Guatemala 17 de Septiembre de 2022
Diego Alvarez - 19498
Guido Padilla - 19200
Oscar Paredez - 19109

Data Science

Laboratorio 5 Análisis de Sentimientos

Github: https://github.com/diego59x/lab5AnalisisSentimientos

- Limpie y preprocese los datos. Describa de forma detallada las actividades de preprocesamiento que llevó a cabo.

  Se aseguró que cada tweet fuera de tipo string, luego de esto se removieron palabras que no influyen en el análisis de un tweet.

  Se removieron: emails, números, números de teléfono, conjunciones (inglés y español), direcciones, urls, emojis, tags de html, signos de puntuación, caracteres especiales.

```
text = df["text"].str.upper()

for i in range(len(text)):
    text[i] = nt.TextFrame(str(text[i])).remove_emails()
    text[i] = nt.TextFrame(str(text[i])).remove_numbers()
    text[i] = nt.TextFrame(str(text[i])).remove_phone_numbers()
    text[i] = nt.TextFrame(str(text[i])).remove_stopwords(lang = "en")
    text[i] = nt.TextFrame(str(text[i])).remove_btc_address()
    text[i] = nt.TextFrame(str(text[i])).remove_urls()
    text[i] = nt.TextFrame(str(text[i])).remove_stopwords(lang = "es")
    text[i] = nt.TextFrame(str(text[i])).remove_emojis()
    text[i] = nt.TextFrame(str(text[i])).remove_html_tags()
    text[i] = nt.TextFrame(str(text[i])).remove_puncts()
    text[i] = nt.TextFrame(str(text[i])).remove_special_characters()
```

```
0                    DEEDS REASON EARTHQUAKE ALLAH FORGIVE
1                         FOREST FIRE NEAR RONGE SASK CANADA
2       RESIDENTS ASKED SHELTER PLACE NOTIFIED OFFICER...
3        PEOPLE RECEIVE WILDFIRES EVACUATION ORDERS CA...
4       GOT SENT PHOTO RUBY ALASKA SMOKE WILDFIRES POU...
                                ...
7608    GIANT CRANES HOLDING BRIDGE COLLAPSE NEARBY HOMES
7609    ARIAAHRARY THETAWNIEST CONTROL WILD FIRES CALI...
7610                               M  UTCKM VOLCANO HAWAII
7611    POLICE INVESTIGATING EBIKE COLLIDED CAR LITTLE...
7612    LATEST HOMES RAZED NORTHERN CALIFORNIA WILDFIR...
```

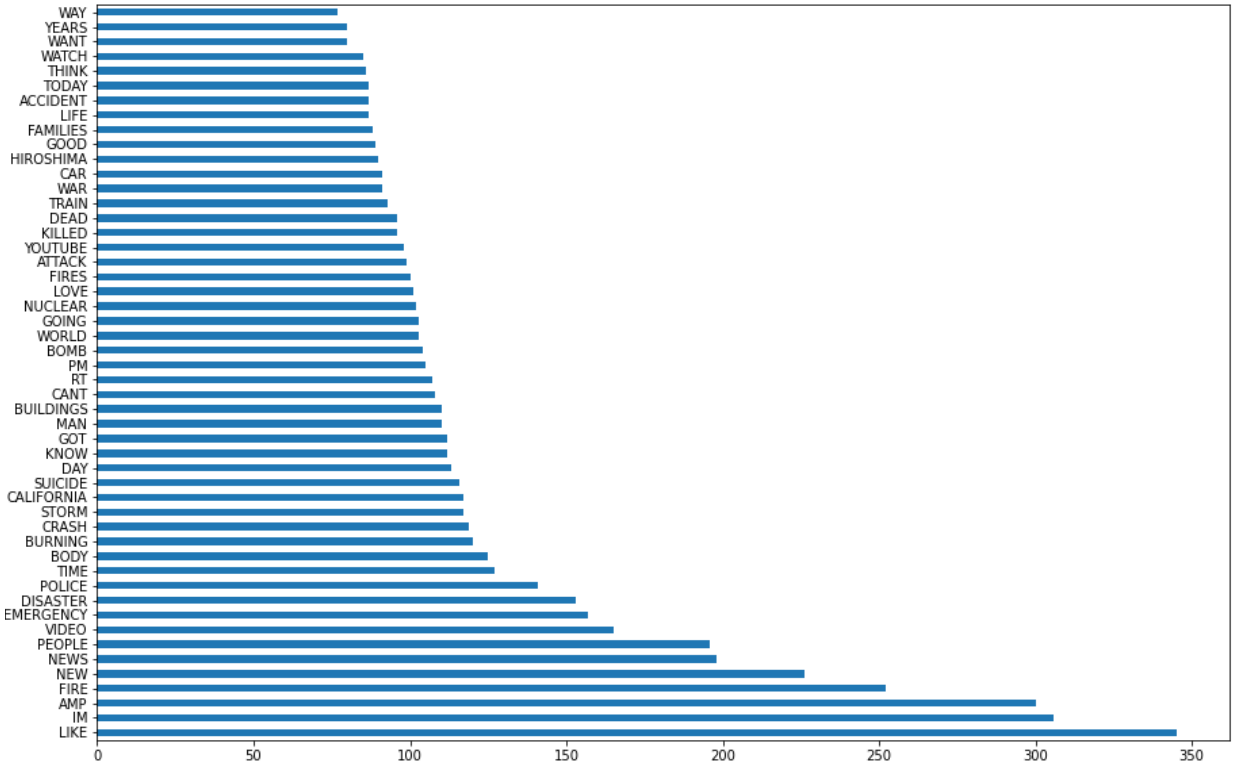
- Obtenga la frecuencia de las palabras tanto de los tweets de desastres como de los que no.

  Aquí podemos ver algunas de las palabras con sus frecuencias.

```
CATEGORIA = nan  || PALABRA =  Series([], )
CATEGORIA = ablaze  || PALABRA =  ABLAZE    28
CATEGORIA = accident  || PALABRA =  ACCIDENT    32
CATEGORIA = aftershock  || PALABRA =  AFTERSHOCK    19
CATEGORIA = airplane%20accident  || PALABRA =  ACCIDENT    35
CATEGORIA = ambulance  || PALABRA =  AMBULANCE    36
CATEGORIA = annihilated  || PALABRA =  ANNIHILATED    31
CATEGORIA = annihilation  || PALABRA =  ANNIHILATION    22
CATEGORIA = apocalypse  || PALABRA =  APOCALYPSE    28
CATEGORIA = armageddon  || PALABRA =  ARMAGEDDON    37
CATEGORIA = army  || PALABRA =  ARMY    33
CATEGORIA = arson  || PALABRA =  ARSON    29
CATEGORIA = arsonist  || PALABRA =  ARSONIST    17
CATEGORIA = attack  || PALABRA =  ATTACK    31
CATEGORIA = attacked  || PALABRA =  ATTACKED    35
CATEGORIA = avalanche  || PALABRA =  AVALANCHE    26
CATEGORIA = battle  || PALABRA =  BATTLE    23
CATEGORIA = bioterror  || PALABRA =  FEDEX    34
CATEGORIA = bioterrorism  || PALABRA =  BIOTERRORISM    24
CATEGORIA = blaze  || PALABRA =  BLAZE    20
CATEGORIA = blazing  || PALABRA =  BLAZING    28
CATEGORIA = bleeding  || PALABRA =  BLEEDING    35
CATEGORIA = blew%20up  || PALABRA =  BLEW    33
CATEGORIA = blight  || PALABRA =  BLIGHT    25
CATEGORIA = blizzard  || PALABRA =  BLIZZARD    18
...
CATEGORIA = wounds  || PALABRA =  WOUNDS    32
CATEGORIA = wreck  || PALABRA =  WRECK    51
CATEGORIA = wreckage  || PALABRA =  WRECKAGE    39
CATEGORIA = wrecked  || PALABRA =  WRECKED    39
```

  Aquí tenemos una nube de palabras la cual nos indica mejor qué palabras resaltan más que otras en cuestión de desastres.

Esta tabla de frecuencias respalda la imagen anterior.



- ¿Qué palabras cree que le servirán para hacer un mejor modelo de clasificación?
  Fire, IM, New, News, Disaster, California, People, Storm, Car, Suicide, Collapse, Video, Attacj, Accident, Time, Emergency, Police, Today.
- ¿Vale la pena explorar bigramas o trigramas para analizar contexto?
  Si vale la pena, pero de esta forma caemos en obviar circunstancias o combinaciones de palabras por parte de los usuarios, no todos escriben igual y

con orden. Por lo que se pueden referir a que ha ocurrido un desastre pero agrega un comentario personal o un dato secundario, con el cual se obvie este.

- Haga un análisis exploratorio de los datos para entenderlos mejor, documente todos los análisis

Dataset statistics

Number of variables	5
Number of observations	7613
Missing cells	2594
Missing cells (%)	6.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	297.5 KiB
Average record size in memory	40.0 B

Variable types

Numeric	1
Categorical	4

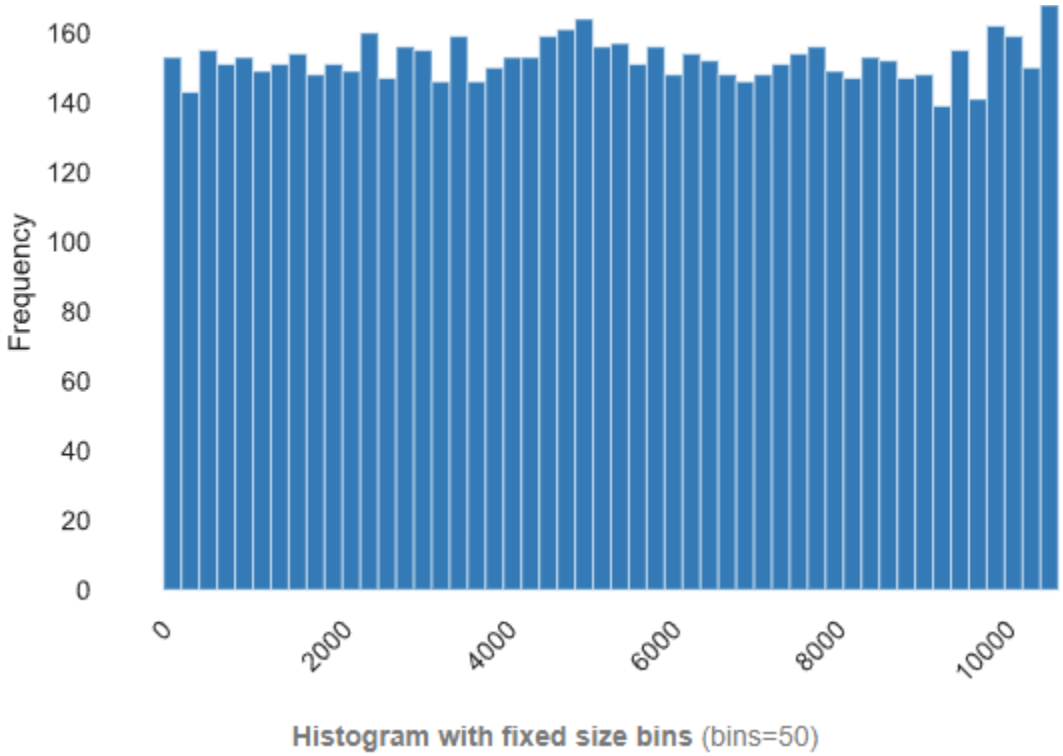
Variables

id

Real number ($\mathbb{R}_{\geq 0}$)

UNIFORM

UNIQUE



Histogram with fixed size bins (bins=50)

keyword

Categorical

HIGH CARDINALITY

UNIFORM

Value	Count	Frequency (%)
fatalities	45	0.6%
armageddon	42	0.6%
deluge	42	0.6%
sinking	41	0.5%
damage	41	0.5%
harm	41	0.5%
body%20bags	41	0.5%
twister	40	0.5%
collided	40	0.5%
windstorm	40	0.5%
Other values (211)	7139	94.5%

location

Categorical

HIGH CARDINALITY

MISSING

Value	Count	Frequency (%)
	377	3.5%
usa	255	2.4%
new	209	1.9%
the	176	1.6%
ca	148	1.4%
york	138	1.3%
london	110	1.0%
united	95	0.9%
uk	94	0.9%
in	91	0.8%
Other values (3291)	9156	84.4%

text

Categorical

`HIGH CARDINALITY`
`UNIFORM`

Value	Count	Frequency (%)
like	345	0.5%
im	306	0.5%
amp	300	0.5%
fire	252	0.4%
new	226	0.3%
news	198	0.3%
people	196	0.3%
video	165	0.3%
emergency	157	0.2%
disaster	153	0.2%
Other values (16731)	63331	96.5%

target

Categorical

Value	Count	Frequency (%)
0	4342	57.0%
1	3271	43.0%

En este set de datos tenemos las variables: id, keyword, location, text y target. De las cuales nos interesan únicamente Keyword, Location y Text, debido a que estas se relacionan adecuadamente a nuestro objetivo de estudio; clasificar tweets buenos y malos. Keyword nos indica la categoría del tweet, Location desde que país fue enviado y Text el contenido.

- Teniendo en cuenta la cantidad de palabras positivas y negativas del tweet determine qué tan positivo, negativo o neutral es el mismo.
  Utilizando la librería de nltk se pudo clasificar cada tweet en relación al porcentaje de la cantidad de palabras positivas, negativas y neutrales en el mismo. A su vez se obtuvo un cuarto valor llamado compound, el cual nos indica con mayor precisión si el tweet es clasificado como bueno, malo o neutral. Se estableció un rango de 0.66, donde los tweets con un valor menor a -0.33 serán negativos, mayores a 0.33 positivos y los que estén entre -0.33 y 0.33 serán neutrales.

```
{'neg': 0.0, 'neu': 0.656, 'pos': 0.344, 'compound': 0.2732}
{'neg': 0.324, 'neu': 0.676, 'pos': 0.0, 'compound': -0.34}
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

- Luego de analizar los datos determine:
    - ¿Cuáles son los 10 tweets más negativos?¿En qué categoría están?

Keyword es la categoría.

	id	keyword	location	text	target	score	polaridad
7472	10689	wreck	NaN	WRECK WRECK WRECK WRECK WRECK WRECK WRECK WREC...	0	-0.9879	negative
6414	9172	suicide%20bomber	NaN	ABUBARAA SUICIDE BOMBER TARGETS SAUDI MOSQUE D...	1	-0.9686	negative
6411	9166	suicide%20bomber	NaN	SUICIDE BOMBER KILLS SAUDI SECURITY SITE MOSQU...	1	-0.9623	negative
6393	9137	suicide%20bomb	Worldwide	TH DAY JUL NIGERIA SUICIDE BOMB ATTACKS KILL...	1	-0.9595	negative
6407	9159	suicide%20bomber	Worldwide	KILLED SARABIA MOSQUE SUICIDE BOMBING SUICIDE ...	1	-0.9552	negative
472	682	attack	portland, oregon	ILLEGAL ALIEN RELEASED OBAMADHS TIMES CHARGED ...	1	-0.9538	negative
1540	2225	chemical%20emergency	Las Vegas, Nevada	BOMB CRASH LOOT RIOT EMERGENCY PIPE BOMB NUCLE...	1	-0.9524	negative
6930	9940	trouble	NaN	CSPAN PREZ MR PRESIDENT BIGGEST TERRORIST TROU...	1	-0.9493	negative
2932	4213	drowned	Pembroke NH	LAKE SEES DEAD FISH ME POOR LITTLE GUY WONDER ...	0	-0.9477	negative
6438	9211	suicide%20bombing	NaN	REMEMBERING REBECCA ROGA PHILIPPINES MURDERED ...	1	-0.9451	negative

    - ¿Cuáles son los 10 tweets más positivos? ¿En qué categoría están?

Keyword es la categoría.

	id	keyword	location	text	target	score	polaridad
6992	10028	twister	NaN	CHECK WANT TWISTER TICKETS VIP EXPERIENCE SHAN...	0	0.9682	positive
3163	4541	emergency	Renfrew, Scotland	BATFANUK ENJOYED TODAY GREAT FUN EMERGENCY NON...	0	0.9423	positive
3382	4844	evacuation	Renfrew, Scotland	BATFANUK ENJOYED TODAY GREAT FUN EMERGENCY NON...	0	0.9423	positive
6292	8989	storm	NaN	TODAYS STORM PASS LET TOMORROWS LIGHT GREET KI...	1	0.9403	positive
2238	3198	deluge	NaN	MEDITATIONBYMSG PPL GOT METHOD MEDITATION UP A...	0	0.9287	positive
6295	8994	stretcher	NaN	FREE EBAY SNIPING RT LUMBAR EXTENDER STRETCHER...	0	0.9260	positive
6560	9386	survived	Puerto Rico	DUCHOVBUTT STARBUCKSCULLY MADMAKNY DAVIDDUCHOV...	0	0.9217	positive
1856	2668	crush	San Diego, Texas.	LOVE LOVE LOVE REMEMBER CRUSH	0	0.9186	positive
1909	2744	crushed	Trinidad & Tobago	DISILLUSIONED LEAD CHARACTER CHECK HAPPY LUCKY...	0	0.9136	positive
5033	7176	mudslide	London	IMPRESSIONS GLAD HAT MAN LEAVING LIEU INTEREST...	0	0.9100	positive

    - ¿Son los tweets de la categoría que indica que habla de un desastre real más negativos que los de la otra categoría?

Si, debido a que las palabras negativas en el texto están asociadas a desastres podemos determinar que si la negatividad de este tweet es alta es debido a que hay un desastre. Cabe mencionar que depende del banco de palabras, ya que si son palabras que incitan al odio y no a un desastre natural, pues no podemos determinar si existe o no un desastre.

- Cree una variable que contenga la “negatividad” de cada tweet. Inclúyala en el conjunto de datos y entrene nuevamente el modelo de clasificación de la hoja pasada. ¿La inclusión de esta variable mejoró los resultados del modelo de clasificación?

  Inicialmente se clasificó cada tweet debido al nivel de negatividad que se obtuvo agregando dos diferentes columnas al data frame, el cual indicaba el puntaje del mismo siendo -1 el más negativo y 1 el más positivo. La segunda columna agregada sería la que indicará si el tweet es negativo o positivo o neutral. Con estos datos agregados se filtraron para encontrar los más positivos y más negativos, obteniendo resultados acertados. Al momento de entrenar el modelo con estos nuevos datos se pudo llegar a predecir qué tweets tendrían cierto puntaje y con ello si es negativo o positivo. Sin embargo los resultados se mantuvieron, seguimos con el mismo grado de precisión pero claro que ahora no debemos clasificar por medio de otro modelo, sino usando los datos previamente calificados.