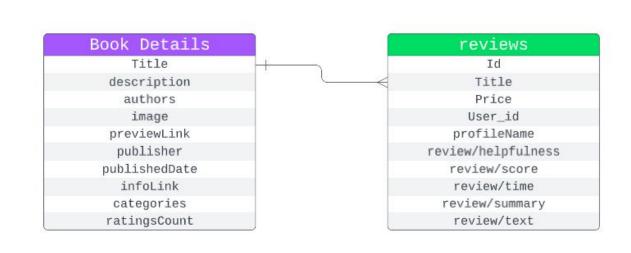
## Sistema de recomendación de libros

Luis Manuel Ambrocio Loreto



#### Dataset

- El conjunto de datos contiene reseñas de libros de amazon, este conjunto de datos contiene 2 archivos, como se ve en la siguiente figura
- Se tienen 3 millones de reviews y 212404 libros



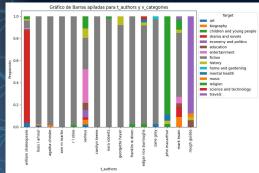


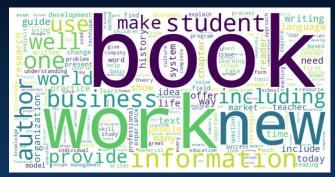
### Objetivo

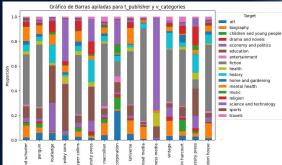
- El objetivo al final del diplomado es tener un sistema de recomendación que ayude a las personas a descubrir nuevos libros basado en sus gustos, entre la información más importante que se requiere para entrenar un sistema de recomendación se encuentra:
  - Tener información sobre calificaciones que usuarios hacen a los libros
  - Tener formas de agrupar a los libros y/o usuarios
- Los objetivos de este parte del proyecto son:
  - Tener una representación numérica de la reseña que el usuario dio a un libro
  - Tener toda la columna de la categoría del libro llena (actualmente tiene muchos datos sin información)
- Se crearán 2 modelos que nos ayuden a cumplir estos objetivos



- Se detectó que la variable de clasificación del libro (ficcion, educacion etc,) estaba muy sucia, esta variable es de mucha importancia al momento de crear recomendaciones.
- Una vez limpios los datos se detectó cerca del 12% de los datos sin información para esta variable, así
  que se decidió crear un modelo ML para imputar estos datos faltantes
- Al hacer el EDA las variables que más influyen en la clasificación son:
  - Autor
  - Editorial
  - Descripción y título del libro
  - Fecha de publicación









# +++

- Entre los modelos utilizados se encuentran:
  - Naive Bayes
  - Random Forest
  - Gradient Boosting
  - Ada Boost
  - Redes Neuronales Artificiales
- El mejor modelo fue una Red Neuronal donde se obtuvo un Accuracy de 0.71 y la arquitectura fue la siguiente:
  - La primera capa con 3 entradas diferentes:
    - Descripción tokenizada seguida de una capa embedding de dimensión 300
    - Título Tokenizado seguido de una capa de embedding dimension 300
    - Autor, editorial y fecha en One hot encoding
  - Se concatenaron las 3 entrados y se pasaron a una capa oculta con 64 unidades y función de activación ReLu
  - O Capa Dropout con parámetro 0.5 para regularización
  - Capa de salida con 17 unidades y función de activación softmax
  - Optimizador Adama con Learning rate de 0.00003

							М	atriz	de Co	nfusić	ón						
art -	808	37	35	23	12	52	13	31	3	37	33	5	5	22	29	13	22
biography -	25	336	15	9	28	14	14	29	4	60	2	17	22	25	4	10	10
children and young people -	5	8	160	1	3	20	9	58	3	7	5	12	2	5	5	1	13
drama and novels -	5	3	1	25	0	3	0	3	1	0	0	0	0	3	0	0	1
economy and politics -	13	26	5	4	410	76	0	11	20	107	1	34	6	32	32	1	17
education -	43	25	33	6	82	1464	10	26	32	43	11	71	12	54	97	20	27
entertainment -	15	4	7	3	4	4	168	19	1	0	0	4	5	3	2	3	2
fiction -	86	117	122	50	46	71	130	3598	8	112	3	73	6	67	12	21	30
Actual Actual	1	4	3	2	6	19	3	2	352	3	8	39	1	2	14	10	3
¥ history -	50	124	30	11	144	64	8	42	2	814	2	13	5	71	21	15	35
home and gardening -	17	5	6	0	4	8	0	2	4	2	344	4	0	1	8	2	7
mental health -		29	16	1	34	81	16	24	76	9	7	608	5	70	11	7	3
music -		11	3	3	0	2	1	2	0	1	0	1	180	0	0	1	0
religion -		57	32	22	44	82	7	49	8	71	6	96	10	1138	10	4	8
science and technology -		5	15	0	24	94	4	49	26	11	11	23	4	12	1023	3	13
sports -		6	9	0	4	9	5	3	5	1	1	2	0	2	1	171	6
travels -		2	5	0	5	8	3	2	0	8	3	1	1	1	6	20	165
	art	biography	children and young people	drama and novels	economy and politics	education	entertainment	fiction	health	history	home and gardening	mental health	music	religion	science and technology	sports	travels
	_			_				P	euicte	u			_			_	





- En el modelo 2 se busco predecir el Score que un usuario dio al libro a partir de su reseña.
- El objetivo es tomar la salida de la última capa oculta de una red neuronal como una representación numérica de la reseña.
- Se obtuvo un Accuracy con la siguiente arquitectura e hiperparametros
  - O Review tokenizado como entrada:
    - Vocab\_size: 200000
    - Max\_sequence\_length: 300
  - Capa embedding de dimension 300
  - Capa oculta de 64 unidades y función de activación ReLu
  - Dropout con parametro 0.4
  - Capa de salida con 5 unidades y función de activación softmax
  - Optimizador Adam con learning rate de 0.00002





- Se combinaron los score 1 y 2 como score 1 (no le gusto), el 3 como score 2 (más o menos), y el 4 y 5 como score 3(le gusto).
- Se mantuvo la misma arquitectura, solo cambiando la última capa de 5 unidades a 3 unidades, se tuvo un accuracy de 0.87

