

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 procesamiento 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 ni 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:

datos																											
	id	tweet	state	location	s1	s2	s3	s4	s5	w1	w2	w3	w4	k1	k2	k3	k4	k5	k6	k7							
1	1	Jazz for a Rainy Afternoon: (link)	oklahoma	Oklahoma	0.000	0.000	1.000	0.000	0.000	0.800	0.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00							
2	2	RT: @mention: I love rainy days.	florida	Miami-Ft. Lauderdale	0.000	0.000	0.000	1.000	0.000	0.196	0.000	0.804	0.000	0.000	0.000	0.000	0.000	0.000	0.00								
3	3	Good Morning Chicago! Time to kick the Windy City in t...	idaho		0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00								
4	6	Preach lol! : RT @mention: #aliwantis this type of weat...	minnesota	Minneapolis-St. Paul	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00								
5	9	@mention good morning sunshine	rhode island	Purgatory	0.000	0.000	0.000	0.403	0.597	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00								
6	10	Rather be storm chasing.	district of columbia	Washington, DC	0.000	0.000	0.600	0.000	0.400	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00								
7	11	#WEATHER: 1:54 pm : 61.0F. Feels 60F. 29.98% Humidity: ...	michigan	St Louis, MI, USA	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.203	0.176	0.000	0.376	0.00								
8	13	Father. Son. Holy ghost. Amen... hope I come out alive =...	connecticut	Hartford & New Haven	0.000	0.193	0.000	0.000	0.807	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.00							
9	14	Mild Monday best weather day of the week in Oklahom...	oklahoma	Oklahoma City	0.000	0.000	0.000	1.000	0.000	0.599	0.000	0.401	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.00							

Visualización alternativa:

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


Uso de levels():

```
> head(levels(datos$location))
[1] ""
[3] "--Unknown--"
[5] "- jersey city [:]"
      "-- bull cityyyy !"
      "- columbus (:)"
      "-Behind U wit Sum Shades On--"
```

Uso de subset() para quitar las locaciones vacías:

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

Visualización de datos2:

Data	
▶ datos	77946 obs. of 28 variables
▶ datos2	67018 obs. of 28 variables

datos2 x	
id	tweet
1	Jazz for a Rainy Afternoon: (link)
2	RT: @mention: I love rainy days.
4	Preach lol! : RT @mention: #alliwantis this type of weat...
5	@mention good morning sunshine
6	10 Rather be storm chasing.
7	11 #WEATHER: 1:54 pm : 61.0F. Feels 60F. 29.98% Humidity. ...
8	13 Father. Son. Holy ghost. Amen... hope I come out alive #...
9	14 Mild Monday best weather day of the week in Oklahom...
10	15 With the snow forecast for Tahoe this weekend, maybe t...
11	16 No signs of the appocolypse yet... sunshine and warm w...
12	17 Outside: 62.7 F, Humidity: 71%, Wind: (121) 11 mph, Pres...

state	location	s1	s2	s3	s4	s5	w1	w2	w3	w4	k1	k2	k3	k4	k5	k6	k7
oklahoma	Oklahoma	0.000	0.000	1.000	0.000	0.000	0.800	0.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00
florida	Miami-Ft. Lauderdale	0.000	0.000	0.000	1.000	0.000	0.196	0.000	0.804	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00
minnesota	Minneapolis-St. Paul	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.60
rhode island	Purgatory	0.000	0.000	0.000	0.403	0.597	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00
district of columbia	Washington, DC	0.000	0.000	0.600	0.000	0.400	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00
michigan	St Louis, MI, USA	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.203	0.176	0.000	0.376	0.00	0.42
connecticut	Hartford & New Haven	0.000	0.193	0.000	0.000	0.807	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.00	0.00
oklahoma	Oklahoma City	0.000	0.000	0.000	1.000	0.000	0.599	0.000	0.401	0.000	0.000	0.000	0.000	0.194	0.000	0.00	0.60
california	Sacramnto-Stkton-Modesto	0.208	0.000	0.792	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00
kansas	Lawrence, Kansas	0.000	0.000	0.185	0.815	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.414	0.000	0.00	0.00
new mexico	Farmington, MN	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.390	0.00	0.40

Reemplazo de datos:

Inicial

```
Alabama 75 1.119102e-03
ALABAMA 4 5.968546e-05
Alabama 2 2.984273e-05
Alabama State University 1 1.492136e-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