Taller integrado de ciencias de datos

Proyecto 1 – Sebastián Farías

Tareas a realizar durante cada fase de un PCD

Identificar el problema:

El problema, en términos generales, consta de encontrar una respuesta a el comportamiento de tweets referentes a sentimientos y clima.

El desafío es analizar tweets y determinar si tienen un sentimiento positivo, negativo o neutral; si el clima ocurrió en el pasado, presente o futuro; y qué tipo de clima hace referencia el tweet.

Fuente de datos:

Cada tweet es revisado por múltiples calificadores, y se espera cierta cantidad de desacuerdo en las etiquetas. El puntaje de confianza explica dos factores: la mezcla de etiquetas que los evaluadores dieron un tweet y la confianza individual de cada evaluador. Como algunos evaluadores son más precisos que otros (por ejemplo, prestan más atención, se toman el trabajo más en serio, etc.), estos evaluadores cuentan más en el puntaje de confianza. Tiene más confianza en que un tweet se refería al pasado si confía en la persona que le está diciendo.

Recopilar datos:

En Kaggle obtenemos los datos para trabajar los cuales son train.csv y test.csv

train.csv: tweets con ubicación de estado y locación, además de su respectiva calificación.

test.csv: tweets con ubicación de estado y locación.

Preparar datos:

Integrar

Transformar

Limpiar

Filtrar

Agregar

Herramientas de procesado de datos a utilizar

Rstudio proporciona diferentes herramientas para procesar los datos, como es el caso de janitor.

Ventajas:

tabyl(): entrega el número y porcentaje de apariciones de un mismo valor en una columna.

levels(): entrega los datos almacenados según su valor, al igual que tabyl(), pero sin su número de repeticiones no porcentaje.

replace(): modificar información del dataset.

subset(): elimina o guarda datos con restricciones deseadas.

Rstudio permite visualizar el dataset en una ventana de tipo excel de forma completa.

Desventajas:

tabyl(): no permite conocer el id correspondiente a cada caso.

levels(): no nos deja conocer información de identificadores de posición de las consultas.

replace(): permite modificar uno por uno y no un grupo.

Rstudio nos permite ver sólo los primeros 1000 elementos de nuestras consultas, además, al visualizar el dataset completo en una ventana no podemos editarlo directamente.

Dataset

Instancias: 77946

Atributos: 28

Tipo de datos: data.frame

numeric

factor

valores perdidos: datos con locación vacía.

valores con ruido: datos con caracteres no correspondientes en tweet, estado o locación.

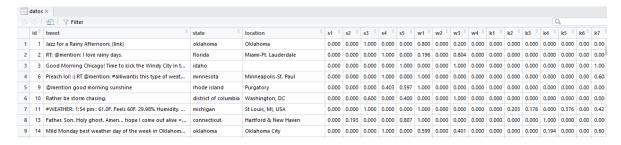
outliers: datos con ubicación correspondiente a una locación fuera de Estados Unidos.

Limpieza de datos

Carga de datos en R:

```
datos <- read.csv(file = "train.csv", header = TRUE, sep = ",")
```

Visualización del dataset:



Visualización alternativa:

```
> head(datos)
id
1 1
2 2
3 3
4 6
5 9
6 10
```

```
tweet
1
    Jazz for a Rainy Afternoon: {link}
2
       RT: @mention: I love rainy days.
                                            Good Morning Chicago! Time to kick the Win
dy City in the nuts and head back West!
4 Preach lol! :) RT @mention: #alliwantis this type of weather all the time.. I live for
 beautiful days like this! #minneapolis
         @mention good morning sunshine
6
              Rather be storm chasing.
                                  location s1 s2 s3
                                                        s4
                 state
                                                              5.5
                                                                   w1 w2
                                                                            w3 w4
                                  oklahoma 0 0 1.0 0.000 0.000 0.800 0 0.200
              oklahoma
               florida Miami-Ft. Lauderdale 0 0 0.0 1.000 0.000 0.196
                                                                       0 0.804
                                            0 0 0.0 0.000 1.000 0.000
             minnesota Minneapolis-St. Paul 0 0 0.0 1.000 0.000 1.000 0 0.000 0
          rhode island
                                           0 0 0.0 0.403 0.597 1.000
                                 Purgatory
                                                                       0.000
                            Washington, DC 0 0 0.6 0.000 0.400 0.000 0 1.000 0
6 district of columbia
                      k7 k8
  k1 k2 k3 k4 k5 k6
                            k9 k10 k11 k12
                                               k13 k14 k15
                 0 0.000 0 0.000
                                            0.000
             0
  0 0
        0 0
               0
                 0 0.000
                          0 0.000
                                        0
                                            0 0.000
                                                      0
                                                          0
                                    1
               0
                 0 1.000
                          0.000
                                    0
                                            0.000
                                                      0
                                                          0
                          0 0.196 0 0
  0 0 0 0 0 0 0.604
                                           0 0.201
                                                      O
                                                         0
                         0 0.000 0 0
                 0 0.000
                                            0 1.000
        0 0 0 0 0.000 0 0.000 0 0
                                            1 0.000
Uso de tabyl():
> tabyl(datos$location)
                             datos$location
                                                n
                                                       percent
                                            10928 1.401996e-01
                            -- bull cityyy !
                                                2 2.565879e-05
                                                1 1.282939e-05
                                --Unknown--
                               - Columbus (:
                                               1 1.282939e-05
                           - jersey city [:
                                               1 1.282939e-05
                -Behind U Wit Sum Shades On-
                                               2 2.565879e-05
                               -Houston , TX
                                               1 1.282939e-05
                          -illinois-chicago-
                                               1 1.282939e-05
                         -Phoenix; ĂZ
[chicago] â~œâ™¡â~ž
                                                1 1.282939e-05
                                                1 1.282939e-05
                                       [WV]
                                               2 2.565879e-05
                                               1 1.282939e-05
                                      4, MS
                                     72401
                                                1 1.282939e-05
> tabyl(datos$state)
          datos$state
                               percent
                        n
              alabama 1577 0.020231955
               alaska 714 0.009160188
              arizona 1575 0.020206297
             arkansas 808 0.010366151
           california 3765 0.048302671
             colorado 1712 0.021963924
          connecticut 1548 0.019859903
             delaware 749 0.009609217
 district of columbia 782 0.010032587
              florida 3659 0.046942755
              georgia 1519 0.019487851
               hawaii 765 0.009814487
                idaho 752 0.009647705
```

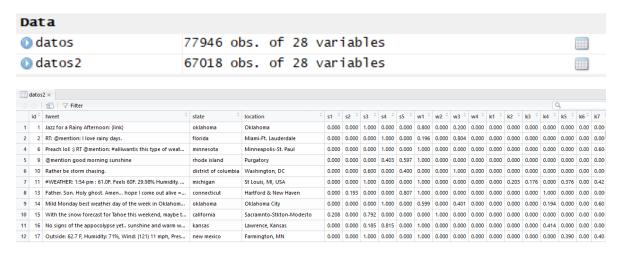
Con esto podemos concluir que existen locaciones vacías que representan el 14% del dataset, pero hay no estados vacíos.

Uso de levels():

Uso de subset() para quitar las locaciones vacías:

```
datos2 <- subset(datos, location != '')</pre>
```

Visualización de datos2:



Reemplazo de datos:

Inicial

```
Alabama 75 1.119102e-03
ALABAMA 4 5.968546e-05
Alabama 2 2.984273e-05
Alabama, Montgomery 2 2.984273e-05
Alabama, Tuscaloosa 1 1.492136e-05
Alabama, USA 1 1.492136e-05
Alabama...=) 1 1.492136e-05
Alabaster, AL 4 5.968546e-05
Alamogordo, NM 3 4.476409e-05
```

Reemplazo

```
datos2$location[datos2$location=="Alabama...=)"]<- "Alabama"
```

Final

Alabama	76 1.134024e-03	1.134058e-03
ALABAMA	4 5.968546e-05	5.968724e-05
Alabama	2 2.984273e-05	2.984362e-05
Alabama State University	1 1.492136e-05	1.492181e-05
Alabama, Montgomery	2 2.984273e-05	2.984362e-05
Alabama, Tuscaloosa	1 1.492136e-05	1.492181e-05
Alabama, USA	1 1.492136e-05	1.492181e-05
Alabama=)	0 0.000000e+00	0.000000e+00
Alabaster, AL	4 5.968546e-05	5.968724e-05
Alamogordo, NM	3 4.476409e-05	4.476543e-05