Organización de Datos 75.06/95.58
Trabajo Práctico Nº 1

DatUs

Apellido y Nombre	Padrón	Correo electrónico
Boada Ignacio	95212	ignacio.boada@outlook.com
Goñi Mauro Ariel	87646	maurogoni@gmail.com
Perez Cristian	95536	cfperez@fi.uba.ar
Perez Machado Axel	101127	axelmpm@gmail.com

`Github:`
`https://github.com/axelmpm/TP1_DATOS_1C_2020`
`Kaggle:`
`https://www.kaggle.com/axelmpm`

Indice

1 Introducción **4**
1.1 Hipótesis de Trabajo 4
1.2 Supuestos 4
1.3 Objetivo del informe 4

2 Target **5**
2.1 ¿Cómo se distribuyen las palabras claves en los tweets con desastre? 5
2.1.1 TOP 10 mas frecuentes 6
2.1.2 TOP 10 menos frecuentes 8
2.1.3 NaN como keyword 9
2.2 Porcentaje de Desastres por Keyword 10

3 Usuarios **12**
3.1 ¿Que proporción de menciones hay en los tweets? 12
3.2 ¿Cuantas menciones tienen los tweets? 13
3.3 ¿Cuales fueron los usuarios con mas menciones? 14
3.4 ¿Promedio de palabras en tweets con menciones? 15

4 Análisis del lenguaje de los tweets **16**
4.1 ¿Qué porcentaje de inglés tienen los tweets? 16
4.2 ¿Los tweets con mayor porcentaje de inglés tienen más desastres? 18
4.3 ¿Existe alguna relación entre el porcentaje de ingles, la longitud de los tweets y los desastres? 19

5 Métricas del Desastre: Negatividad, Importancia y Longitud **20**
5.1 La Negatividad de un Tweet 20
5.1.1 Que es un Token? 20
5.2 Continuando la definición de Negatividad 21
5.3 Pre preocesamineto de texto 22
5.4 El núcleo del asunto. Como asignamos negatividad a los tokens? 23
5.5 Visualizando los resultados de la negatividad 24
5.6 La ecuación mas peligrosa de la historia 25
5.6.1 Que es chico? 25
5.7 Conclusiones de la métrica de Negatividad 26
5.8 La métrica de la Longitud 27
5.9 Distribución de longitudes 27
5.10 Efectividad de la Longitud como medición del desastre 28
5.11 Combinando métricas. Longitud y Negatividad 28
5.12 Preliminares 29
5.13 Distribución de tweets en Longitud y Negatividad 30
5.14 Efectividad de la combinación Longitud y Negatividad 31

5.15 Conclusión de la Longitud como métrica del desastre 32
5.16 La Importancia de un tweet como métrica del desastre 32
5.17 Distribución de la Importancia de los tweets 33
5.18 Efectividad de la Importancia como métrica del desastre 34
5.19 Combinando métricas. Importancia y Negatividad 34
5.20 Distribución de tweets en Importancia y Negatividad 35
5.21 Efectividad de la combinación Importancia y Negatividad 36
5.22 Conclusión de la Importancia como métrica del desastre 36

6 Locación **37**
6.1 ¿La cantidad de tweets por continente es igual? 37
6.2 ¿Cómo se distribuyen los desastres reales por continente? 38
6.3 ¿Cuantas palabras claves hay por continente? 39
6.4 ¿Cual es la palabra clave mas popular en cada continente? 40
6.5 Cantidad de tweets por país 41
6.6 ¿Qué países tienen mayor concentración de tweets con desastres? 42
6.7 ¿Cuales son los países con menos palabras claves? 43
6.8 ¿Cómo se distribuye la longitud de los tweets por país? 44
6.9 ¿Qué ciudades tienen la mayor cantidad de tweets? 45
6.10 Ciudades Con mayor proporción de desastres 46

7 Tops **47**
7.1 ¿Cuales son los Hashtags mas utilizados? 47
7.2 ¿Cual es la frecuencia de los hashtags mas usados si llevamos a su palabra raiz? 48
7.3 ¿Cuales son los Hashtags 15 hashtags mas usados? 49
7.4 ¿Cuales son los Hashtags 15 hashtags menos usados? 50
7.5 ¿Como se encuentran distribuidos los hashtags en el léxico? 51
7.6 ¿Cual es la frase mas frecuente? 52

8 Curiosidades **53**
8.1 ¿Existen tweests Repetidos? 53
8.2 Palabra mas comun en keywords 55
8.3 ¿El tweet menciona al keyword? 56
8.4 Tweets que contienen alguna URL en el texto 57
8.4.1 Cantidad de Tweets con URL 57
8.4.2 Target en tweets con URL 58
8.4.3 ¿Como se distribuyen los tweets con URL en las keywords? 59
8.5 ¿Los tweets con preguntas tienen desastres reales? 61

9 Conclusión **62**
9.1 Conclusiones del análisis 62
9.2 Conclusiones del análisis de Target 62

9.3 Conclusiones del análisis de Usuarios 62
9.4 Conclusiones del Lenguaje de los Tweets 62
9.5 Conclusiones finales de la Longitud, Negatividad e Importancia 63
9.6 Conclusiones del análisis de Locación 63
9.7 Conclusiones del análisis de Tops 63
9.8 Conclusiones del análisis de Curiosidades 64

10 Bibliografía y Referencias **65**

1 Introducción

1.1 Hipótesis de Trabajo

El presente informe consiste en el análisis exploratorio del set un set de datos de tweets, los que pueden o no relacionarse a desastres reales.

Nos hemos enfocado en Pandas, Phyton y Jupiter notebook como herramientas para abordar el análisis de los datos del set.

Inicialmente, se realizaron tareas de reconocimento de los datos, que información contenían y de que tipo, posteriormente un pre-procesamiento de los mismos y una serie de plots simples para darnos una idea de con que volumen de datos y como estaban constituidas cada columna. Luego de depurar los datos se generó el dataset utilizado en los análisis posteriores.

El código se encuentra disponible en:
`https://github.com/axelmpm/TP1_DATOS_1C_2020`

1.2 Supuestos

Para poder procesar los datos e indentificar caracteristicas de los mismo, nos basamos en diversas librerias, alguna creada por nosotros (datuslib) y otras ya existente, tales como, NLTK (libreria de procesamiento del lenguaje natural), geonamescache (para ampliar información acerca de una locación), geopandas (que nos brinda simplicidad a la hora de graficar), entre otras, que nos han permitido abordar los datos de forma mas eficiente.

1.3 Objetivo del informe

El informe realizado busca resaltar caracteristicas particulares de lo tweets, los procedimientos llevados a cabo para abordar los datos y como se realizó el procesamiento de los mismos, a fin de poder brindar información relevante para determinar si un tweet se corresponde con un desastre real o no.

2 Target

En esta sección, se mostrará un análisis realizado sobre los tweets que se refieren a desastres reales.

2.1 ¿Cómo se distribuyen las palabras claves en los tweets con desastre?

Se quiere saber cómo se distribuyen las palabras clave, es decir, cuáles se repiten más y cuáles menos dentro de los tweets con desastres reales ya que esto puede servir para tener en cuenta en caso de querer predecir si un tweet se trata de un desastre real o no. Para este análisis se filtraron las palabras claves que no están en inglés ya que se busca analizar únicamente los tweets en inglés, que representan el 99.8% de los datos.

Figure 1: Frecuencia de las palabras clave de los tweets con desastre.

En la imagen se puede ver que las palabras *wreckage*y *outbreak* son las más frecuentes. Una posible explicación para estas palabras es que "destrucción" (wreckage) es una palabra muy general para desastres, ya que casi cualquier desastre causa destrucción, ya sea un tornado, terremoto, incendio, inundación, etc, por lo tanto muchas personas decidieron usarla como palabra clave al hablar de un desastre real, y en el caso de un brote (outbreak) hay muchas personas afectadas y se necesita informar a la gente distintas medidas de seguridad para evitar su propagación por lo que mucha gente decidió usar twitter para transmitir esa información,

2.1.1 TOP 10 mas frecuentes

Se quizo realizar un TOP 10. Para hacerlo, habia keywords que eran NaN y se reemplazo por Nothing(nuevo keyword), entonces se filtro por target (1).

Figure 2: Keyword mas usados con desastres reales.

Se puede observar que el keyword mas comun es Nothing, es decir ningun (NaN). Sacando este detalle (NaN es este grafico es un keyword llmado Nothing), este TOP esta en resonancia con el grafico anterior, *wreckage*y *outbreak* son las más frecuentes.

Como un analisis mas de Keyword, se puede comparar el TOP 10 de keywords mas usados con desastres con el siguiiente TOP.

Figure 3: Keyword mas usados en el dataset.

Que es el mismo TOP pero del dataset completo.
Puede verse que Nothing o no poner keyword es lo "mas usado" en relacion a keywords en ambos casos.

2.1.2 TOP 10 menos frecuentes

De igual forma se realizo un TOP 10 con los menos frecuentes.

Figure 4: Keyword menos usados con desastres reales.

2.1.3 NaN como keyword

Viendo que NaN es la keyword, si es que se puede usar como keyword, mas usada. Se procedio a ver en que proporcion esta dividida en Desastres y no Desastres.

Figure 5: Target en NaN.

Sabiendo que 61 tweets son NaN y solo el 69% son Desastres. Entonces 42 tweets del total son Desastres, por lo que la Figura 2 de la seccion 2.1 se corrobora con esto, ya que si se tiene en cuenta que los keywords de la Figura 3 que estan debajo de Nothing(NaN) no tienen todos sus tweets con Desastres, el 42 estaria a la cabeza si solo se hablaran solo de los de Desastres(figura 2).

2.2 Porcentaje de Desastres por Keyword

Si quiero averiguar si un tweet es Desastre o no y solo tengo su keyword, ¿Puedo tener una nocion de que es?.
Partiendo de esa pregunta, se analizo que keyword tiene mas probabilidad (casos favorables sobre casos totales) de ser Desastre del dataset y en base a eso se puede tener una idea para seguir.
Como candidata a ser mas desastres que no desastre, se tomo que debe tener mas del 90% de los tweets en esa keyword como desastre.

Figure 6: Keywords con Porcentaje de aciertos mayores.

Como conclusion, con el set de datos que tenemos, se puede considerar a los tweets con KEYWORD **"wreckage", "debris" y "derailment"** como sospechosos (en el TP2) a ser Desastres reales, ya que su % de aciertos es del 100%. Mientras que a los demas del barplot tambien se los puede tomar como posibles desastres, pero con menor certeza.

De igual forma se realiza con los de menor porcentaje,mostrando solo aquellos que tienen un porcentaje menor al 10 en aciertos.

s

Figure 7: Keywords con Porcentaje de aciertos menores.

Aca, ya por tener un pequeño % de aciertos, todas las keywords no entregan datos seguros para predecir si es desastre y se deberia encarar por otro feature.
Pero me da una nocion de que si la keyword es **"aftershock"**, seguro **no** es Desastre, ya que tiene un 0% de aciertos.

3 Usuarios

En esta sección, nos enfocaremos en la búsqueda de características de los tweets que contienen usuarios, que es posible que nos brinden información relevante.

3.1 ¿Que proporción de menciones hay en los tweets?

Se quiere determinar la proporción de usuarios mencionados en los tweets, para saber cuantos que proporción de cada situación (no desastre y desastre) cuenta con menciones.

Figure 8: Relación de las menciones de acorde a tipo de evento

Como se aprecia en el gráfico precedente, notamos que la mayoría de los tweets que hacen referencia a desastres reales no contienen menciones, y la proporción de estas es menor que en los no desastres.

3.2 ¿Cuantas menciones tienen los tweets?

Queremos determinar cuantas menciones tienen los tweets, observando si se aprecia algún patrón en los tweets de cada situación.

Figure 9: Cantidad de menciones en tweets de ambos tipos

Como se puede ver, el ámbito de convivencia de la mayoría de los tweets que tienen menciones se restringe casi en su totalidad a menciones entre 1-3 usuarios, siendo muy poco frecuentes una cantidad mayor.

3.3 ¿Cuales fueron los usuarios con mas menciones?

Lo que buscamos es encontrar alguna relación entre los usuarios mas mencionados con los desastres.

Figure 10: Menciones a usuarios discriminadas por desastre y no desastre

Se puede observar que el usuario con mayor frecuencia es "youtube", el mismo posee un gran numero de menciones en tweets que no son desastre pero también cuenta con una frecuencia significativa de tweets considerados desastre, en esta senda, podemos observar que medios de información cuentan con mas menciones en eventos que son desastres al igual que los usuarios "usagov" y "potus" (que es la cuenta del actual presidente de USA).

3.4 ¿Promedio de palabras en tweets con menciones?

Buscamos evaluar la longitud promedio de los tweets luego de quitarles las stopwords, por lo que nos quedamos con los tweets con menciones y queremos observar como varían entre desastres y no desastres.

Figure 11: Promedio de palabras separado por cantidad de menciones

Podemos observar que se aprecia una menor cantidad de menciones en tweets con desastres, así como también un promedio de palabras mayor para tweets con 1 o 2 menciones, mientras que en las otras, es mayor el promedio de palabras de los twees de no desastres.

4 Análisis del lenguaje de los tweets

En esta sección se busca mostrar el análisis realizado sobre los tweets para poder considerar que el lenguaje utilizado es casi exclusivamente el inglés.

4.1 ¿Qué porcentaje de inglés tienen los tweets?

Se quiere saber en qué idiomas están los textos de los tweets para poder realizar un análisis de la información que contienen. Como se observó algunos tweets en inglés, se decidió primero analizar el porcentaje de inglés que contienen. Para esto se aplicó una función que limpie las palabras de caracteres inválidos y se utilizó un diccionario inglés para verificar si la palabra pertenece a ese idioma.

Figure 12: Porcentaje de ingles en el texto de los tweets

El histograma está volcado hacia la derecha, con una cola a izquierda. Se puede observar que la mayoría de los tweets tiene más de un 70% de las palabras en inglés. Si se considera

que un tweet con más de la mitad de las palabras en inglés es un tweet en inglés, solamente quedan 150 tweets que no están inglés, de los 7613 tweets a analizar, lo que representa menos de un 0.02% de los mismos, por lo que se pueden considerar despreciables, y concluir que los textos de los tweets están en inglés.

4.2 ¿Los tweets con mayor porcentaje de inglés tienen más desastres?

Se quiere saber cómo se distribuye el porcentaje de inglés respecto al hecho que los tweets tengas desastres reales para analizar si existe alguna relación.

Figure 13: Relación entre el porcentaje de inglés y los desastre en en los tweets

En el gráfico se puede observar la comparación entre los tweets con desastre y sin desastre según su porcentaje de inglés, con un histograma donde se ve la concentración de tweets por cantidad, y la densidad para mostrar con mayor precisión la distribución. Como conclusión se puede comentar que los tweets sin desastres se concentran entre los que tienen más de un 90% de inglés, mientras que aquellos que tienen desastres reales tienen una mayor concentración entre los tweets con un porcentaje de inglés entre un 60% y 80%.

4.3 ¿Existe alguna relación entre el porcentaje de ingles, la longitud de los tweets y los desastres?

Se quiere averiguar si hay alguna relación entre la longitud de los tweets y el porcentaje de inglés en los tweets, para lo cuál se realizó un scatter plot que relacione ambas variables.

Figure 14: Porcentaje de ingles vs Longitud de los textos de los tweets y su relación con los desastres

El gráfico muestra que no existe una clara relación entre el porcentaje de inglés, la longitud de los tweets y el hecho que estos traten sobre desastres reales. Se puede observar una concentración de tweets sin desastre entre los tweets con menos de 40 caracteres y más de un 60% de inglés.

5 Métricas del Desastre: Negatividad, Importancia y Longitud

En esta sección se aventura el analizar diferentes potenciales "métricas naturales" para los tweets y confirmar o refutar la correlación que pueden tener con si el tweet es o no un desastre.

Entre estas mencionadas "métricas naturales" contamos con:

- La longitud del tweet en caracteres de su campo "text"
- La Importancia del tweet (que definiremos mas adelante)
- La Negatividad de un tweet (que también definiremos luego)

5.1 La Negatividad de un Tweet

Dado que una de las principales características que llaman la atención en los tweets es su campo target, es decir si es o no un desastre, nosotros consideramos la posibilidad del hecho de que un tweet fuese un desastre estaría necesariamente relacionado con un "tono" mas negativo o lúgubre. Es decir un tweet que tiene un tono animado y feliz, tendería a ser menos desastroso que un tweet en un tono triste o alarmante.

En esta linea se define la negatividad de un tweet. Nuestra intención original (y la que usamos para guiar la definición final) fue asignar una negatividad a cada palabra (o token) y definir la negatividad de un tweet como la suma de las negatividades de las palabras o tokens que lo componen.

5.1.1 Que es un Token?

Al hechar un vistazo a lo que la gente escribe en los campos de texto de los tweets uno se encuentra de todo. No solo faltas de ortografía, sino otros idiomas, links de internet, memes, chistes, jerga, etc. Proponemos entonces una generalización de la noción de palabra en un concepto que llamamos "token". Token seria como un conjunto de caracteres que tiene sentido en si mismo y que no es una oración sino una unidad. Una palabra bien escrita es un token pero no todo token es una palabra.

5.2 Continuando la definición de Negatividad

El problema con el enfoque inocente de asignar negatividades a tokens y luego sumar, es la arbitrariedad de al tarea.

- Con que criterio se asigna la Negatividad?
- No es una tarea subjetiva del asignador la de definir arbitrariamente un numero de negatividad a cada token?
- Como asignar negatividad a CADA token del data set?
- No hay palabras que solo son negativas EN CONJUNTO con otras?

Para resolver estos problemas se recurrió al desarrollo de una clase que llamamos Context que realiza varias tareas pero que a grandes rasgos se pueden descomponer en las siguientes:

- Extraer de cada tweet los textos a usar
- Extraer de cada texto los tokens
- Identificar "el significado" de cada token y mapearlos a tokens mas "esenciales" reduciendo así el espacio a etiquetar
- Crear una serie de maps de tokens que cargan con la información contextual de cada token. Es decir que en el se conoce por ejemplo cuantas veces apareció cierto par de tokens en el data set
- Asignar usando el map previamente mencionado, una negatividad a cada token

5.3 Pre preocesamineto de texto

Tal como indican los primeros tres pasos de la clase Context, primero se extrajeron los textos de los tweets. Esto se realizo bajo la concatenación de los campos "text", "location" y "keyword" (en ese orden), habiendo previamente mapeado los NaNs y Nones a espacios en blanco.

Acto seguido se realizo un split por espacios en blanco, para luego separar cada parte spliteada en subpartes dependiendo de si esta contenida ciertos caracteres especiales o no.

Es decir, si se tenia "hola3que/tal". Este cacho de texto se procesaba a la siguiente lista de tokens ["hola","que","tal"].

Luego siguió el mapeo de tokens a tokens mas "esenciales". Esto se podría haber hecho de forma mucho mas inteligente pero nuestro enfoque fue mas practico y acá lo que se hizo fue:

- Mapear todas las direcciones web a un flag WEB ADDRESS
- Mapear todas los acrónimos a un flag ACRONIMS
- Mapear todas los espacios en blanco de diferentes tipos a un flag EMPTY
- Mapear todas los tokens con solo números a un flag NUMERIC
- Mapear todas los tokens que no eran NUMERIC y que no tenían ni una letra del alfabeto a un flag NON ALPHABETICAL
- Mapear a todos los tokens que no resultaron en flags a su versión en minúscula y que no poseían caracteres especiales. Si poseían caracteres especiales, se los quitaban y se aplicaba recursivamente toda la limpieza de nuevo.

En un momento planteamos la posibilidad de mapear los tokens a su equivalente mas similar en el ingles usando las distancia de Hamming (pero en una versión modificada). El mapeo que conseguíamos no nos convenció del todo.

5.4 El núcleo del asunto. Como asignamos negatividad a los tokens?

Una vez que tenemos cada tweet mapeado a unos tokens procesados lo que hacemos es extraer la frecuencia de cada par de tokens que aparece en el data set. Es decir de la frase "hola que tal hola" extraemos (hola, que) 2, (hola, tal) 2, (hola, hola) 1, (que, tal) 1

El numero que acompaña a cada par de tokens es lo que llamamos el "refuerzo" del par que es mas alto cuando mas común es usar una palabra en conjunto con la otra.

La idea del refuerzo es la de poder definir una negatividad contextual de un par de palabras. Si se tiene por ejemplo la palabra "persona", esta no necesariamente es una palabra negativa. Pero si "persona" se la acompaña con "muerta", no solo tendremos la negatividad intrínseca a la palabra "muerta", sino que parte de la negatividad de "muerta" se transferirá (por el contexto o el refuerzo) a la palabra "persona".

De esta forma se define la negatividad contextual asociada al par (persona,muerta) como el refuerzo de ese par dividido la suma de los refuerzos de persona con las demás palabras con las que se relaciona. De alguna forma estamos hablando de P(persona / muerta)P(muerta)

Para conocer la negatividad intrínseca de un token simplemente se toma la P(desastre / se usa esa palabra). Es decir que la negatividad de la palabra "terremoto" es la cantidad que se uso "terremoto" en tweets de desastres dividido la cantidad de tweets de desastres.

Esta forma de asignar la negatividad intrínseca es mas objetiva que una simple asignación a ojo y es escalable a datasets mas grandes y nuevos.

Finalmente la negatividad de un tweet es la suma de la negatividad intrínseca de los tokens que lo componen mas suma de las negatividades contextuales de cada par de tokens en ese texto.

5.5 Visualizando los resultados de la negatividad

Se presenta ahora un plot que muestra la relación entre el desastre o no de un tweet y su negatividad. Se busca con esto poner a prueba la métrica nuestra de negatividad.

Figure 15

Cada barra es la P(desastre / negatividad = n). Es decir que si tomamos un tweet cualquiera cuya negatividad es de por ejemplo 27, entonces se quiere saber cual es la probabilidad de que con ese nivel de negatividad se trate de un desastre. Eso es lo que representaría la barra asociada al 27.

Antes de comentar nada sobre este plot se puede apreciar un comportamiento sospechoso en los valores altos de negatividad. Es raro que en un delta chico de negatividad la función pase de probabilidad 1 a 0 dramáticamente. En la siguiente subseccion se analiza esta causa.

5.6 La ecuación mas peligrosa de la historia

Se sabe de la historia (y de la teórica) que el sacar conclusiones de conjuntos de datos "chicos" es riesgoso. Esto es conocido como no hacer uso de "La ecuación mas peligrosa de la historia" que habla de la confianza que uno puede tener sobre una conclusión basado en el tamaño de los datos.

Analizamos ahora la distribución en frecuencias de los tweets del data set para inspeccionar si los valores anómalos de probabilidad observados en la sección anterior no corresponden a set de datos "chicos".

5.6.1 Que es chico?

Se podrían tomar varios criterios, pero nosotros nos quedamos con la de decir que una clase de datos es chica si su cardinalidad es inferior que la raíz cuadrada del total de datos en el dataset.

Figure 16: La linea roja marca el valor de la raíz cuadrada de los datos y por tanto, el nivel de truncamiento

Como se puede ver, la mayoría de los datos se ubican en clases de negatividad menores a la 60. Esto quiere decir que todas las clases de negatividad que superen a 60 serán consideradas "chicas" y por tanto poco confiables. Por esta razón se decide descartar los valores mayores a 60 y volver a realizar el plot anterior.

5.7 Conclusiones de la métrica de Negatividad

Figure 17: La linea roja marca ahora el valor de 1/2 de probabilidad.

Ahora si, una vez filtrados las clases con pocos datos, se dejan de ver comportamientos anómalos y ya se puede apreciar el comportamiento de la negatividad con la probabilidad de desastre de los tweets.

La razón por la que se miran probabilidades es porque se entiende que ninguna herramienta o métrica es capaz de determinar con cien por ciento de certeza si un tweet es desastre o no por lo que se mira a niveles de probabilidad su efectividad.

Se puede ver que esta métrica artesanal de negatividad es monótona creciente con la "desastrocidad" de los tweets, es decir, cuando mas alto da la negatividad, mas probable es que se trate de un desastre.

Esto muestra que esta negatividad puede ser utilizada como indicador y potencialmente de predictora de desastre de un tweet basándose en su contenido.

Particularmente puede verse como a partir del nivel 20 de negatividad la probabilidad de desastre supera el 50 por ciento bruscamente. Uno podría clasificar un tweet como desastre o no desastre usando a este valor como medida.

5.8 La métrica de la Longitud

La definición de esta métrica ya fue introducida al comienzo de este capitulo así que ahora solo nos vamos a limitar a realizar un análisis de distribución en frecuencias y probabilidad condicional por clases de longitud, tal como se hizo con la métrica de Negatividad, para evaluar su efectividad.

5.9 Distribución de longitudes

Figure 18: La linea roja marca el valor de la raíz cuadrada de los datos y por tanto, el nivel de truncamiento

De este gráfico se puede ver que la mayoría de los datos superan el valor de truncamiento. Esto quiere decir que casi todas las clases son "confiables", o dicho de otra forma, no son chicas.

5.10 Efectividad de la Longitud como medición del desastre

Figure 19: La linea roja marca ahora el valor de 1/2 de probabilidad.

Acá se extrae una observación interesante. Ninguna longitud es particularmente propia de los tweets de desastre. Es decir, no aparenta haber alguna longitud (salvo la de 140 caracteres) que haga que al presentarse un tweet con esta longitud, uno pueda afirmar que es probable que se trate de un desastre.

Es mas, si esta métrica indicase algo, indicaría un no desastre. Es decir que si nos aparece un tweet de menos de 70 caracteres probablemente sea un tweet de no desastre.

5.11 Combinando métricas. Longitud y Negatividad

Dado el éxito de la Negatividad y el fracaso para medir desastres (pero el éxito para medir no desastres) de la Longitud, nos preguntamos que pasaría si mezcláramos estas métricas en una sola combinada.

De esta forma nuestras clases ya no serian tweets con longitud l o negatividad n, sino tweets con negatividad n y longitud l a la vez. De este modo cada clase queda identificada con un par ordenado en un plano y la efectividad seria la probabilidad condicional asociada que formaría un campo escalar.

5.12 Preliminares

Primero se muestra una relación mas modesta entre estas dos métricas para poder saber que esperar de la relación explicada en la sección anterior.

Figure 20

Se presenta entonces unos plots en formato de Facet Grid que muestran la proporciones de tweets de cierta longitud que usan o no palabras negativas y si son o no desastres.

En este caso la noción de negatividad esta relajada dado que se considera negativo o no al tweet si usa o no alguna palabra negativa, donde en este caso solo se considera negativa a una palabra si pertenece a un conjunto de palabras que bajamos de internet.

5.13 Distribución de tweets en Longitud y Negatividad

Tal como se dijo antes, ahora las clases de datos forman regiones en el plano identificadas por un par (n,l), por lo que visualizar la distribución requiere otro tipo de plots.

Figure 21

Las regiones mas oscuras corresponden a zonas mas pobladas por los tweets. Por ejemplo, del plot se puede ver que la mayoría de los tweets tienen una negatividad de al rededor de entre 0 y 10 (osea no desastres) y con una distribución mas o menos uniforme en longitud.

Esto es útil para luego poder interpretar las conclusiones que saquemos sobre la efectividad de combinar Negatividad y Longitud como métrica del desastre y no caer en el mal uso de "la ecuación mas peligrosa de la historia". Nuevamente, las zonas mas pobladas son las mas confiables.

5.14 Efectividad de la combinación Longitud y Negatividad

Figure 22

Se pueden ver valores altos de probabilidad en la derecha superior del gráfico. Lamentable mete esas zonas no son las mas pobladas por los datos por lo que el nivel de confianza en esta probabilidad es bajo.

Si hubiese que arriesgar con quedarse con alguna zona, la mas prometedora es el centro. Esto es porque el centro posee un poblamiento de los datos moderado y corresponde a una zona de probabilidad elevada.

Esto significaría que un tweet con una longitud de entre 125 y 50 caracteres y de una negatividad de entre 20 y 40 puntos tiene con cierta confianza moderada, una probabilidad significativa de ser un desastre.

En el plot de distribución puede verse una región aislada de concentración de datos en la región derecha superior de los datos correspondiente a los 140 caracteres y a una negatividad de aproximadamente 50 puntos. Sobre esta región la probabilidad de desastre puede verse cercana a 1. Esto nos recuerda al pico de probabilidad observado en el plot 19 donde se tenia que los tweets de 140 caracteres tenían una chance mayor de ser desastres que el resto.

Además es notable de observar de nuevo la monotonía de crecimiento en probabilidad en el eje de la negatividad y la uniformidad de la probabilidad en el eje de la longitud.

También puede notarse que en las regiones de negatividad baja hay también una probabilidad baja (y uniforme) de ser un desastre. Esto es consistente con el plot 19 y el 17.

La dificultad de asignar una métrica confiable radica en parte sobre el hecho de que la longitud en si no es una buena métrica.

5.15 Conclusión de la Longitud como métrica del desastre

Visto la uniformidad de la distribución en probabilidad a lo largo de las clases de longitud y habiendo visto como se distribuyen las probabilidades cuando se tiene en cuenta la negatividad. Podemos concluir que la longitud no solo no sirve para distinguir si un tweet es desastre o no, sino que tampoco sirve para distinguir si es un no desastre o no.

Esto ultimo quedo evidenciado al ver el plot 22, en donde se vio que la probabilidad de desastre crecía solo cuando crecía la negatividad y era uniforme dado una negatividad fija.

La probabilidad de desastre parece depender solo de la negatividad al comparar estas dos métricas.

5.16 La Importancia de un tweet como métrica del desastre

Ahora se analizará la importancia del tweet y su efectividad para ver si un tweet es o no un desastre.

La motivación para incluir a la importancia como una métrica "natural" o "intuitiva" es porque pensamos que para nosotros un desastre es algo negativo pero también es algo importante. No solo lo que es negativo es un desastre, de lo contrario cualquier cosa negativa (sin importar lo trivial que fuese) seria un desastre.

Definimos la importancia de un tweet como la cantidad de caracteres "!", "?" y mayúsculas que usa dividido la cantidad de tokens en el tweet.

Esta ultima división se realiza para no favorecer en importancia a los tweets largos solo por ser largos.

5.17 Distribución de la Importancia de los tweets

Como siempre y para realizar un análisis confiable primero verificamos cuales son las zonas de baja frecuencia o lo que llamamos las clases "chicas".

Figure 23

Nuevamente observamos que la mayoría de los tweets se distribuyen en importancias bajas por lo que solo consideraremos confiables a conclusiones sacadas con tweets cuyo nivel de Importancia sea 1.2 o menos.

Se puede ver además que tenemos muchos tweets con baja Importancia (de hecho es la mayoría). Todavía no sabemos si la Importancia esta bien definida como tal, que constituye una importancia baja o alta o si es una buena métrica del desastre, pero en caso de serlo podríamos explicar esta observación argumentando que hay mas tweets de no desastres que de desastres por lo que de ahí se podría originar la falta de Importancia de la mayoría.

5.18 Efectividad de la Importancia como métrica del desastre

Figure 24

Aquí nuevamente se nos muestra la probabilidad de que un tweet con cierto nivel de importancia i sea un desastre.

Lo que se concluye de este plot es que la probabilidad es siempre inferior a 1/2 y que por mas de tener un máximo en el centro, no presenta mucha varianza, por lo que pareciera que la Importancia como métrica no importaría para determinar si un tweet es desastre o no.

5.19 Combinando métricas. Importancia y Negatividad

Nuevamente decidimos combinar métricas. En este caso repetimos el uso de la métrica de Negatividad porque viene siendo la que mas da en el blanco y porque la combinación Negatividad-Importancia nos parece la semánticamente mas razonable.

El análisis a realizar es análogo a la otra combinación realizada antes. Pares (n,i) corresponden ahora a clases de datos y cada clase tiene asociada una frecuencia (que le da una confiabilidad a las conclusiones sacadas sobre ella) y una probabilidad de desastre.

5.20 Distribución de tweets en Importancia y Negatividad

Figure 25

En este caso es interesante como la mayoría de los datos se distribuyen en baja negatividad (en una asociada al no desastre) y en la baja importancia. Esto da cuenta de que quizás Importancia y Negatividad estén hablando de lo mismo. Es decir, si tanto Importancia y Negatividad cuando son bajas, el desastre tiene a ser nulo, esto habla de una relación de monotonía creciente.

5.21 Efectividad de la combinación Importancia y Negatividad

Figure 26

Nuevamente (y ahora mas que antes) este gráfico confirma la sospecha mencionada en la sección anterior. Efectivamente tanto la Importancia como la Negatividad son buenas métricas de si un tweet es desastre o no.

La justificación es una monotonía creciente en Negatividad pero también un ligera monotonía creciente en Importancia y todo esto sobre la región de mayor población de los datos.

5.22 Conclusión de la Importancia como métrica del desastre

Contrario a lo que se había percibido al principio del análisis y a la luz de estos últimos plots, se concluye una leve esperanza en Importancia como una métrica útil para medir la "desastrocidad" de un tweet.

Cabe también destacar que existe terreno para la mejora de esta métrica, en cuanto como definirla para extraer el máximo jugo de lo que uno se refiere cuando habla de que algo es importante. Por ejemplo en ningún momento hacemos referencia a las palabras usadas o a la cantidad de palabras o a la cantidad de sustantivos o nombres, etc.

Lo que seguro queda claro, en caso de que la Importancia no haya sido convincente como métrica, es que una vez mas la Negatividad es claramente una buena métrica y este ultimo plot lo vuelve a validar.

6 Locación

En esta sección, queremos determinar características asociadas a la ubicación desde donde se generó el tweet.

6.1 ¿La cantidad de tweets por continente es igual?

Queremos conocer la cantidad de tweets que hay por continente.

Figure 27: Cantidad de tweets por continente

Podemos observar que hay una notable cantidad de tweets del set de datos que provienen de América del Norte, por lo que al analizar desastres debemos considerar que habrá un sesgo en cuanto a que serán desastres que ocurren principalmente en América del Norte, y no son una representación de los desastres globales.

6.2 ¿Cómo se distribuyen los desastres reales por continente?

Buscamos conocer cuál es la distribución de los tweets con desastres reales, divididos por continente, de manera proporcional.

Figure 28: Distribución de tweets con desastres reales, por continente

Rápidamente se puede ver que los tweets de la Antártica son en un 100% desastres reales, pero esto se debe a que son únicamente dos tweets, por lo que no resultan de valor para un análisis estadístico. Podemos observar que los tweets de Asia y África tienen una mayor proporción de tweets con desastre a los tweets de América.

6.3 ¿Cuantas palabras claves hay por continente?

Queremos observar la cantidad de palabras claves distintas que tiene cada continente.

Figure 29: Cantidad de palabras claves por continente

Como se puede apreciar en la gráfica, vemos que América del Norte cuenta con la mayor cantidad de palabras claves seguido de Europa y Asia.

6.4 ¿Cual es la palabra clave mas popular en cada continente?

Queremos observar cual es la palabra clave mas popular en cada continente

Figure 30: Palabra clave mas popular en cada continente

Podemos ver cuales son las distintas palabras claves mas populares en cada continente, como por ejemplo, tanto en Norte América y Sur América la palabra clave mas popular es "collision"

6.5 Cantidad de tweets por país

Queremos conocer la cantidad de tweets que tiene cada país.

Figure 31: Cantidad de tweets por país

Realizamos un gráfico con colores variados para poder apreciar las diferentes cantidades de tweets. Podemos observar que en casi ningún país fuera de Norte América supera los 200 tweets, mientras que en Estados Unidos hay más del cuádruple de tweets que en cualquier otro país.

6.6 ¿Qué países tienen mayor concentración de tweets con desastres?

Buscamos conocer los países que tienen la mayor concentración porcentual de tweets con desastres.

Figure 32: Distribución de los tweets con desastre por país

Podemos observar que hay países pequeños con un 100% de tweets con desastres, esto seguramente se deba a que tienen unos pocos tweets los cuales contienen desastres. También podemos ver que, aunque Estados Unidos tiene la mayor cantidad de tweets, aproximadamente la mitad no tienen desastre, una proporción similar a la Argentina.

6.7 ¿Cuales son los países con menos palabras claves?

Queremos observar cuales son los países con menor cantidad de palabras claves mayor a 1, es decir, queremos descartar aquellos que pudieran ser atípicos.

Figure 33: Países con menor cantidad de palabras claves

Como podemos observar, se aprecian una gran cantidad de países con frecuencias bajas.

6.8 ¿Cómo se distribuye la longitud de los tweets por país?

Queremos conocer la distribución de la longitud de los tweets en los países que tienen más tweets.

Figure 34: Distribución de la longitud de los tweets en los 5 países con más tweets

En el gráfico podemos ver que en los 5 países que concentran la mayor cantidad de tweets, la media de la longitud de los mismos está entre 100 y 125 caracteres.

6.9 ¿Qué ciudades tienen la mayor cantidad de tweets?

Buscamos saber qué cantidad de tweets provienen de cada ciudad del set de datos.

Figure 35: Cantidad de tweets por ciudad

Podemos observar que hay una gran cantidad de tweets provenientes de las costas de Estados Unidos, una explicación posible es que están las ciudades con mayor densidad de población como Los Ángeles en la costa oeste y Nueva York en la costa Este.

6.10 Ciudades Con mayor proporción de desastres

Queremos conocer qué ciudades concentran la mayor proporción de tweets con desastres reales.

Figure 36: Proporción de desastres por ciudad

En el gráfico podemos ver que en la Antártica no hay ciudades con tweets, esto se debe a que en el set de datos no debe estar informada la ciudad. Y podemos ver que las ciudades con pocos tweets (la cantidad de tweets depende del tamaño del círculo) tienen 100% de tweets con desastre, esto se debe a que es probable que de pocos tweets todos tengan desastre o ninguno tenga. Considerar que las ciudades pequeñas tienen el 100% de los tweets con desastres sería un error de interpretación de los datos.

7 Tops

En esta sección se busca abordar los tweets desde la perspectiva de la popularidad, es decir, aquellos que contienen la mayor o la menor cantidad de eventos.

7.1 ¿Cuales son los Hashtags mas utilizados?

Buscamos resaltar aquellos hashtags que han sido utilizados en mayor medida por los usuarios de Twitter, por lo que se tendrán en cuenta todos los mensajes en los que se haya usado el símbolo "#".

Figure 37: Frecuencia de #Hashtags.

Se puede aprecia a simple vista que sobresale el hashtag "news" por sobre todos los demas, siendo este el mas utilizado en los tweets, luego le siguen "hot","prebreak" y "best" con un uso similar entre ellas pero notoriamente menor a "news".

7.2 ¿Cual es la frecuencia de los hashtags mas usados si llevamos a su palabra raiz?

Buscamos apreciar que sucede con la frecuencia de los hashtags sobre el total de los tweets si reducimos la palabra a su origen, tal como vemos en la siguiente gráfica del logo de Twitter.

Figure 38: Frecuencia de hashtags llevados a su origen.

Vemos con claridad que hay ciertos hashtags que continuan sobresaliendo sobre el resto en gran medida, lo cual puede ayudarnos al momento de detectar un evento ya que hashtags que son variaciones de una misma palabra raiz se sumarán, mejorando así las posibilidades de aislar ciertos patrones.

7.3 ¿Cuales son los Hashtags 15 hashtags mas usados?

Veremos cuales son aquellos hashtags mas usados, pero con la finalidad de agrupar aquellos que su significado es similar, hemos decidido llevar la palabra a su origen, logrando asíobservar la utilizacion de los hashtags mas frecuente.

Figure 39: Los 15 hashtags mas populares.

En la imagen, se puede observar que la palabra "new" es la mas popular, pero la frecuencia de los restantes hashtags se ha incrementado considerablemente al hacer uso de reducir las palabras a su palabra raiz. al analizar el significado de las palabras vemos que la mayoria de las mismas se encuentran contenidas en un pequeño grupo de eventos los cuales pueden estar relacionadas a desastres reales.

7.4 ¿Cuales son los Hashtags 15 hashtags menos usados?

Observaremos los hashtag con menor uso de todos los tweets, para este análisis nos quedaremos con los hashtags que se hayan utilizado mas de una vez, a fin de descartar aquellos hashtags que solo aparecen una vez.

Figure 40: Los 15 hashtags menos populares.

En la imagen, se puede observar que la palabra que las palabras tienen una frecuencia de uso simili, donde se pueden observar alguna palabras palabras muy variadas relacionadas a catástrofes, ciudades, o personas, entre otros. Sin un marcado patrón.

7.5 ¿Como se encuentran distribuidos los hashtags en el léxico?

Al agrupar todos los tweets nos concentramos en ver como los 15 hashtags mas populares se encuentran distribuidos en el léxico armado por las palabras de los tweets encadenadas una detrás de otra llevadas a su termino raiz.

Figure 41: Distribución de los 15 hashtags mas populares en el léxico.

Como era de esperarse, vamos a encontrar concentraciones en la dispersión de los hashtags mas usados, dado que ante un evento, es esperable que las personas utilicen dicho hashtag, por lo que nos define eventos a considerar al observar la creciente utilización de dichos hashtags.

7.6 ¿Cual es la frase mas frecuente?

Buscamos aquella frases que son mas frecuentes, se tomo como medida de longitud de la frase la media de la cantidad de palabras(17 palabras).

Figure 42: Frases mas frecuentes.

En este caso se aprecia que las frases mas populares corresponden a cuando el primer ministro de Malasia confirma que los restos pertenecen a MH370.

8 Curiosidades

En esta seccion, se estaran dando datos curiosos, que a lo mejor no aporten informacion importante, que se fue viendo en el proceso del Analisis Exploratorio.

8.1 ¿Existen tweests Repetidos?

Al analizar los tweets, no a profundidad, es decir no viendo que palabra, contexto, etc tiene el tweet, sino fijandose estrictamente en que el tweet sea el mismo en alguna de las 7500 filas del set, se pudo observar que existen una pequeña cantidad que son repetidos.

Hay 110 tweets repetidos en total con relacion al tweets en si(texto), separados en 69 tweets.
Es decir:

- tweet x_1 se repite y_1 veces
- ....
- ....
- tweet x_{69} se repite y_{69} veces

con:
$\sum_{i=1}^{69} x_i = 69$
$\sum_{i=1}^{69} y_i = 110$

Para simplificar, solo se mostraran los 10 mas repetidos, ya que los menos repetidos son los que tienen 2 repeticiones y son mayoria.

Figure 43: Tweets mas recurrentes.

Puede observarse,hay un tweets que se repite bastante, unas 10 veces.

El tweet completo es:
11-Year-Old Boy Charged With Manslaughter of Toddler: Report: An 11-year-old boy has been charged with manslaughter over the fatal sh...

No sabria decir si es por retweetear que se repite o si es por que basicamente todos pusieron lo mismo, ya que del dataset no se puede averiguar eso.

8.2 Palabra mas comun en keywords

Analizando las palabras de los keywords, no frase por frase (que serian los keywords en si), sino las palabras que forman esa frase.
Por ejemplo, unos keyword serian:
'fire', 'fire truck', 'first responders'

Se analizaria las palabras distintas, si los keywords fuesen un solo texto:
"fire fire truck first responders"

En ese contexto se pudo armar el siguiente WordCloud.

Figure 44: Palabras en Keywords.

A primera vista pareceria que emergency o fire son los keywords mas frecuentes, pero no es asi. Lo que se muestra, como dije antes, son las palabras de una concatenacion de todos los keywords.
Al parecer tanto "emergency" o "fire" son las palbras mas usadas, ya que varias keywords contienen esta palabra en si misma. Como por ejemplo:

para emergency: (chemical-emergency, emergency, emergency-plan, emergency-services, radiation-emergency)

para fire: (fire, fire-truck, forest-fire, wildfire)

No es tanto que exista una keyword muy usada (Nan o Nothing es la keyword mas usada) que contenga esa palabra, sino que existen unas cuantas que quiza lo tengan pocas veces pero al ser muchas keywords terminan siendo mayoritario en apariciones.

8.3 ¿El tweet menciona al keyword?

En este analisis no existe mucha explicacion, simplemente se busco si en alguna parte del texto del tweet se mencionaba alguna vez al keyword por el que fue tageado.

Figure 45: Keyword en texto.

Como pequeña conclusion,en 900 de las 7613 (el 12% del total) tweets el keyword no esta en el texto y los restantes 6700 (88% del total) contiene en alguna parte del texto a la keyword.
Entre los 900 podrian entrar los que no tenian keyword(NaN)

8.4 Tweets que contienen alguna URL en el texto

En esta seccion se filtrara y trabajara solo con tweets que tenga alguna direccion URL en el texto, sea www.xxxx o http://xxxx.

8.4.1 Cantidad de Tweets con URL

Para empezar, lo mas simple a ver es que cantidad de tweets del total del dataset tiene URL's.
En la siguiente Grafica se muestra esto.

Figure 46: Tweets con URL.

Como se puede observar, un poco mas de la mitad de todos los tweets tienen una URL, por lo que como es una cantidad bastante alta de tweets, vale la pena analizarlos un poo mas a fondo.

8.4.2 Target en tweets con URL

Sabiendo que la cantidad de tweets es suficiente como para obtener algun patron o conclusion, se procedio a analizar el target, es decir si e desastre (1) o si no lo es (0), en los tweets con URL.

Figure 47: Target en tweets con URL.

De la figura, se puede sacar que de los tweets que tiene URL's suelen ser poco mas de la mitad de target 1 (desatre) con 54,7% y el resto de target 0 (no desastre) con 45,3%. Mientras que los que no tiene URL son mayoria de target 0 (no desatres) con 70% y el resto de target 1 (desastres) con 30%.
Se puede obtener una idea de que si no tiene url es mas probable que no sea desastre (target 0). En cambio si tiene URL, no se puede estar muy seguro (50/50) y se deberia analizar otra cosa, como palabras en columna text del dataset.

8.4.3 ¿Como se distribuyen los tweets con URL en las keywords?

Este analisis quiza sea el que menos datos aporte, ya que la distribucion cambia cada vez que sesgo el dataset, en este caso me centre en los tweets con URL pero si me hubise centrado en todos los tweets, la distribucion seria distinta (no sesgaria nada). A pesar de esto, el analisis es interesante de ver.
Basicamente lo que se trato de hacer fue contar los tweets que tiene cada keyword y agruparlos por cantidad para poder encontrar su distribucion aproximada.

Por ejemplo:

$Keyword_1$,$Keyword_2$,$Keyword_3$ tiene 10 tweets.

$Keyword_4$,$Keyword_5$,$Keyword_6$ tiene 20 tweets.

Figure 48: Distribucion de cantidad tweets en keywords(tweets con URL)

Aca se puede ver la distribucion de cantidad de tweets por keyword es masomenos uniforme entre 20 y 35.
En el rango que abarca(de 1 a 36) hay al menos un keyword que tiene esa cantidad de tweets.

Ahora, realizando el mismo analisis, pero esta vez al dataset entero se puede notar como cambia la distribucion.

Figure 49: Distribucion de cantidad tweets en keywords(todo el dataset)

Se puede observar que los keywords tienen una cantidad de tweets, en su mayoria, de 30 a 40 y es mas pronunciada la diferencia con respecto a los tweets solo con URL, por ejemplo de 45 a 60 ningun keyword tiene esa cantidad de tweets.

8.5 ¿Los tweets con preguntas tienen desastres reales?

Se quiere saber si los tweets que contienen preguntas pueden hablar sobre desastres reales, o si son meramente personas preguntando algo, y en caso de querer predecir si un tweet es sobre un desastre real o no, poder decidir si se descartan los tweets con preguntas o no. Consideramos un tweet con pregunta a cualquier tweet que tenga al menos una palabra en inglés y un signo de pregunta después de esta.

Figure 50: Densidad de los tweets con preguntas según tratan sobre desastres reales o no.

En el gráfico se observa que hay preguntas con desastre, por lo que no se pueden descartar las preguntas al realizar un análisis de desastre sobre de los tweets, y que las preguntas sin desastre tienen una notoria concentración en los tweets con una longitud cercana a los 150 caracteres.

9 Conclusión

9.1 Conclusiones del análisis

Luego de analizar las características del set de datos, contamos con una perspectiva mas amplia y vemos que hay muchos factores a tener en cuenta a la hora de definir si un tweet se corresponde con un desastre real o no, las mismas las podemos representar en los siguientes puntos:

9.2 Conclusiones del análisis de Target

- Los keywords **"wreckage", "debris" y "derailment"** son los mas probables a ser Desastres.
- Por contraparte, el keyword **"aftershock"** es mas improbable a ser Desastre.
- El keyword mas usado, tanto condicionado a solo Desatres como en el set completo, es **"el no usar keyword"**, que paradojicamente es una Keyword(NaN, renombrada a Nothing).
- Los Keywords en tweets en Desastres que le siguen es frecuencia a **Nothing/NaN** son *wreckage*y *outbreak* ya que son palabras sinonimos de Desastre, destruccion, etc.

9.3 Conclusiones del análisis de Usuarios

- Se aprecian que la cantidad de menciones a usuarios representa el 26.21% del total de los tweets, separándolos en tweets de desastres reales, vemos aun que dicho porcentaje es aun menor.
- Podemos ver los tweets en esencia mencionan entre 1 y 3 usuarios, con alguna salvedad de casos que contienen hasta 8.
- Se observa que la plataforma "youtube" es el usuario mas mencionado tanto en casos tanto en desastres y no desastres.
- Los tweets que tienen entre 1 y 2 menciones a usuarios cuentan con un promedio mayor de palabras en los casos de desastres reales, mientras que 3 o mas menciones a usuarios es mayor el promedio de palabras de tweets de no desastres.

9.4 Conclusiones del Lenguaje de los Tweets

- Podemos observar que la gran mayoría de los tweets están en ingles o contienen mas del 50% de sus palabras en dicho idioma, quedando por fuera del lenguaje solo un 0.02% del total de los tweets.

- Se aprecia que la mayor concentración de tweets acerca de desastre contienen un porcentaje de palabras en ingles entre el 60% y el 80% lo cual nos puede dar un indicio de que una persona bajo una situación de stress puede tipear de manera incorrecta con mas frecuencia que cuando no se quiere comunicar un desastre.

9.5 Conclusiones finales de la Longitud, Negatividad e Importancia

- La Negatividad fue muy bien definida y tuvo buenísimos resultados.
- La Longitud resulto ser completamente indiferente a la "desastrocidad" de un tweet.
- La Importancia tiene buena perspectiva de servir y tiene espacio para ser mejorada pero no convence del todo.
- Las conclusiones sobre cada métrica tienen cierto nivel de confianza dado que se cuido de hacerse sobre conjuntos de datos no muy chicos.
- Hay cierto sesgo en los datos y las conclusiones que se hacen porque hay mas tweets de no desastres que sobre desastres.

9.6 Conclusiones del análisis de Locación

- Se puede apreciar que Norte América concentra la mayor cantidad de tweets, quien también concentra la mayor cantidad de palabras claves distintas.
- Se aprecia que a pesar de existir una notoria diferencia en cuanto a la cantidad de tweets por país, esta diferencia no se ve reflejada en términos de las palabra clave mas popular, por lo que se encuentra muy diversificados la representación de los tweets a través de las palabras claves.
- Podemos apreciar no hay una marcada diferencia en la longitud de los tweets entre los países con mas tweets, pero si la distribución de los mismos es muy diferente en Estados Unidos, quien es el que mas tweets posee, y los demás países
- Se aprecia una gran concentración de tweets de las ciudades con mayor densidad poblacional de los Estados Unidos.

9.7 Conclusiones del análisis de Tops

- Se pueden apreciar que las palabras pueden tener muchas variaciones pero que en el fondo representan lo mismo, ya sea porque esta escrita en plural, en pasado o cualquier otro tipo de conjugación.
- Se observa que al procesar y llevar las palabras a un mismo significado crecen considerablemente la frecuencia de los términos.

- Los hashtags mas populares, tanto los primeros como los últimos, tienen similitud entre si pero con una marcada diferencia entre ambos grupos..
- Como era de esperarse se tiene una gran cantidad de hashtags poco populares.
- Se puede apreciar que en algunos casos, los hashtags, aparecen como oleadas, por lo que nos pueden dar indicios de eventos.

9.8 Conclusiones del análisis de Curiosidades

- Existen muy poco tweets repetidos, siendo el de mayor frecuencia aquel que contiene 10 repeticiones.
- Se aprecia que en la gran mayoria de los tweets (88%) la palabra clave se encuentra en el texto del tweet.
- Mas de la mitad de los tweets tiene una URL.
- En una primera aproximación, se puede observar que la mayoría de los tweets que son desastres contienen una URL, lo que nos lleva a pensar que se está compartiendo alguna pagina donde se encuentra algún desastre real.
- Vemos que los tweets de preguntas no referidos a desastres tienen una gran concentración en longitudes cercanas a los 150 caracteres.

10 Bibliografía y Referencias

- Organización de Datos. Apunte del Curso. October 19, 2018v2.0
- Pandas Cookbook: Recipes for Scientific Computing, Time Series Analysis and Data Visualization Using Python Book by Theodore Petrou v1.0.1