Final
Tema: exploración, entrenamiento y evaluación de modelos
Fecha entrega: 9:00 pm Noviembre 19 de 2024

Punto 1 (130)

- Se tiene el siguiente dataset de canciones de reggaeton. La directora de ICBF quiere hacer un evento publico y utilizar una playlist que no contenga canciones con palabras sensibles para los niños. Ella tiene esta tabla con frases que considera sensibles y quiere que entrenemos un modelo de clasificación.

Number	Título de la Canción	Segmento de la Letra	Contenido Sensible para Niños	Conjunto de Datos
1	Tusa	Me llama en la noche, siempre con fuego y pasión	Sí	Training
2	Bichota	Camino en la noche con fuerza, pasión y fuego	Sí	Training
3	Ay, DiOs Mío!	Siento el fuego y la pasión cada noche	Sí	Training
4	Mi Cama	La noche es nuestra, encendemos la pasión y el fuego	Sí	Training
5	Punto G	En la noche, tu cuerpo y el mío sienten la pasión y el fuego	Sí	Training
6	China	La música suena en la noche y el perreo sigue con fuego	Sí	Training
7	Mi Cama Remix	El ritmo y el fuego nos encienden en la pista en la noche	Sí	Training
8	Toda de Ti	Quiero toda tu pasión, fuego y amor en la noche	Sí	Training
9	Otra Noche	Otra noche de pasión y fuego bajo las estrellas	Sí	Training
10	Por Ti	Por ti, la noche arde con pasión, amor y fuego	Sí	Testing
11	Dime Qué Quieres	Dime qué deseas y cumpliré tu pasión en la noche con fuego	Sí	Testing
12	A Solas	A solas en la noche, con pasión, amor y sin miedo	Sí	Testing
13	Provenza	Bailamos en la playa, riendo y celebrando la vida	No	Training
14	Ocean	Tu amor es un océano que me envuelve con calma	No	Training
15	El Makinón	El ritmo nos mueve bajo el sol y la luna	No	Training
16	200 Copas	Celebramos juntas, riendo y compartiendo la noche	No	Training
17	Ahora Me Llama	Libre y feliz, ahora todo es diferente	No	Training
18	Créeme	Mis promesas son sinceras y llenas de amor	No	Training
19	Secreto	Nuestro secreto es un juego inocente	No	Training
20	Culpables	Culpables de este amor, sin preocupaciones	No	Training

21	**El Barco**	Navegamos hacia nuevos comienzos, con esperanza	No	Training
22	**Leyendas**	Somos leyendas, bailando bajo las estrellas	No	Training
23	**Beautiful Boy**	Tu sonrisa es mi paz y mi alegría	No	Training
24	**Guilty**	Por amarte, me declaro culpable y feliz	No	Training
25	**Esperándote**	Te espero, sin prisa y con amor verdadero	No	Training
26	**No Quiero Más**	Risas y sueños, sin lágrimas ni miedo	No	Training
27	**Me Llama Remix**	Tu voz resuena en mi corazón con amor	No	Training
28	**Te Quiero Pa' Mí**	Te quiero conmigo, sin mentiras ni dudas	No	Testing
29	**Tus Besos**	Tus besos son mi refugio y mi felicidad	No	Testing
30	**Cuando Te Vea**	Cuando te vea, seremos felices, sin preocupaciones	No	Testing

- Exploración, crear un gráfico de barras con las 10 palabras más frecuentes. Calcular la correlación con el si es sensible. Si el contenido es sensible es 1 y sino es sensible es 0 (10 puntos)
- Crear la matriz tf-idf. Ordene las columnas por las palabras mas comunes (50 puntos)
- Utilizar la matriz tf-idf y crear un algoritmo gradiente descendiente para clasificar los textos. error = y *log (1/(1+ e(-g)) + (1-y)*log (1 - 1/(1+ e(-g)) donde g = b1 tf-idf(palabra1)+ b2 tf-idf(palabra2)+b3 tf-idf(palabra3) ….+ bn tf-idf(palabran)+b0 (80 puntos)

Hint: La derivada esta en uno de los notebook para la regresion logistica or en el siguiente blog articulo

$$\frac{\partial(J(\theta))}{\partial(\theta)} = \frac{1}{m} X^T[\ h_\theta(x) - y]$$

Basicamente el gradiente es b1 = b1 - lr * (x1* (sigmoid(x)-y))/m donde m es el numero de datos

- Escriba el algoritmo de gradiente descendiente con un learning rate de 0.4, un bache de 2 y vamos a tomar las muestras en orden (una sensible y otra no sensible). E inicializar los valores en 0 (40 puntos)

- Hacer las pruebas de escritorio para el gradiente por lo menos 6 iteraciones. Y llenar la siguiente tabla. (40 puntos)

iteracion	b1 (palabra 1 frecuente)	b2 (palabra 2 frecuente)	b3 (palabra 3 frecuente)	b0 (intercepto)
1 (observaciones 1-13)				
2 (observaciones 2-14)				
3 (observaciones 3-15)				
4 (observaciones 4-16)				
5 (observaciones 5-17)				
6 (observaciones 6-18)				
7 (observaciones 7-19)				
8 (observaciones 8-20)				
9 (observaciones 9-21)				
10 (observaciones 23-27)				

Repita el entrenamiento con los mismos datos hasta que el error en training sea pequeño.

Punto 2 (90)

Ahora utilizar los siguientes embedings a cambio de TF-ID. Estos son embedings semanticos. Se debe crear una matriz con todas las palabras y colocar estos embdeings y depues entrenar otro modelo de la misma forma que en el punto 1.

- Este modelo sera llamado mV1_0

Y consiste en utilizar un embedding semantico de la siguiente tabla. Es decir cada palabra tiene este valor y sino se encuentra toma el valor de 0:

Palabra	Embedding Promedio
noche	0.6
fuego	0.5
pasión	0.7
amor	0.8
bajo	0.0604835580197238
estrellas	0.08119550732594460
feliz	0.06699710685271450
miedo	-0.5

preocupaciones	-0.03010725441448190
quiero	-0.0029724325835689400
riendo	0.06067223051315760
ritmo	-0.06956178170621670
alegría	0.7
amarte	0.017439131381306600
arde	0.02988509981869250
bailamos	0.0028371308986503800
bailando	-0.005393766995690460
besos	-0.06317301484555990
cada	0.046576184640074500
calma	0.060401976352805600
camino	-0.014900773405456400
celebramos	-0.004649521197028000
celebrando	0.0991526383710418
comienzos	-0.059132503936802100
compartiendo	0.057208382402024400
conmigo	-0.06563344867237910
corazón	-0.06037460912757930
cuerpo	0.03525351433841200
culpable	0.06522211491767320
culpables	-0.08741284756674930
cumpliré	0.09863395498618160
declaro	-0.08765080719915800
deseas	-0.048995848460708000
diferente	-0.08754317216012630
dime	0.03606400506723460
dudas	-0.03413176324639590
encendemos	0.06915032890857260
encienden	-0.05139235111677530
envuelve	0.08827922080899680
esperanza	-0.09643029313483610
espero	0.0842241599807508
felices	-0.0856976355900374
felicidad	0.8
fuerza	-0.09755197461816010
hacia	-0.07553077491990210

inocente	0.06586253627225020
juego	-0.03958390891568330
juntas	0.03667812833050680
leyendas	0.052099991985988800
libre	-0.028429924584411400
llama	0.06844139235373310
llenas	-0.005766399688165950
luna	0.07931643025408350
lágrimas	-0.016908285090785500
mentiras	-0.07279846396634390
mis	0.09385787060953830
mueve	0.06417418444204430
mío	-0.055707430484414000
música	0.4
navegamos	-0.0958723946991972
nuestra	0.0710322620204088
nuestro	0.09147719529194070
nuevos	0.07662425881469370
océano	-0.08007088100071250
otra	-0.036508273527103400
paz	0.0428530101898115
perreo	-0.8
pista	0.037036590272521200
playa	-0.06848711312422030
prisa	-0.01544788290323850
promesas	-0.004127321941265130
qué	-0.07921250843891660
refugio	0.07982013445832780
resuena	-0.09074942660835970
risas	-0.09498927531780310
secreto	-0.037455115142717400
seremos	0.0266162439625843
siempre	0.08463242535720920
sienten	0.09205380583698690
siento	0.09509870833958930
sigue	0.0866952399183814
sinceras	0.06813143280390070

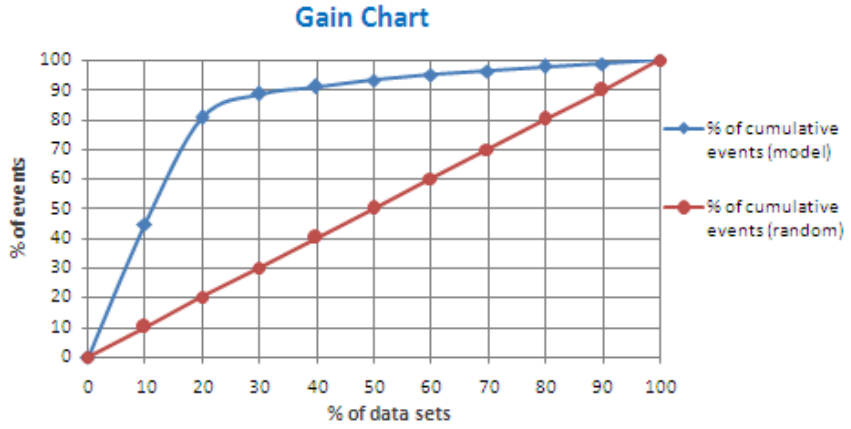
sol	0.05729862649169060
solas	-0.038203042946617800
somos	-0.06206790175816720
sonrisa	-0.09021203329893160
suena	-0.05097926073140070
sueños	0.5
ti	0.05078417215100500
toda	-0.05152616401175960
todo	-0.038062545199441900
tus	0.0557894450187974
vea	0.03694831518667350
verdadero	0.0444968680568619
vida	-0.06543174109622590
voz	-0.06365143538614950

- Cree los 2 modelos a mano y coloca cuales son los coeficientes que aprendio la regresion logistica. Si utiliza librerias se resta el 20% de la nota

Punto 3 Evaluacion, (90)

Para los 2 modelos mV2_0 y mV1_0. Calculate (40 puntos):

- f1, precision, recall y matriz de confusion para los 2 modelos (10 puntos)
- crear gain chart y medir la capacidad de predecir correctamente los mensajes sensibles para los 2 modelos. (30 puntos)



- Cual modelo es mejor y utilizando el gain chart cuantos falsos positivo tienes que tener si se quieren detectar al menos el 90% de las canciones con contenido sensible. (10)_

Punto 4 Sustentar (90),

Al finalizar el examen vamos a hacer una llamada de 5 minutos para sustentar el codigo.