


```
In [74]: Counter({'NO_DATA': 48,
'BIRMINGHAM': 23,
'HERS': 32,
'AMERICA': 20,
'PHILADELPHIA_PA': 22,
'LONDON': 115,
'RHODE_ISLAND': 22,
'WELLSFIRE': 88,
'ALBERTA': 56,
'LIVE_MS': 13,
'MIDWEST': 42,
'OREGON': 23,
'ANGLES': 70,
'GAINESVILLE_FL': 25,
'INDIANA': 82,
'CANADA': 70,
'SOUTH_ASIA': 56,
'COLUMBUS_OHIO': 28,
'IT': 20,
'TRINIDAD_TOBAGO': 16,
'CALGARY_CANADA': 98,
'TWITTER': 6,
'CAPE_TOWN': 25,
'SAN_FRANCISCO': 112,
'IRELAND': 41,
'NASHVILLE_TN': 20,
'UK': 42,
'WALKER_COUNTY_ALABAMA': 11,
'FORTLAUD_ORE': 15,
'SAN_ANTONIO_TX': 21,
'BROOKLYN_NY': 23,
'LIVERPOOL': 30,
'LIVINGSTON_IL': 33,
'NEW_YORK': 170,
'LOUISIANA': 19,
'MANCHESTER': 32,
'WELLINGTON': 19,
'GLOBAL': 19,
'CHICAGO_IL': 77,
'BANGALORE_INDIA': 23,
'GARRETT': 12,
'LEICESTER': 23,
'WATSON_KENYA': 31,
'SINGAPORE': 29,
'KOWI': 10,
'MAD_HELL': 14,
'MORICH_JAPAN': 21,
'KERTU': 31,
'UNITED_STATES': 66,
'MUMBAI': 58,
'BAILEY_DURHAM_NC': 32,
'PENNSLVANIA': 41,
'INDONESIA': 22,
'PALO_ALTO_CALIFORNIA': 24,
'AIRES_ARGENTINA': 32,
'HEAD': 20,
'PERTH_WESTERN_AUSTRALIA': 19,
'FINANCIAL_NEWE_VIEWS': 13,
'SILICON_VALLEY': 11,
'FOTTERS_HANDS': 15,
'BRISBANE': 30,
'T': 29,
'NYC': 41,
'CARRY': 25,
'WEST_COAST': 23,
'BAFFLY_MARRIED_KIDS': 21,
'RID': 16,
'ONDO': 13,
'WILLAMSBURG': 9,
'ATLANTA': 100,
'WINTER_PARK_COLORADO': 19,
'GUA_BUCK_W_COV': 4,
'CT': 10,
'LAGOS': 45,
'JOHANNESBURG': 20,
'PITTSBURGH': 13,
'ROAD_WILLONWAINES_CLUB': 46,
'NIJERIA': 49,
'HOUSTON': 56,
'AMERICAN_MASTELAND_MV': 20,
'TORRANCE': 19,
'VIRGINIA_UNITED_STATES': 23,
'CALIFORNIA': 60,
'VICTORIA_BC': 19,
'SYDNEY_AUSTRALIA': 30,
'OK': 13,
'REPUBLIC_TEXAS': 17,
'FRENCH_27': 27,
'RIGHT_YOU': 17,
'HUNTSVILLE': 11,
'GRANCO': 31,
'TEXAS': 28,
'ILLINOIS': 22,
'ASHEVILLE_NC': 14,
'AUSTIN_TX': 22,
'BIRM': 10,
'OKLAHOMA_CITY': 29,
'VEGAS_NEVADA': 30,
'LONDON_ENGLAND': 39,
'MACLESFIELD': 9,
'FLORIDA': 32,
'TAYLOR_SWIFT': 17,
'NEWCASTLE_OK': 37,
'MACONIA': 23,
'KENYA': 36,
'PERTSHIRE': 16,
'MEMPHIS_TN': 28,
'PARIS': 29,
'WASHINGTON_DC': 73,
'UK_GERMANY': 15,
'INTERNET': 30,
'NORTH': 30,
'SS': 13,
'ELAK_RIDGE': 13,
'KAMA_FRANCE': 13,
'DALLAS_TX': 28,
'PORT_JERVIS_MV': 20,
'DUBLIN': 19,
'SEATTLE_WA': 62,
'DEL': 11,
'LYTHAM_ST_MANS': 15,
'MADISON_WI': 28,
'GOLD_COAST_AUSTRALIA': 18,
'HTX': 8,
'CENTRA_COAST_CALIFORNIA': 26,
'BUY_MONEY': 18,
'DENVER_COLORADO': 31,
'BRASIL': 22,
'PHILIPPINES': 15,
'METRVILLE': 17,
'PAINTCOM': 27,
'QUEENS': 16,
'UTAH': 15,
'MACON_GA': 23,
'MELBOURNE_AUSTRALIA': 30,
'MOON': 19,
'FRESNO': 13,
'DC': 7,
'ROCKY_MOUNTAINS': 16,
'GOTHAM_CITYUSA': 17,
'CONCAC_WF': 8,
'SECRET': 16,
'CHINA': 23,
'BALTIMORE_MD': 17,
'TENNESSEE': 24,
'SOCHI_RUA_RU': 6,
'NORWAY': 7,
'VANCOUVER_BC': 25,
'FLANRY_BATH': 20,
'ONTARIO_CANADA': 17,
'MFP': 6,
'IRA_BROCH_HIT': 15,
'IRAQAFGHANISTAN_RSA_BAGHDAD': 9,
'NEW_ENGLAND': 29,
'MING_CITY': 18,
'THAILAND': 24,
'KIDNEY': 4,
'VENTURA': 14,
'UNITED_KINGDOM': 26,
'PATRSON_NEW_JERSEY': 20,
'NOTTINGHAM_ENGLAND': 12,
'NJ': 11,
'ISTANBUL': 7,
'WACO_TEXAS': 15,
'ENGLAND_UNITED_KINGDOM': 13,
'TULSA_OK': 12,
'CHEVY_CHASE_MD': 11,
'LOWELL': 4,
'REDDING_CALIFORNIA': 15,
'PORTO': 11,
'UNWOOOTDOOS_TOR': 7,
'KUALA_LUMPUR_MALAYSIA': 8,
'IM_SENT': 9,
'MICHIGAN': 15,
'US': 15,
'CONVENTRY': 19,
'HONG_KONG': 10,
'SCOTTSDALE_AZ': 12}}
```

```
In [76]: palabras = []
frecuencias = {}
for i in ifrequencies.keys():
    palabras.append(i)
    frecuencias.append(i)

for k in ifrequencies.values():
    frecuencias.append(k)

In [77]: index1 = frecuencias.index(max(frecuencias))
pal = palabras[index1]
frecuencias.pop(index1)

index2 = frecuencias.index(max(frecuencias))
pa2 = palabras[index2]
frecuencias.pop(index2)

index3 = frecuencias.index(max(frecuencias))
pa3 = palabras[index3]
frecuencias.pop(index3)

print("La palabra más frecuente es: ", pal)
print("La segunda palabra más frecuente es:", pa2)
print("La tercera palabra más frecuente es:", pa3)

La palabra más frecuente es: NO_DATA
La segunda palabra más frecuente es: LIVINGSTON_IL
La tercera palabra más frecuente es: PHILADELPHIA_PA
```

```
In [78]: n = 100
words = palabras[0:n]
freq = frecuencias[0:n]

# Plot histogram using matplotlib bar().
indexes = np.arange(len(words))
width = 0.7

fig = plt.figure(figsize = (25, 12 ))
plt.bar(indexes, freq, width)
plt.xticks(indexes + width * 0.5, words)
plt.xticks(rotation = 90)
plt.show()
```



```
Text
In [88]: # convertimos todo a un string
Text = " ".join(review for review in data.text.astype(str))

In [91]: Frequencies = Counter(Text.split())

In [92]: Frequencies
```

[illegible]

```
'WEBS': 8,
'GRUNT': 8,
'HISTORY': 25,
'INSTANTLY': 2,
'SMASH': 5,
'ABILITY': 4,
'ANNILILATE': 1,
'ROMANTITY': 3,
'TREVOUS': 2,
'WENT': 33,
'MINUS': 1,
'FACT': 12,
'STOPPED': 9,
'QUICKLY': 9,
'SHORT': 12,
'BALL': 21,
'TORNAIL': 1,
'G': 43,
'ORXX': 1,
'SYMBOL': 1,
'SABIAN': 4,
'PENINSULA': 1,
'HUNTERS': 7,
'THEY': 1,
'READY': 18,
'BUCS': 1,
'PHILLIPDUNCAN': 1,
'BREAFASTONS': 1,
'NIGRES': 2,
'WEATHER': 51,
'PHILIP': 1,
'THOUGHT': 31,
'FORECAST': 9,
'DOMAIN': 1,
'SOPHISTICATON': 1,
'CLOSEL': 1,
'UPPTHEMINUTE': 1,
'FEAT': 12,
'GRAT': 1,
'STORMBEARD': 1,
'STEELORD': 1,
...]]
```

```
In [93]: palabras = []
frecuencias = {}
for i in Frequencias.keys():
    palabras.append(i)

for k in Frequencias.values():
    frecuencias.append(k)
```

```
In [94]: index1 = frecuencias.index(max(frecuencias))
pal = palabras[index1]
frecuencias.pop(index1)

index2 = frecuencias.index(max(frecuencias))
pa2 = palabras[index2]
frecuencias.pop(index2)

index3 = frecuencias.index(max(frecuencias))
pa3 = palabras[index3]
frecuencias.pop(index3)

print("La palabra más frecuente es: ", pal)
print("La segunda palabra más frecuente es:", pa2)
print("La tercera palabra más frecuente es:", pa3)

La palabra más frecuente es: LIKE
La segunda palabra más frecuente es: IM
La tercera palabra más frecuente es: PUBLICATIONS
```

```
In [95]: n = 100
words = palabras[0:n]
freq = frecuencias[0:n]

# Plot histogram using matplotlib bar().
indexes = np.arange(len(words))
width = 0.7

fig = plt.figure(figsize = (25, 12 ))
plt.bar(indexes, freq, width)
plt.xticks(indexes + width * 0.5, words)
plt.xticks(rotation = 90)
plt.show()
```



Hallazgos y conclusiones

En Twitter y las redes sociales el lenguaje es metafórico y no textual, por lo que es difícil realizar conclusiones precisas al respecto. Analizando la frecuencia y el orden de las frases más comunes, se obtendrá una clasificación para las expresiones comunes y aquellas que sean referentes a desastres naturales reales. Luego de unificar la localidad del tweet, es decir, su origen, es posible concluir que, para la ubicación de los tweets:

- La palabra más frecuente es: NO_DATA
- La segunda palabra más frecuente es: LIVINGSTON_IL
- La tercera palabra más frecuente es: PHILADELPHIA_PA

es decir, es más común no contar con una ubicación precisa, y Livingston y Philadelphia tienen una alta cantidad de registros en esta área.

Por otro lado, de la dispersión de los datos, podemos observar que la mayoría no son referentes a desastres reales. Este hecho es importante y será utilizado más adelante al plantear el modelo, para conocer así qué clase de patrones podemos esperar en tweets "falsos".

Además, podemos observar según las gráficas, que el tipo de desastre o elemento más registrado son:

- La palabra más frecuente es: BOMBING
- La segunda palabra más frecuente es: FATALITIES
- La tercera palabra más frecuente es: CRASHED

por lo que estos desastres deben tomarse en cuenta a la hora de explorar y así saber qué podemos esperar de los datos.