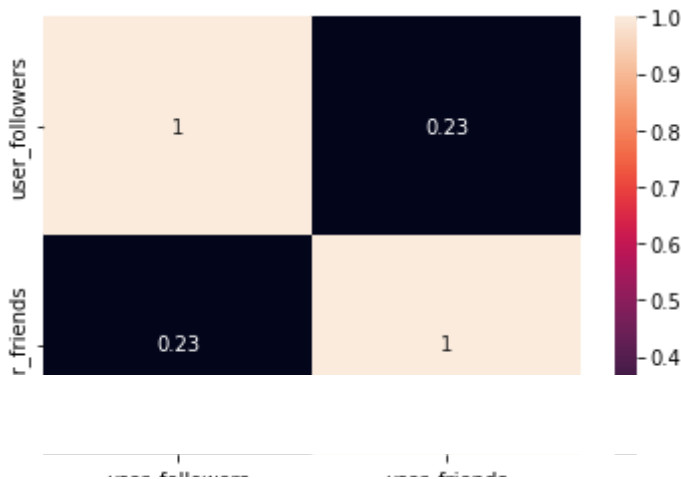
```
print(df_nfamosos.iloc[:,[4,5]].corr())
```

```

                user_followers  user_friends
user_followers      1.000000     0.228873
user_friends        0.228873     1.000000

```

```
sns.heatmap(df_nfamosos.iloc[:,[4,5]].corr(), annot=True)
plt.show()
```



Este si parece tener un poco más de correlación, es decir que en este rango si afecta un poco más la actividad que tengas en la aplicación.

Conclusión

Pudimos observar que en su mayoría, los famosos son un poco más activos en twitter en cuanto a publicaciones. Nos dimos cuenta, que en esta epoca de Covid, la mejor oportunidad, que tampoco garantiza el exito, es ser activo en cuanto a likes. Entre más activo seas, más probabilidad tienes de ser exitoso, pero al ya ser famoso, los likes que des no te garantizan más exito.

Karen Paula Mayorga Guerrero

Objetivo: Determinar el impacto en variables user followers, user friends y user favorites en inicios y finales de las bases de datos para los usuarios de twitter por medio de diferentes análisis a través de la primera base y última base de datos del año 2020 gracias a las herramientas proporcionadas por la materia esta semana se logrará llevar a cabo el análisis. Y, lo anterior se logrará trabajando en un lapso de 8 horas el día 23 de marzo.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#inicio de la documentacion de tweets
data1 = pd.read_csv("covid19_tweets_inicio.csv", index_col = 'date')
data1.head(5)
```

	user_name	user_location	user_description	user_created	user_follower
date					
2020-07-25 12:27:21	Viñe	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	62
2020-07-25 12:27:17	Tom Basile	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	225
2020-07-25 12:27:14	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	927
2020-07-25 12:27:10	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[] #Cavs ...	2019-03-07 01:45:06	19
2020-07-25	DIPR-J&K	Jammu and Kashmir	Official Twitter handle of	2017-02-12 06:45:15	10100

#final de la documentacion de tweets

```
data2 = pd.read_csv("covid19_tweets_final.csv", index_col = 'date')
data2.head(5)
```

#Se crea una data en base a la dara inicial y la data final

```
superData=pd.merge(data1,data2,left_on="user_name", right_on="user_name")
superData.head(5)
```

#Gráfica para comparar en las bases de datos las variables user_followers, user_friends

```
Datos1=data1.iloc[:,[4,8]]
print(Datos1.plot())
```

```
Datos2=data2.iloc[:,[4,8]]
print(Datos2.plot())
```

```
Datos3=data1.iloc[:,[5,8]]
print(Datos3.plot())
Datos4=data2.iloc[:,[5,8]]
print(Datos4.plot())
```

```
Datos5=data1.iloc[:,[6,8]]
print(Datos5.plot())
Datos6=data2.iloc[:,[6,8]]
print(Datos6.plot())
```

```
#Análisis comparativo para las 3 variables
```

```
print(data1['user_followers'].describe())
print("\n")
print(data2['user_followers'].describe())
print("\n")
```

```
print(data1['user_friends'].describe())
print("\n")
print(data2['user_friends'].describe())
print("\n")
```

```
print(data1['user_favourites'].describe())
print("\n")
print(data2['user_favourites'].describe())
print("\n")
```

```
Datos11=superData.iloc[:,[4,15]]
print(Datos11.corr())
print("\n")
```

```
Datos12=superData.iloc[:,[4,16]]
print(Datos12.corr())
print("\n")
```

```
Datos13=superData.iloc[:,[4,17]]
print(Datos13.corr())
print("\n")
```

```
sns.heatmap(superData.iloc[:,[4,15]].corr(), annot=True)
plt.show()
```

```
sns.heatmap(superData.iloc[:,[4,16]].corr(), annot=True)
plt.show()
```

```
sns.heatmap(superData.iloc[:,[4,17]].corr(), annot=True)
plt.show()
```

Double-click (or enter) to edit

Double-click (or enter) to edit

José Antonio Pacheco

Objetivo SMART: Determinar la evolución de los temas de interés en Twitter (mediante hashtags de los tweets de los usuarios) a lo largo del tiempo durante la pandemia de Covid-19.

- Específico: Utilizar los hashtags de cada tweet de la base de datos para determinar cuáles son los más usados durante un tiempo determinado de la pandemia de Covid19.
- Medible: Se determinarán los hashtags o temas más populares cuando superen un valor determinado de en cuántos tweets aparezca ese hashtag.
- Alcanzable: Se emplearán herramientas para el análisis de datos aprendidos en el curso para poder determinar la popularidad de los temas en Twitter en un contexto determinado.
- Relevante: Se busca aplicar el conocimiento adquirido en el curso y poner a prueba el criterio para interpretar los datos, y así alcanzar la subcompetencia "SING0202A ", que busca determinar la interacción existente entre dos variables.
- A tiempo: Se estima que la resolución de este reto tomará menos de 8 horas en total. Lo cual es tiempo suficiente para presentarlos a las 12:00pm del dÃa siguiente.

Para determinar los temas más relevantes en Twitter durante la pandemia de Covid19, tomaremos como estándar más de 1000 tweets dentro de toda la base de datos en los que cada tema haya sido mencionado en hashtags. No tomaremos en cuenta aquellos temas que no hayan sido mencionados en hashtags en menos de 1000 tweets en nuestra base de datos, ya que esta presenta un total de 74436 tweets, y aquellos temas presentes en menos de 1000 tweets son irrelevantes.

```
# Obtenemos todos los tweets relacionados con el Covid19 = 42673
```

```
data = data.fillna("")  
dataframecovid = data[data["hashtags"].str.contains("COVID19") | data["hashtags"].str.  
dataframecovid
```

	user_name	user_location	user_description	user_created	user_follower
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	927
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud # [] #Cavs ...	2019-03-07 01:45:06	19
4	DIPR-J&K	Jammu and Kashmir	🔪 Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	10100
5	🎹 Franz Schubert	Новороссия	🎵 #Новороссия #Novorossiya #оставайсядома #S...	2018-03-19 16:29:52	118
6	hr bartender	Gainesville, FL	Workplace tips and advice served up in a frien...	2008-08-12 18:19:49	7995
...
24613	Srikanth	Away from Hypocrisy		2010-12-13 13:53:12	523
24615	Hindustan Times	India	One of India's largest media companies. Latest...	2009-04-29 10:11:34	764961
24616	sam mostyn	sydney, australia	#sustainability #climatechange #genderequality...	2009-04-07 14:14:09	1437
24617	Caspian 🐱🐱	Dragons Lair	I'm a fire-breathing DRAGON cat! 🔥 Never know ...	2019-01-14 03:28:39	34

```
# Obtenemos todos los tweets relacionados con los cubrebocas/mascarillas = 1414
```

```
data = data.fillna("")
dataframemask = data[data["hashtags"].str.contains("facemask") | data["hashtags"].str.
dataframemask
```

	user_name	user_location	user_description	user_created	user_followers
71	BestValueButtons	Denver, CO	We are a fabric outlet and specialty button Et...	2018-05-30 19:45:27	121
103	Thomas Faires	Toronto, Ontario, Canada	...yup, still a husband, vegetarian, computer ...	2011-04-14 12:36:09	73
147	Illustrated London		Buy a print for your wall and look out onto an...	2020-07-25 12:04:11	
169	Christine Murphy Estes, MM, MA, CCC-SLP	New York, NY	Speech-Language Pathologist, Voice Specialist,...	2018-10-07 02:31:06	63
212	coley	Grays, England	26, Essex UK. bearded. long haired. engaged. p...	2015-12-11 22:20:21	1
...
24449	America		"The world will not be destroyed by those who ...	2014-05-30 12:14:01	73
24455	AJ 🗨️ 🌸 🍏 🇺🇸	East Coast, USA	👋 I'm a Corona Virus killer!.... Click Follow!...	2019-08-21 22:32:59	73
24493	why_i_march	United States	An Ind who believes all ppl should be treated ...	2017-02-10 00:30:47	190
24508	Latinos for America	Austin, TX		2016-03-11 07:46:57	10
24604	lellekrafts	Nairobi, Kenya	Sun seeking, woke adventurer, spreading knowle...	2018-07-02 05:47:36	1

```
# Obtenemos todos los tweets relacionados con Donald Trump = 1215
```

```
data = data.fillna("")
dataframetrump = data[data["hashtags"].str.contains("TRUMP") | data["hashtags"].str.contains("donald trump")]
dataframetrump
```

	user_name	user_location	user_description	user_created
103	Thomas Faires	Toronto, Ontario, Canada	...yup, still a husband, vegetarian, computer ...	2011-04-14 12:36:09
234	BeautifulSkinYourIn		Daily Tips from our Nurses, Dietitians, Aesthe...	2016-11-17 14:30:31
412	marklaverdure	Los Angeles & Billings	I have passion for living life to it's fullest...	2011-03-04 15:17:55
437	Mx. Joe~Anthony Sierra Let us pray #COVID19 away	#StGeorge~#StatenIsland~#NYC	Hope to be insightful	2013-05-03 03:46:03
448	marklaverdure	Los Angeles & Billings	I have passion for living life to it's fullest...	2011-03-04 15:17:55
...
24486	bennyvids		21k on Tik Tok @bennyvids I Follow me for qual...	2020-07-09 20:20:14
24492	Robert J - Wear a Damn Mask	Atlanta, GA	Artist, writer, pianist, humanitarian, progres...	2013-01-23 19:52:38
24495	Maher Marzouk	Suffern, NY	CEO at Intech Investment.	2010-07-07 14:22:33
24506	Terence L. Mathious	Saginaw, MI Metro-ATL, GA	BVHS-Saginaw Alumni; Xavier of Louisiana Alumn...	2011-09-20 22:36:34
24520	Scott	Oregon, USA	Husband, Father, Technologist, Executive, Entr...	2012-11-07 04:51:16

313 rows x 13 columns

```
# Creamos un dataframe solamente para ver en cuántos tweets no se usaron hashtags = 21
dataframenot = data.loc[(data["hashtags"]=="")]
dataframenot
```


	user_name	user_location	user_description	user_created	user_follower
0	Viñeta	astroworld	wednesday addams as a disney princess keepin i...	2017-05-26 05:46:42	62
1	Tom Basile 🇺🇸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2009-04-16 20:06:23	225
7	Derbyshire LPC			2012-02-03 18:08:10	60
10	Voice Of CBSE Students			2020-07-14 17:50:30	
14	DailyaddaaNews	New Delhi	Breaking news alerts from India.	2016-10-22 09:18:42	54
...
24619	Chrissea Tea	Greater Vancouver, British Columbia	Hello 😊 Pics with #iPhone #treelover 🍀🍁🌲🌲🌲 #fl...	2010-12-28 07:01:05	100
24621	Derik Van Derbeken	Los Angeles, CA	Photog, VO, Foodie, Amateur Sleuth, 💖 - Rusty G d...	2009-03-12 23:01:59	40
24622	_____	Cincinnati, OH	Big heart, sharp tongue. ...	2009-03-01 04:17:32	15
24623	Shannon	Ohio, USA	Capricorn. 24.	2016-12-04 23:54:12	1
24624	RIVIERA BEACH FIRE RESCUE	600 West Blue Heron Blvd, Rivi	Official information from Riviera Beach Fire R...	2015-03-31 13:13:24	130

7024 rows x 13 columns

Obtenemos todos los tweets relacionados con otros temas = 7700

```
len(data.iloc[:, 8]) - len(dataframenot.iloc[:, 8]) - len(dataframecovid.iloc[:, 8]) -  
  
2590
```

La base de datos con la que estamos trabajando solamente tiene datos de tweets escritos desde el 25 de julio de 2020 hasta el 4 de agosto de ese año, los cuales son solamente 11 días.

Podemos obtener las menciones de cada tema relevante elegido por hashtags por cada uno de estos días.

```
# Obtenemos los tweets sobre Covid19 escritos el 25 de julio de 2022  
  
dataframecovid_25jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-25")]  
dataframecovid_25jul
```

	user_name	user_location	user_description	user_created	user_followers
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	2009-02-28 18:57:41	9275
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud # [] #Cavs ...	2019-03-07 01:45:06	197
4	DIPR-J&K	Jammu and Kashmir	Official Twitter handle of Department of Inf...	2017-02-12 06:45:15	101009
	Franz		#Новороссия #Novorossiya	2018-03-10	

Obtenemos los tweets sobre Covid19 escritos el 26 de julio de 2022

```
dataframecovid_26jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-26")]
dataframecovid_26jul
```

	user_name	user_location	user_description	user_created	user_fol
17176	Mr. Toro Güero, #WearAMask #HighTransmissibility	Blocked by @RealJamesWoods	*Blunt* #Agave #advocate #diffusist #tequila #...	2010-12-08 06:21:51	
17179	Afrobodies	South Africa	South African recombinant alpaca antibody prod...	2015-08-20 19:10:38	
17181	ChinAfrica Magazine - South Africa		China's only monthly magazine on China- Africa ...	2019-11-04 09:22:50	

Obtenemos los tweets sobre Covid19 escritos el 27 de julio de 2022

```
dataframecovid_27jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-27")]
dataframecovid_27jul
```

	user_name	user_location	user_description	user_created	user_followers	user_
17185	DanKlink1892	USA	USA MAGA Free	2017-04-10		

Obtenemos los tweets sobre Covid19 escritos el 28 de julio de 2022

```
dataframecovid_28jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-28")]
dataframecovid_28jul
```

	user_name	user_location	user_description	user_created	user_followers	user_
17185	DanKlink1892	USA	USA MAGA Free	2017-04-10		

Obtenemos los tweets sobre Covid19 escritos el 29 de julio de 2022

```
dataframecovid_29jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-29")]
dataframecovid_29jul
```

	user_name	user_location	user_description	user_created	user_followers	user_
--	-----------	---------------	------------------	--------------	----------------	-------

Obtenemos los tweets sobre Covid19 escritos el 30 de julio de 2022

```
dataframecovid_30jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-30")]
dataframecovid_30jul
```

user_name	user_location	user_description	user_created	user_followers	user_id
24617	Corona 🇲🇽 🇲🇽	Dragonair	DRAGONair 🇲🇽	2019-01-14	

```
# Obtenemos los tweets sobre Covid19 escritos el 31 de julio de 2022
```

```
dataframecovid_31jul = dataframecovid[dataframecovid["date"].str.contains("2020-07-31")]
dataframecovid_31jul
```

user_name	user_location	user_description	user_created	user_followers	user_id
-----------	---------------	------------------	--------------	----------------	---------

```
# Obtenemos los tweets sobre Covid19 escritos el 01 de agosto de 2020
```

```
dataframecovid_01ago = dataframecovid[dataframecovid["date"].str.contains("2020-08-01")]
dataframecovid_01ago
```

user_name	user_location	user_description	user_created	user_followers	user_id
-----------	---------------	------------------	--------------	----------------	---------

```
# Obtenemos los tweets sobre Covid19 escritos el 02 de agosto de 2020
```

```
dataframecovid_02ago = dataframecovid[dataframecovid["date"].str.contains("2020-08-02")]
dataframecovid_02ago
```

user_name	user_location	user_description	user_created	user_followers	user_id
-----------	---------------	------------------	--------------	----------------	---------

```
# Obtenemos los tweets sobre Covid19 escritos el 03 de agosto de 2020
```

```
dataframecovid_03ago = dataframecovid[dataframecovid["date"].str.contains("2020-08-03")]
dataframecovid_03ago
```

user_name	user_location	user_description	user_created	user_followers	user_id
-----------	---------------	------------------	--------------	----------------	---------

```
# Obtenemos los tweets sobre Covid19 escritos el 04 de agosto de 2020
```

```
dataframecovid_04ago = dataframecovid[dataframecovid["date"].str.contains("2020-08-04")]
dataframecovid_04ago
```

user_name	user_location	user_description	user_created	user_followers	user_id
-----------	---------------	------------------	--------------	----------------	---------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 25 de julio de 2022
```

```
dataframemask_25jul = dataframemask[dataframemask["date"].str.contains("2020-07-25")]
dataframemask_25jul
```

	user_name	user_location	user_description	user_created	user_follow
71	BestValueButtons	Denver, CO	We are a fabric outlet and specialty button Et...	2018-05-30 19:45:27	1
103	Thomas Faires	Toronto, Ontario, Canada	...yup, still a husband, vegetarian, computer ...	2011-04-14 12:36:09	
147	Illustrated London		Buy a print for your wall and look out onto an...	2020-07-25 12:04:11	
169	Christine Murphy Estes, MM, MA, CCC-SLP	New York, NY	Speech-Language Pathologist, Voice Specialist,...	2018-10-07 02:31:06	
212	coley	Grays, England	26, Essex UK. bearded. long haired. engaged. p...	2015-12-11 22:20:21	
...	
16745	John Earl Burnett	Los Angeles	Asset Management/Economic Consulting - Film, T...	2009-05-25 18:26:29	3
16809	Christina Headrick	Arlington, Virginia	writer/designer/creative & other job is mom; 3...	2009-08-14 03:01:17	
16813	Jo Anna Van Thuyne 💖💖	New York, NY	Actor. Comedian. Filmmaker. Podcaster. Host of...	2009-04-18 17:59:15	10
16831	PROMOrx	USA	Since 2000, we've been prescribing branded pro...	2009-03-10 23:31:31	-
16843	Arab News	Saudi Arabia	Established in 1975, Arab News is the Middle E...	2009-08-27 02:15:51	285

345 rows x 13 columns

Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 26 de julio de 2022

```
dataframemask_26jul = dataframemask[dataframemask["date"].str.contains("2020-07-26")]
dataframemask_26jul
```

	user_name	user_location	user_description	user_created	user_followers
17195	Wild Goose #TNCM		Veteran market strategist. Cantillon not Law! ...	2014-12-03 18:15:49	1892
17299	Demelza Klass ❤️	United Kingdom	Changing the world one person at a time. #chan...	2017-06-11 18:07:57	834
17333	Gifts Consultant		Hand-picked #gifts for him, her, kids, home, o...	2013-06-12 11:31:29	53
17366	Patrick Henningesen	USA	Independent global affairs analyst, journalist...	2008-10-23 22:03:55	47896
17423	Deep Mahajan		Vice President- MNC	2020-03-22 15:14:21	42
...
24449	America		"The world will not be destroyed by those who ...	2014-05-30 12:14:01	788
24455	AJ 🇺🇸 🌸 🍏	East Coast, USA	👋 I'm a Corona Virus killer!.... Click Follow!...	2019-08-21 22:32:59	785
24493	why_i_march	United States	An Ind who believes all ppl should be treated ...	2017-02-10 00:30:47	1902
24508	Latinos for America	Austin, TX		2016-03-11 07:46:57	160
24604	lellekrafts	Nairobi, Kenya	Sun seeking, woke adventurer, spreading knowle...	2018-07-02 05:47:36	39

150 rows × 13 columns

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 27 de julio de 2022
```

```
dataframemask_27jul = dataframemask[dataframemask["date"].str.contains("2020-07-27")]
dataframemask_27jul
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 28 de julio de 2022
```

```
dataframemask_28jul = dataframemask[dataframemask["date"].str.contains("2020-07-28")]  
dataframemask_28jul
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 29 de julio de 2022
```

```
dataframemask_29jul = dataframemask[dataframemask["date"].str.contains("2020-07-29")]  
dataframemask_29jul
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 30 de julio de 2022
```

```
dataframemask_30jul = dataframemask[dataframemask["date"].str.contains("2020-07-30")]  
dataframemask_30jul
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 31 de julio de 2022
```

```
dataframemask_31jul = dataframemask[dataframemask["date"].str.contains("2020-07-31")]  
dataframemask_31jul
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 01 de agosto de 2022
```

```
dataframemask_01ago = dataframemask[dataframemask["date"].str.contains("2020-08-01")]  
dataframemask_01ago
```

user_name	user_location	user_description	user_created	user_followers	user_
-----------	---------------	------------------	--------------	----------------	-------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 02 de agosto de 2022
```

```
dataframemask_02ago = dataframemask[dataframemask["date"].str.contains("2020-08-02")]  
dataframemask_02ago
```



```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 03 de agosto de 2022

dataframemask_03ago = dataframemask[dataframemask["date"].str.contains("2020-08-03")]
dataframemask_03ago
```

user_name	user_location	user_description	user_created	user_followers	user_...
-----------	---------------	------------------	--------------	----------------	----------

```
# Obtenemos los tweets sobre cubrebocas/mascarillas escritos el 04 de agosto de 2022

dataframemask_04ago = dataframemask[dataframemask["date"].str.contains("2020-08-04")]
dataframemask_04ago
```

user_name	user_location	user_description	user_created	user_followers	user_...
-----------	---------------	------------------	--------------	----------------	----------

```
# Obtenemos los tweets sobre Donald Trump escritos el 25 de julio de 2022

dataframetrump_25jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-25")]
dataframetrump_25jul
```

	user_name	user_location	user_description	user_created
103	Thomas Faires	Toronto, Ontario, Canada	...yup, still a husband, vegetarian, computer ...	2011-04-14 12:36:09
234	BeautifulSkinYourIn		Daily Tips from our Nurses, Dietitians, Aesthe...	2016-11-17 14:30:31
412	marklaverdure	Los Angeles & Billings	I have passion for living life to it's fullest...	2011-03-04 15:17:55
437	Mx. Joe~Anthony Sierra Let us pray #COVID19 away	#StGeorge~#StatenIsland~#NYC	Hope to be insightful	2013-05-03 03:46:03

Obtenemos los tweets sobre Donald Trump escritos el 26 de julio de 2022

```
dataframetrump_26jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-26")]
dataframetrump_26jul
```

	user_name	user_location	user_description	user_created	user_followers	u
17324	Stilt	San Francisco, CA	Financial services for immigrants, visa holder...	2015-06-02 07:30:16	375	
17375	Driver H. Potter	At the front of a train, baby.	Proud Senior Minion to @Gordon_3417. Railway E...	2009-07-24 13:03:29	4459	
17442	Joshua Sandman			2016-12-03 00:58:13	0	
17445	Stilt	San Francisco, CA	Financial services for immigrants, visa holder...	2015-06-02 07:30:16	375	
17477	r scot	United States	Just be real	2020-04-25 21:52:57	53	

Obtenemos los tweets sobre Donald Trump escritos el 27 de julio de 2022

```
dataframetrump_27jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-27")]
dataframetrump_27jul
```

	user_name	user_location	user_description	user_created	user_followers	user_id
17506	Robert L.	Artist, writer, pianist				

Obtenemos los tweets sobre Donald Trump escritos el 28 de julio de 2022

```
dataframetrump_28jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-28")]
dataframetrump_28jul
```

	user_name	user_location	user_description	user_created	user_followers	user_id
17506	Ierence L.	Saginaw, MI	Album: VANDERBILT	2011-09-20	82	

Obtenemos los tweets sobre Donald Trump escritos el 29 de julio de 2022

```
dataframetrump_29jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-29")]
dataframetrump_29jul
```

	user_name	user_location	user_description	user_created	user_followers	user_id
--	-----------	---------------	------------------	--------------	----------------	---------

Obtenemos los tweets sobre Donald Trump escritos el 30 de julio de 2022

```
dataframetrump_30jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-30")]
dataframetrump_30jul
```

```

    user_name user_location user_description user_created user_followers user_
# Obtenemos los tweets sobre Donald Trump escritos el 31 de julio de 2022

dataframetrump_31jul = dataframetrump[dataframetrump["date"].str.contains("2020-07-31'
dataframetrump_31jul

```

```

    user_name user_location user_description user_created user_followers user_

```

```

# Obtenemos los tweets sobre Donald Trump escritos el 01 de agosto de 2022

dataframetrump_01ago = dataframetrump[dataframetrump["date"].str.contains("2020-08-01'
dataframetrump_01ago

```

```

    user_name user_location user_description user_created user_followers user_

```

```

# Obtenemos los tweets sobre Donald Trump escritos el 02 de agosto de 2022

dataframetrump_02ago = dataframetrump[dataframetrump["date"].str.contains("2020-08-02'
dataframetrump_02ago

```

```

    user_name user_location user_description user_created user_followers user_

```

```

# Obtenemos los tweets sobre Donald Trump escritos el 03 de agosto de 2022

dataframetrump_03ago = dataframetrump[dataframetrump["date"].str.contains("2020-08-03'
dataframetrump_03ago

```

```

    user_name user_location user_description user_created user_followers user_

```

```

# Obtenemos los tweets sobre Donald Trump escritos el 04 de agosto de 2022

dataframetrump_04ago = dataframetrump[dataframetrump["date"].str.contains("2020-08-04'
dataframetrump_04ago

```

```

    user_name user_location user_description user_created user_followers user_

```

Para determinar la evolución de los temas de interés en Twitter durante estos días de la pandemia, no podemos usar el algoritmo de K means porque necesitamos variables con valores numéricos, y en este caso estamos analizando datos de texto. Además de que no es conveniente transformar todos estos datos a su representación numérica, y tampoco sería muy relevante encontrar grupos de datos si pretendemos visualizar la evolución de los temas más populares en Twitter.

Tampoco sería pertinente emplear un heatmap porque no estamos buscando una correlación entre un día y la popularidad de un tema, simplemente se busca establecer la evolución de la popularidad de los temas.

Para visualizar y de esta manera entender la evolución de la popularidad de los temas de discusión en Twitter, es más conveniente utilizar una herramienta para graficar (como matplotlib o seaborn) los datos. En este caso, usaremos matplotlib para realizar un histograma, para poder observar los niveles de popularidad de cada tema en el transcurso de los días en los que se escribieron los tweets.

```
# Creamos un arreglo de los días en los que se escribieron los Tweets (para ajustarlos
dias = [" ", "Jul 25", "Jul 27", "Jul 29", "Jul 31", "Ago 02", "Ago 04"]

# Creamos un arreglo almacenando los valores obtenidos de la cantidad de tweets sobre
covid = [len(dataframecovid_25jul), len(dataframecovid_26jul), len(dataframecovid_27jul)]
print(covid)

# Hacemos un histograma de los valores obtenidos

tiempo = np.arange(len(covid))
esp = 0.3

fig, ax = plt.subplots()
ax.bar(tiempo, covid, esp)
ax.set_title("Tema: Covid19")
ax.set_xlabel("Días")
ax.set_ylabel("Tweets")
ax.set_xticklabels(dias)
```

```
[9625, 4411, 0, 0, 0, 0, 0, 0, 0, 0, 0]
<ipython-input-44-3f2a43c6e54e>:16: UserWarning: FixedFormatter should only be u:
    ax.set_xticklabels(dias)
[Text(-2.0, 0, ' '),
 Text(0.0, 0, 'Jul 25'),
 Text(2.0, 0, 'Jul 27'),
 Text(4.0, 0, 'Jul 29'),
 Text(6.0, 0, 'Jul 31'),
 Text(8.0, 0, 'Ago 02'),
 Text(10.0, 0, 'Ago 04'),

# Creamos un arreglo almacenando los valores obtenidos de la cantidad de tweets sobre

mask = [len(dataframemask_25jul), len(dataframemask_26jul), len(dataframemask_27jul),
print(mask)

# Hacemos un histograma de los valores obtenidos

tiempo = np.arange(len(mask))
esp = 0.3

fix, ax = plt.subplots()
ax.bar(tiempo, mask, esp)
ax.set_title("Tema: Cubrebocas/mascarillas")
ax.set_xlabel("Días")
ax.set_ylabel("Tweets")
ax.set_xticklabels(dias)
```

```
[345, 150, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

```
# Creamos un arreglo almacenando los valores obtenidos de la cantidad de tweets sobre
```

```
trump = [len(dataframetrump_25jul), len(dataframetrump_26jul), len(dataframetrump_27jul), len(dataframetrump_28jul), len(dataframetrump_29jul), len(dataframetrump_30jul), len(dataframetrump_31jul), len(dataframetrump_ago01), len(dataframetrump_ago02), len(dataframetrump_ago03), len(dataframetrump_ago04)]
print(trump)
```

```
# Hacemos un histograma de los valores obtenidos
```

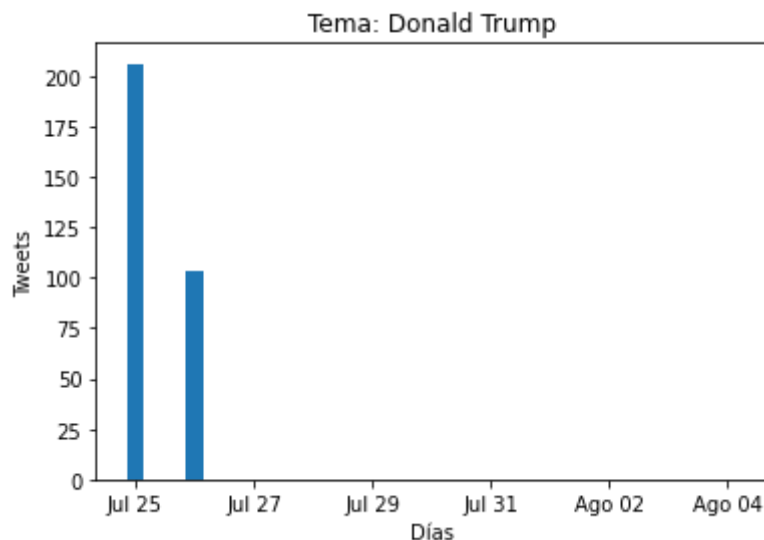
```
tiempo = np.arange(len(trump))
esp = 0.3
```

```
fig, ax = plt.subplots()
ax.bar(tiempo, trump, esp)
ax.set_title("Tema: Donald Trump")
ax.set_xlabel("Días")
ax.set_ylabel("Tweets")
ax.set_xticklabels(dias)
```

```
[206, 103, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

```
<ipython-input-46-5f33d8dd558d>:16: UserWarning: FixedFormatter should only be used with FixedLocator
```

```
ax.set_xticklabels(dias)
[Text(-2.0, 0, ' '),
Text(0.0, 0, 'Jul 25'),
Text(2.0, 0, 'Jul 27'),
Text(4.0, 0, 'Jul 29'),
Text(6.0, 0, 'Jul 31'),
Text(8.0, 0, 'Ago 02'),
Text(10.0, 0, 'Ago 04'),
Text(12.0, 0, ' ')]
```



CONCLUSIONES

A partir de los tres histogramas obtenidos, podemos concluir que la popularidad de los tres temas más relevantes tuvo su nivel más alto el día 25 de julio de 2020. Posteriormente, se observa un

descenso de popularidad de los tres temas alrededor del 29 o 30 de julio. Y finalmente, la gente vuelve a hablar de estos temas en los días posteriores hasta el 04 de agosto de 2020.

Puedo decir que mi objetivo principal de este análisis de datos se cumplió satisfactoriamente, puesto que se determinó adecuadamente qué temas fueron los más relevantes durante estos días mediante los hashtags, y se pudo determinar cómo fue fluctuando la popularidad de cada uno en el transcurso de todos los días esta base de datos.

Adicionalmente, es posible vincular una noticia o acontecimiento importante con el punto de popularidad máximo de cada uno de los tres temas (el cual es el mismo día, el 25 de julio de 2020).

Pablo Ricardo Duran Sanchez

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from datetime import datetime

data = pd.read_csv("covid19_tweets3.csv")
data.head(5)

# In[2]:

data.describe()

# Crear un diccionario que asocie cada país con un número

# In[3]:

country_codes = {}
countries = data['country'].unique()
for i, country in enumerate(countries):
    country_codes[country] = i+1

# Agregar una nueva columna al dataframe con los códigos de país correspondientes
```



```
# In[ ]:
```

```
data['country_code'] = data['country'].map(country_codes)
```

```
# In[5]:
```

```
print(data)
```

```
# Crear un diccionario que asocie cada hashtag con un número
```

```
# In[6]:
```

```
# Crear un diccionario que asocie cada hashtag con un número
```

```
hashtag_codes = {}
```

```
hashtags = data['hashtags'].unique()
```

```
for i, hashtag in enumerate(hashtags):
```

```
    hashtag_codes[hashtag] = i+1
```

```
# Agregar una nueva columna al dataframe con los códigos de país correspondientes
```

```
# In[7]:
```

```
data['hashtag_code'] = data['hashtags'].map(hashtag_codes)
```

```
# In[8]:
```

```
print(data)
```

```
# In[9]:
```

```
data.describe()
```

```
# In[10]:
```

```
data2 = data.loc[:, ["date", "country_code", "hashtag_code"]]
```

```
# In[11]:
```

```
print(data2)

# # Búsqueda de la cantidad óptima de clusters
#

# Calculando qué tan similares son los individuos dentro del cluster
#

# In[12]:

nc = []

for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, max_iter = 300)
    kmeans.fit(data2) #aplico k-means a la base de datos
    nc.append(kmeans.inertia_)

# Graficando los resultados para obtener la "Elbow Curve"

# In[13]:

plt.plot(range(1, 11), nc)
plt.title("Elbow Curve")
plt.xlabel ('Numero de Clusters')
plt.ylabel('WCSS') ##WCSS. Es un indicador de qué tan similares son los individuos dentro
plt.show()

# ## Aplicando el método k-means a la base de datos

# In[14]:

clustering = KMeans(n_clusters = 2, max_iter = 300) #Crea el modelo
clustering.fit(data2) #Aplica el modelo a la base de datos

# In[15]:

KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
n_clusters=2, n_init=10,
random_state=None, tol=0.0001, verbose=0)
```

```
# ## Agregando la clasificación al archivo original
#

# In[16]:

data2['KMeans Clusters'] = clustering.labels_ #Zos resultados del clustering se quarc
data2.head()

# ## Visualizando los clusters que se formaron
#

# In[17]:

X = np.array(data2[["date", "country_code", "hashtag_code"]])
kmeans = KMeans(n_clusters=2).fit(X)
centroids = kmeans.cluster_centers_
print(centroids)

# In[19]:

# Predicting the clusters
labels = kmeans.predict(X)
# Getting the cluster centers
C = kmeans.cluster_centers_
colores=['red', 'green']
asignar=[]
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = fig.add_subplot(111, projection = "3d")
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar, s=60)
ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c=colores, s=1000)

# In[20]:

# Getting the values and plotting it
f1 = data['hashtag_code'].values
f2 = data['country_code'].values

plt.scatter(f1, f2, c=asignar, s=70)
plt.scatter(C[:, 0], C[:, 1], marker='*', c=colores, s=1000)
plt.show()
```

Conclusiones

¿Es posible incluir el uso del algoritmo Kmeans para apoyar con la descripción de lo
#

- Si es posible de hecho ese era el objetivo, poder observar donde se agrupan los da
#

#

#

Para cada herramienta que utilizamos en el curso, mencionar su aplicación en la solu
#

- Mapas de calor: La información podría observarse a partir de mapas de calor pero a
#

- ~~Histogramas: Para el tipo de análisis los histogramas nos darían como resultado ir~~

FileNotFoundError Traceback (most recent call last)

<ipython-input-1-9d8cb07f6a3d> in <module>

23 from datetime import datetime

24

---> 25 data = pd.read_csv("covid19_tweets3.csv")

26 data.head(5)

27

⏏ 5 frames

/usr/local/lib/python3.9/dist-packages/pandas/io/common.py in get_handle(path_or_
storage_options)

784 if ioargs.encoding and "b" not in ioargs.mode:

785 # Encoding

--> 786 handle = open(

787 handle,

788 ioargs.mode,

FileNotFoundError: [Errno 2] No such file or directory: 'covid19_tweets3.csv'

SEARCH STACK OVERFLOW