Recomendador de Subreddit

Laboratorio de Datos - Primer Cuatrimestre 2021

Agustín Bayer **Francisco Valdez** **Nicole Zamonsky**

OBJETIVO

Realizar un análisis de las features de un dataset de Reddit, y emplear un algoritmo que recomiende subreddits a un usuario en base a los subreddits en los que hizo más comentarios.

Tabla de contenidos

01

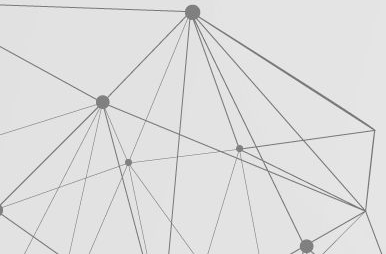
SOBRE EL DATASET

Números del dataset

19.508.509 Comentarios

69.490 Subreddits

37.845 Usuarios diferentes



SOBRE EL DATASET

El dataset consiste en 2 dataframes de 2021 sobre comentarios en Reddit: uno tiene información sobre los distintos subreddits y el otro cuantos comentarios realizaron distintos usuarios en ellos

Dataframe #1

Subreddit	#Suscriptores	+18	Descripción
ChoosingBeggars	2.134.849	No	This subreddit...
Python	809.272	No	News about...
interestingasfuck	8.092.462	No	For anything...
PublicFreakout	3.257.059	No	A subreddit...
ShitMomGroupsSay	258.681	No	Share the...

SOBRE EL DATASET

El dataset consiste en 2 dataframes de 2021 sobre comentarios en Reddit: uno tiene información sobre los distintos subreddits y el otro cuantos comentarios realizaron distintos usuarios en ellos

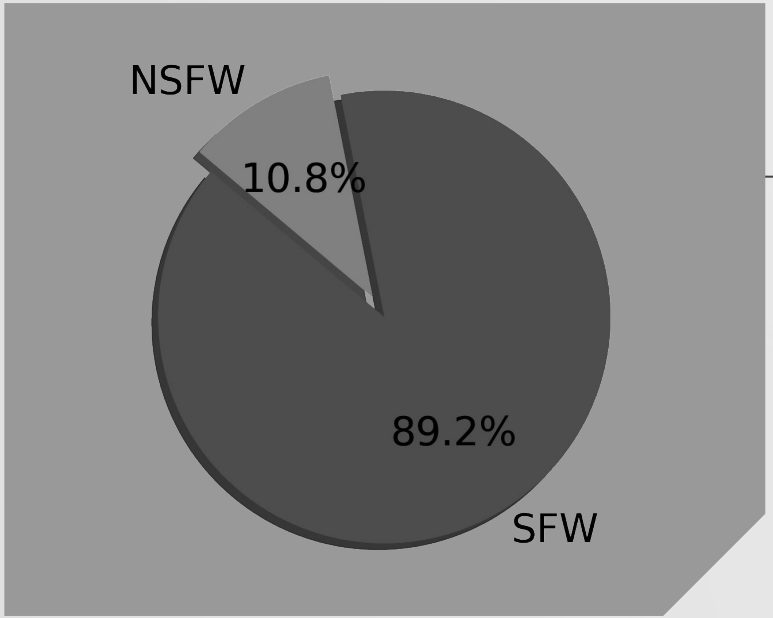
Dataframe #2

Usuario	#Subreddit	#Comentarios
Agus	AskReddit	31
Agus	Movies	1
Nicole	FIFA	17
Nicole	pics	19
C. fulanito	eggs	69

02

EXPLORANDO EL DATASET

Distribución de subreddits SFW y NSFW



La gran mayoría de de los subreddits eran SFW (51618) vs NSFW (6238).

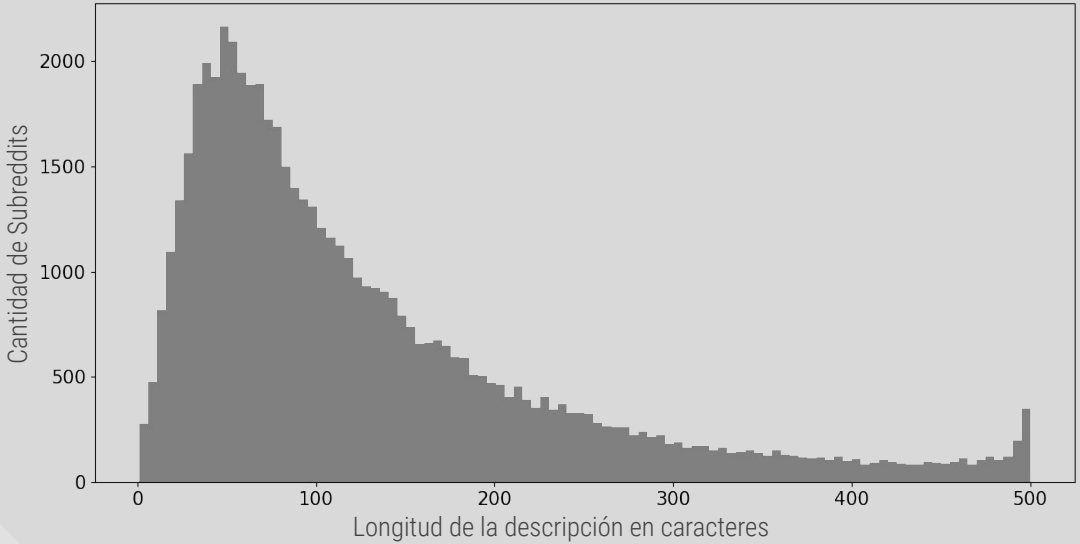
Nubes de palabras

Mostramos la nube de palabras de las descripciones de los subreddits, según la categoría de edad.

SFW

NSFW

Explorando el dataset

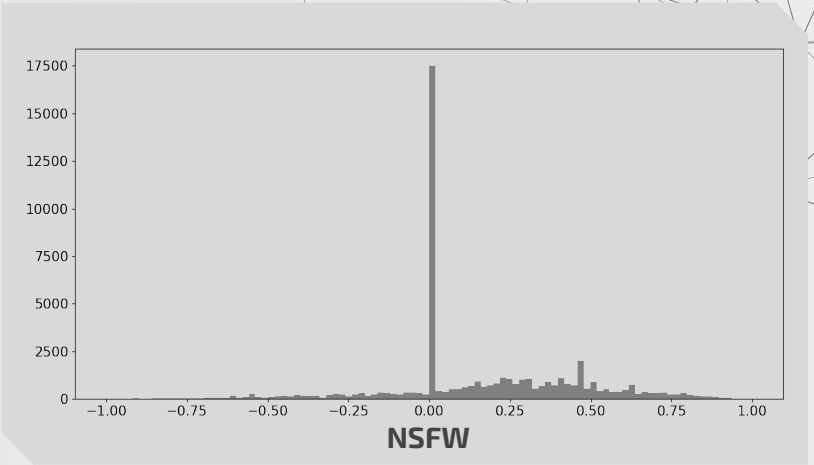


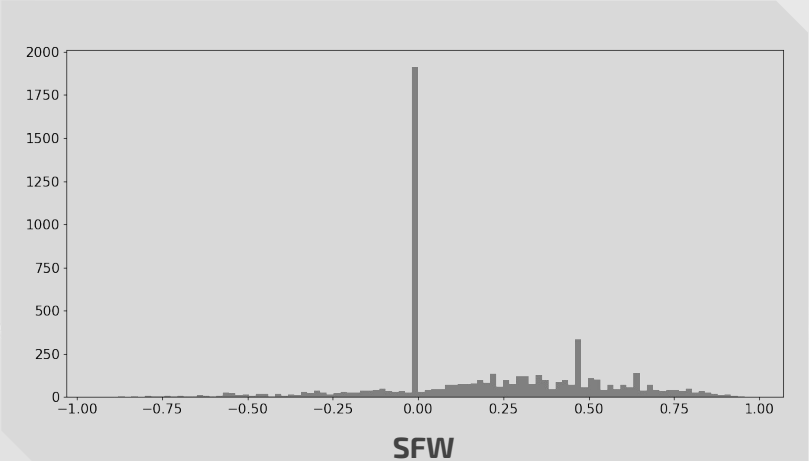
Longitud de la descripción en los distintos Subreddits

Las descripciones de los subreddits están limitadas a 500 caracteres, además vemos que las mismas suelen ser cortas (menores a 200 caracteres).

Sentiment Analysis

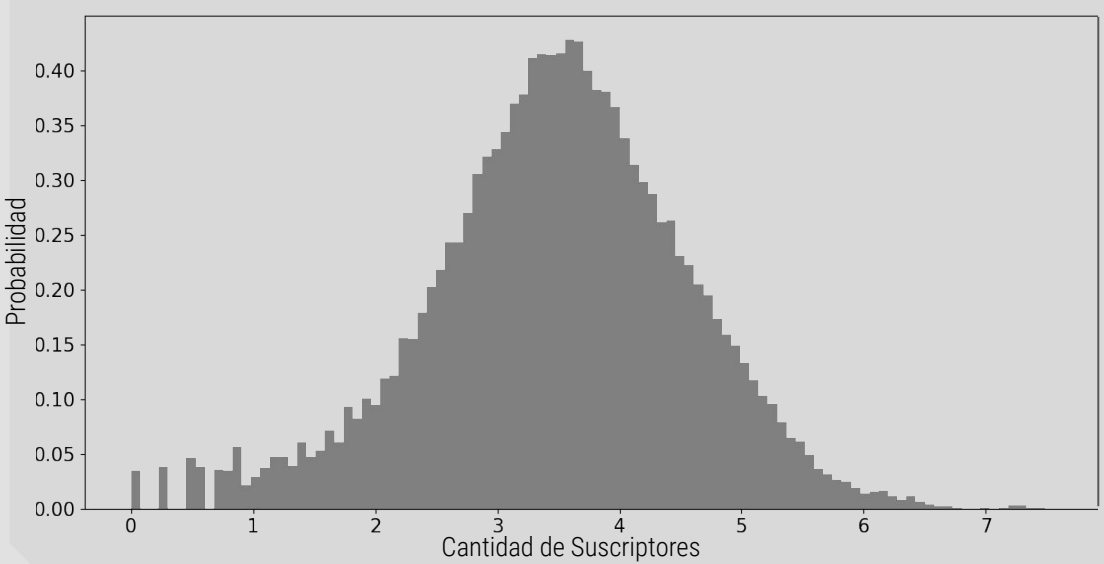
En base a estas categorías realizamos Sentiment Analysis para las descripciones de los subreddits.





Vemos que no hay diferencias significativas en el sentimiento para ambas categorías.

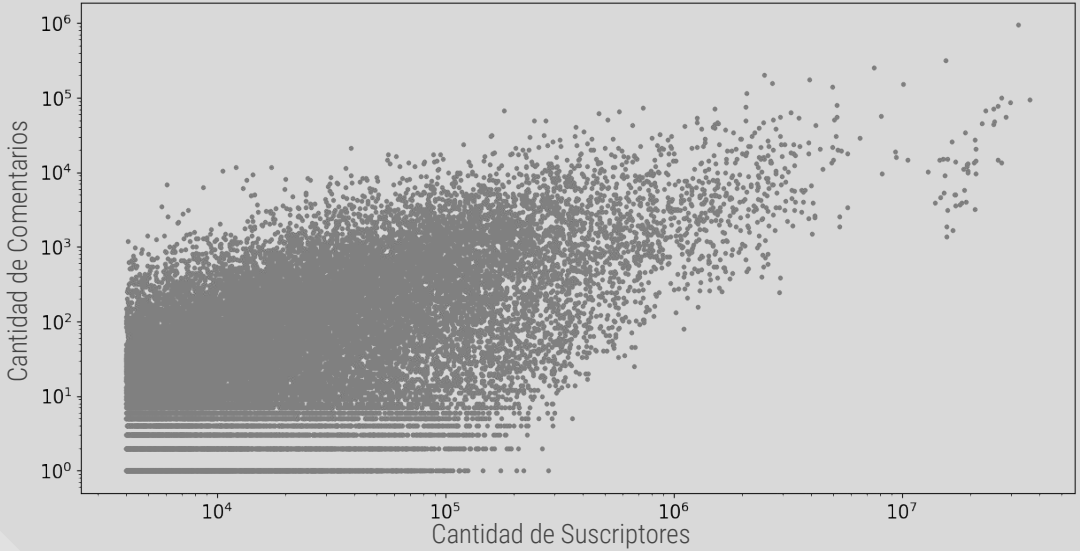
Explorando el dataset



Distribución de la cantidad de suscriptores de los subreddits

El eje x corresponde al logaritmo del número de suscriptores.

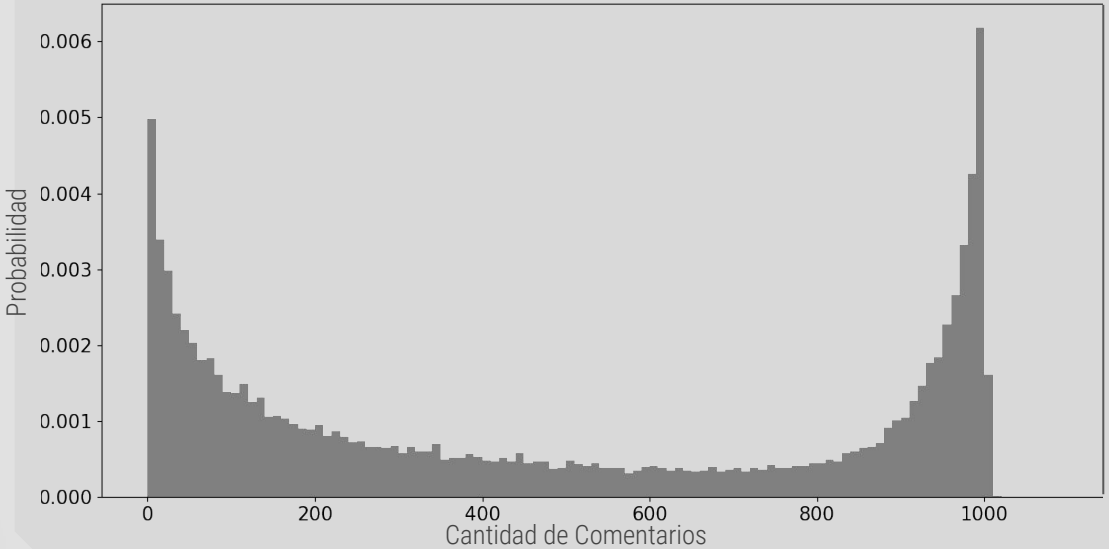
Explorando el dataset



Cantidad de comentarios vs cantidad de suscripciones

Se observa que la cantidad de comentarios tiende a aumentar con la cantidad de suscriptores.

Explorando el dataset



Distribución cantidad de comentarios

Se observa en nuestro dataset dos tipos de usuarios: Un grupo que comenta muy poco (<200 comentarios), y uno que comentó mucho (>800 comentarios).

03

PREPARACIÓN DE LOS DATOS

Para realizar el modelo

PREPARACIÓN DE LOS DATOS

MOTIVACIÓN

Queremos armar un buen modelo predictivo que recomiende subreddits de interés al usuario.

QUÉ NOTAMOS

Notamos que los subreddits con pocos suscriptores contribuían muy poco a la cantidad total de comentarios

DESCARTAR SUBREDDITS

Descartamos comentarios en subreddits con menos de 4000 suscriptores.

ARMAMOS UN DATAFRAME CON USUARIOS Y SUBREDDITS

El dataframe que armamos tenía los subreddits como filas y los usuarios como columnas, y completamos con 1 si comentó en el subreddit y 0 si no. Estos lo convertimos en una **matriz esparsa**

FORMATO DEL DATAFRAME CON EL QUE ARMAMOS LA MATRIZ

	USUARIO 1	USUARIO 2	USUARIO 3	USUARIO 4
ASKREDDIT	0	1	1	0
PYTHON	0	0	0	1
FIFA	1	1	1	0

04

MODELO KNN

PREDICCIÓN CON KNN

Utilizamos un modelo de recomendaciones basado en KNN (k=10). El modelo, dado un subreddit, arroja los subreddits con menor distancia al mismo. Para recomendarle un subreddit a un usuario, se puede elegir el subreddit donde este usuario realizó más comentarios.

Subreddit	Recomendación #1	Recomendación #2	Recomendación #3
AskReddit	pics	funny	gaming

Una versión más refinada de esto, es tomar todos los subreddits y calcular todas las distancias, y tomar peso por la cantidad de comentarios en cada uno. Con esto, devolvemos recomendaciones más heterogéneas en el caso de que un usuario haya hecho similar cantidad de comentarios en dos subreddits muy distintos.

05

Recomendaciones

Usuario 1

Usuario #1	Subreddits	Comentarios
	pennystocks	966
	MoonGangCapital	3
	wallstreetbets	3

Subreddits en los que comentó el usuario 1

Predicciones KNN, los 5 más predichos y sus distancias

	RECOMENDACIONES	DISTANCIA
#1	RobinHoodPennyStocks	0.603
#2	stokes	0.688
#3	smallstreetbets	0.735
#4	StockMarket	0.746
#5	trakstocks	0.802

Usuario 2

Usuario #2	Subreddits	Comentarios
	PewdiepieSubmissions	8
	MarioMaker	6
	TinyHouses	3

Subreddits en los que comentó el usuario 2

Predicciones KNN, los 5 más predichos y sus distancias

	RECOMENDACIONES	DISTANCIA
#1	dankmemes	0.578
#2	memes	0.600
#3	teenagers	0.619
#4	Minecraft	0.677
#5	cursedcomments	0.692

Usuario 3

Usuario #3	Subreddits	Comentarios
	geopolitics	142
	europe	135
	soccer	96

Subreddits en los que comentó el usuario 3

Predicciones KNN, los 5 más predichos y sus distancias

	RECOMENDACIONES	DISTANCIA
#1	worldnews	0.734
#2	mapporn	0.749
#3	CredibleDifference	0.771
#4	AskEurope	0.775
#5	dataisbeautiful	0.789

¿y otros métodos ?

Intentamos utilizar otros métodos: Random Forest y clasificador lineal. En ambos se usaron como features a los subreddits comentados por cada usuario, y como y_data al subreddit más comentado por estos, habiendo eliminado esta información de la matriz de entrenamiento.

- Random Forest resultó no ser efectivo, dado que, para implementar un algoritmo que pueda recomendar efectivamente a cualquier subreddit, necesitaríamos una red con una cantidad de clasificaciones igual a la cantidad de subreddits, que es de 26914, y el algoritmo no nos resultó efectivo en esta tarea.
- Para el clasificador lineal nos quedábamos sin memoria RAM, dado que la matriz de entrenamiento era muy grande dados los datos que teníamos.

06

CONCLUSIONES

CONCLUSIONES FINALES

Estadísticas del dataset

Mostramos las distribuciones características del dataset, sentiment analysis y nube de palabras.

Algoritmo de recomendación

Pudimos emplear un algoritmo recomendador de subreddits basado en KNN. Si bien no pudimos utilizar un método de validación, mostramos que en ejemplos particulares las recomendaciones son razonables.

GRACIAS

¿Preguntas?

Pueden bajar el subreddit en:
https://www.kaggle.com/timschaum/subreddit-recommender