

Introducción

Origen de la información

TED es una organización sin fines de lucro dedicada a difundir ideas, generalmente en forma de conversaciones breves y contundentes.

La organización nació en 1984 por iniciativa de Richard Saul Wurman y Harry Marks como una conferencia donde convergieron Tecnología, Entretenimiento y Diseño (de ahí las siglas TED)

Actualmente, hay más de 3000 charlas TED disponibles para ver online, las cuales cubren una amplia variedad de temas, desde ciencia hasta negocios y asuntos globales.

Planteo del problema

Regresion (Caso 1)

Entrenar y optimizar una red neuronal que sea capaz de predecir la cantidad de visualizaciones que tendrá una charla en base a la información provista en el dataset.

Descripción del dataset

Contamos con un dataset que contiene información sobre todas las charlas TED subidas al sitio web oficial hasta el 21 de septiembre de 2017.

Entre las variables, se incluyen el número de visualizaciones, el número de comentarios, descripciones, oradores y título de cada disertación.

Es un dataset de tamaño pequeño, con gran cantidad de información almacenada de forma textual.

Análisis descriptivo

Archivo: ted_main.csv

Campo: film_date, published_date

- Se extrae mes en columna y se debe transformar a dummies
- Variable por día de la semana



Campo: comments

- Se debe estandarizar el campo.
- Primero se debe realizar el 'train_test_split' y únicamente estandarizar con los datos de train.



Campo: description, title

- Se unen los campos 'description' y 'title'
- Se reemplazan (_) (—) (—) (/) por espacios
- Se reemplaza (') por (')
- Se eliminan (")
- Se expanden las 'contractions' (ej.: don't --> do not)
- Se descartan caracteres de puntuación
- Se reemplazan los términos plurales por su singular
- Se reemplaza el término por su 'lemma'
- Se vectoriza (CountVectorizer, stop_words + min_df=0.01)



Campo: duration

- Se debe estandarizar el campo.
- Primero se debe realizar el 'train_test_split' y únicamente estandarizar con los datos de train.



Campo: event

- El nombre del evento no tiene una estructura consistente.
- El nombre del evento no necesariamente indica que las charlas se realizaron un mismo día.
Ej.: Para TED2006 se realizaron 45 charlas y las mismas variaron entre 01-Feb-2006 hasta 02-Mar-2006.
- La información de Año de la charla se puede obtener del campo 'film_date'.
- Nuevas Features: 3 columnas
- 'event_TED': Empieza con TED
- 'event_TEDx': Empieza con TEDx
- 'event_noTED': No contiene TED



Campo: languages

- Se debe estandarizar el campo.
- Primero se debe realizar el 'train_test_split' y únicamente estandarizar con los datos de train.

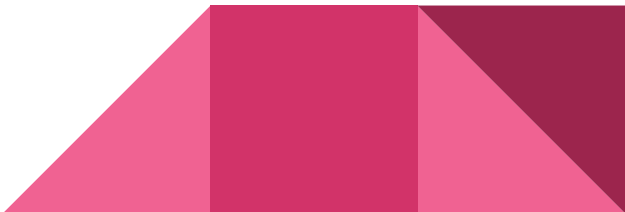


Campo: main_speaker

Son más de 2156 speakers distintos Ej.:

[1 talk, 1880] [2 talks, 202][3 talks, 48][4 talks, 16][5 talks,6][6 talks,2][7 talks, 1][9 talks, 1]

Nuevos features:

- previous_talks': Cantidad de charlas previas
 - previous_talk_views: Cantidad de views en su última charla
 - previous_views_sum: Suma de views de todas sus charlas previas
 - previous_views_max: Máxima cantidad de views en charlas previas
 - previous_views_min: Mínima cantidad de views en charlas previas
- 

Campo: name

- El nombre del speaker ya se encuentra en el campo 'main_speaker'.
- El título de charla ya se encuentra en el campo 'title'.
- Se elimina el campo.



Campo: ratings

- El campo es Texto: "[{ 'id':<...>, 'name':<...>, 'count':<...> }]"
- El campo 'name' tiene un total de 14 valores posibles.
- ["Funny", "Beautiful", "Ingenious", "Courageous", "Longwinded", "Confusing", "Informative", "Fascinating", "Unconvincing", "Persuasive", "Jaw-dropping", "OK", "Obnoxious", "Inspiring"]
- Se genera una columna por cada rating posible



Campo: speaker_occupation

- El campo tiene un total de 1459 valores distintos



Campo: tags

- El campo es Texto: "[tag_1', 'tag_2', ..., 'tag']"
- Existe un número muy alto de posibles valores.
- Se van solo a utilizar únicamente la siguiente lista (26 tags):
- ["technology", "science", "design", "business", "collaboration", "innovation", "social_change", "health", "nature", "environment", "future", "communication", "activism", "children", "personal_growth", "humanity", "society", "identity", "community", "culture", "global_issues", "entertainment", "art", "politics", "economics", "religion"]
- Se genera una columna ('tag_<...>') por cada 'tag' en la lista y una columna adicional ('tag_other') donde se suman todas las restantes.



Campos agrupados (Join)

- df_ratings
- df_tags,
- df_word_count



Campos eliminados (drops)

- film_date
- published_date
- description_title
- event
- main_speaker
- name
- related_talks
- ratings
- tags
- url
- speaker_occupation
- film_yearpublished_year



Dummies

Se crea columnas dummies con los campos:

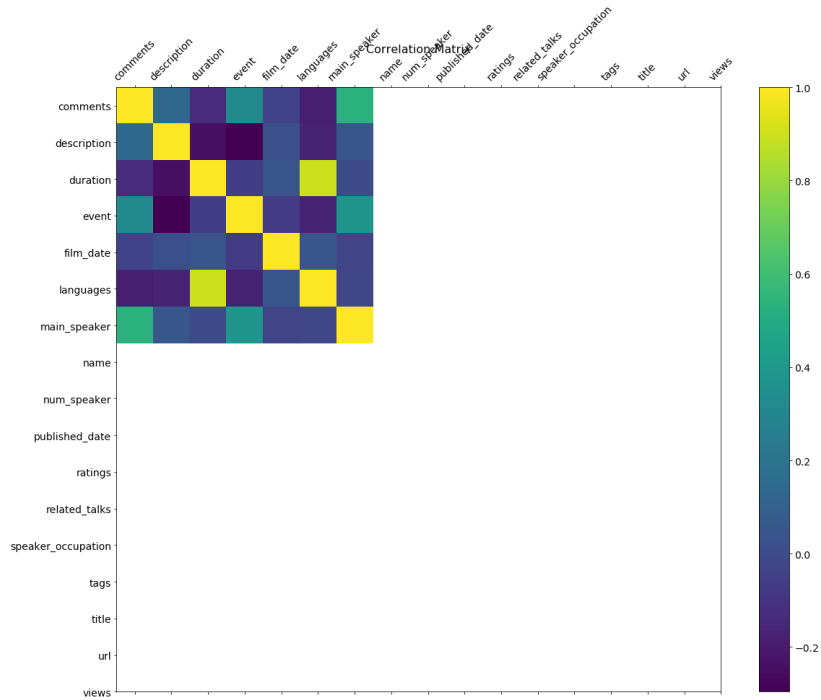
- published_month
- pilm_dayofweek





Visualizaciones preliminares

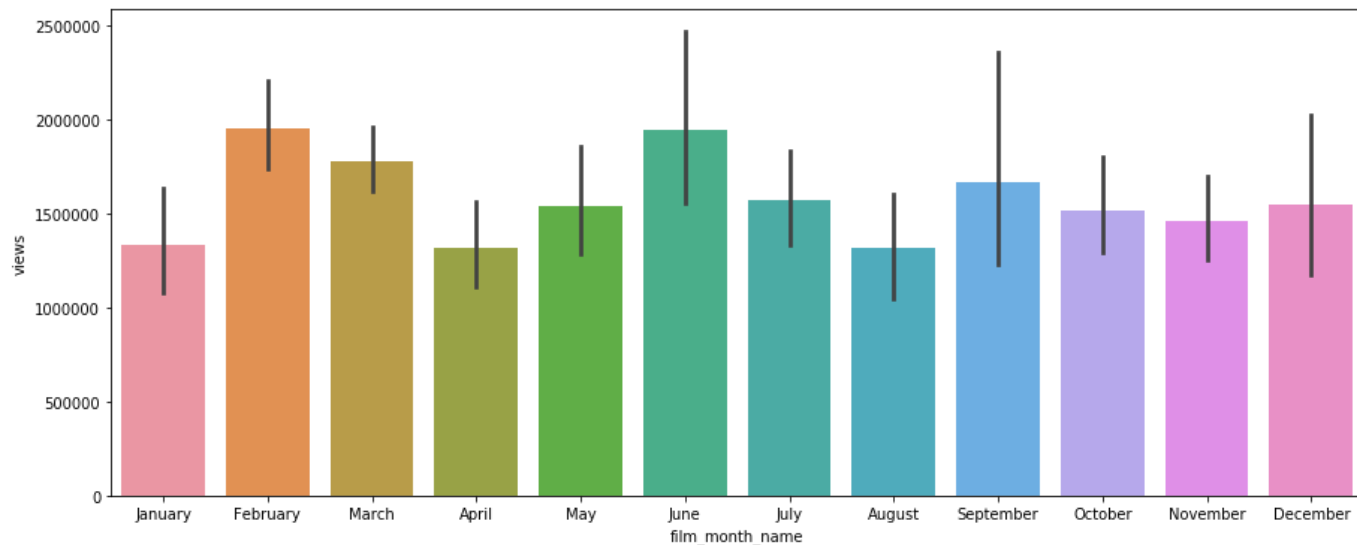
Correlaciones



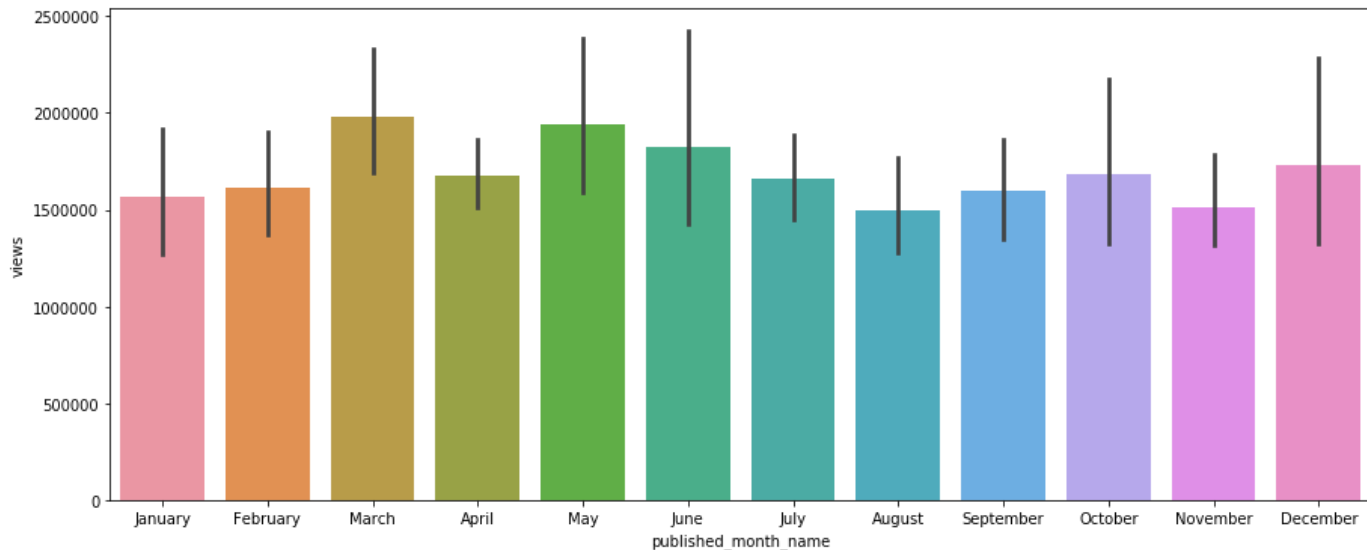
Correlaciones

	comments	duration	film_date	languages	num_speaker	published_date	views
comments	1	0.140694	-0.133303	0.318284	-0.0354889	-0.185936	0.530939
duration	0.140694	1	-0.242941	-0.295681	0.0222572	-0.166324	0.0487404
film_date	-0.133303	-0.242941	1	-0.0619566	0.0402267	0.902565	0.00644673
languages	0.318284	-0.295681	-0.0619566	1	-0.0630999	-0.171836	0.377623
num_speaker	-0.0354889	0.0222572	0.0402267	-0.0630999	1	0.0492399	-0.026389
published_date	-0.185936	-0.166324	0.902565	-0.171836	0.0492399	1	-0.0179197
views	0.530939	0.0487404	0.00644673	0.377623	-0.026389	-0.0179197	1

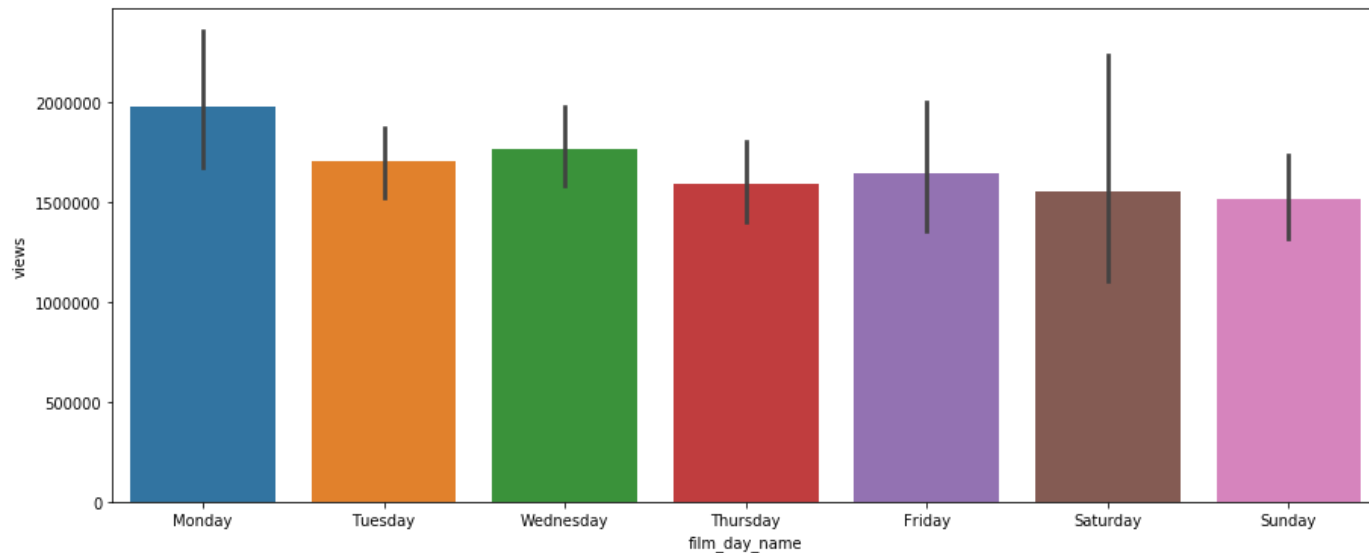
Visualizaciones por mes



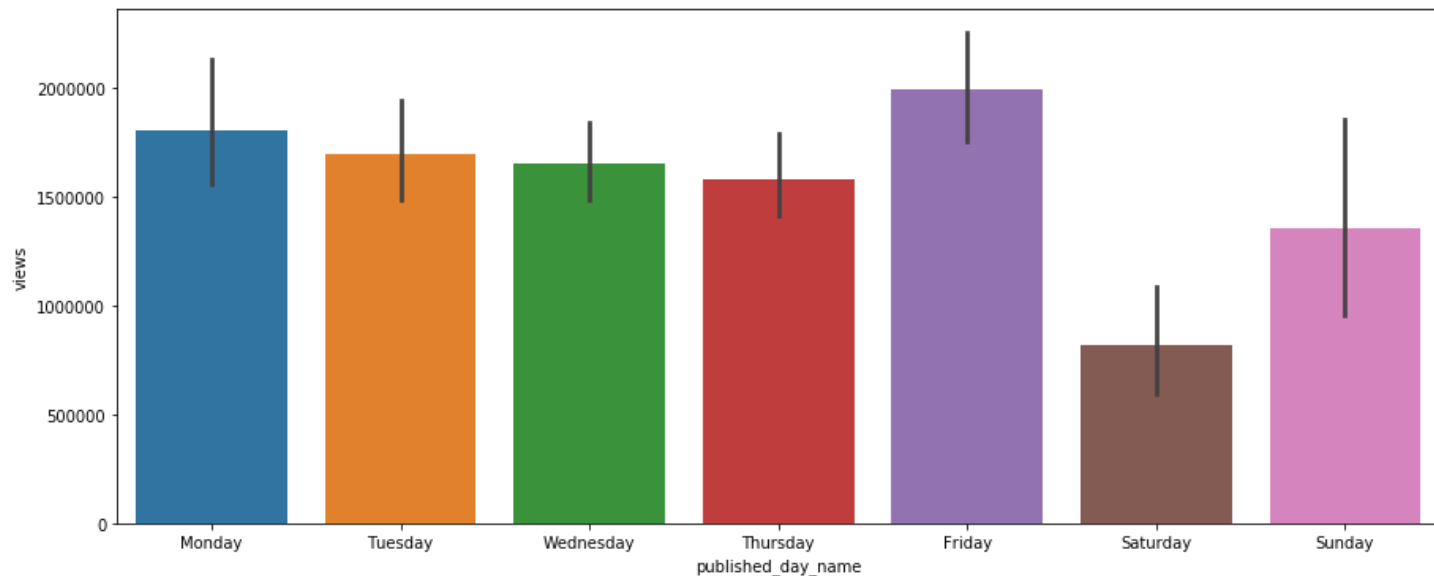
Visualizaciones por mes de publicación



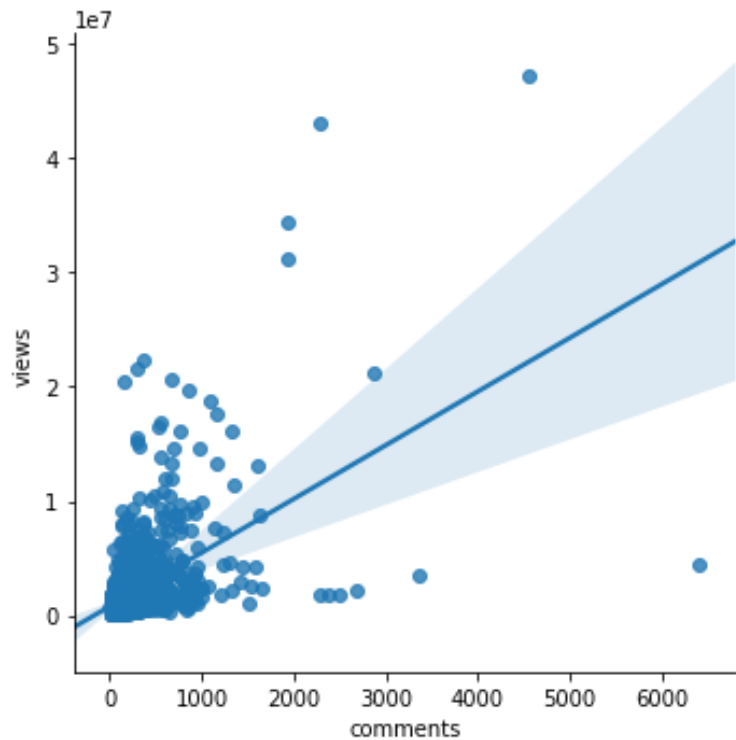
Visualizaciones por día



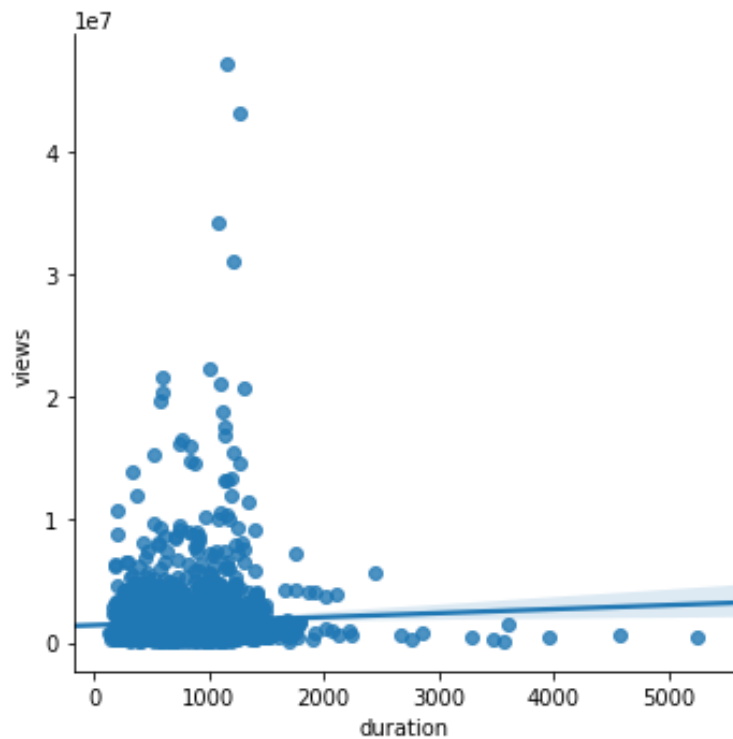
Visualizaciones por día de publicación



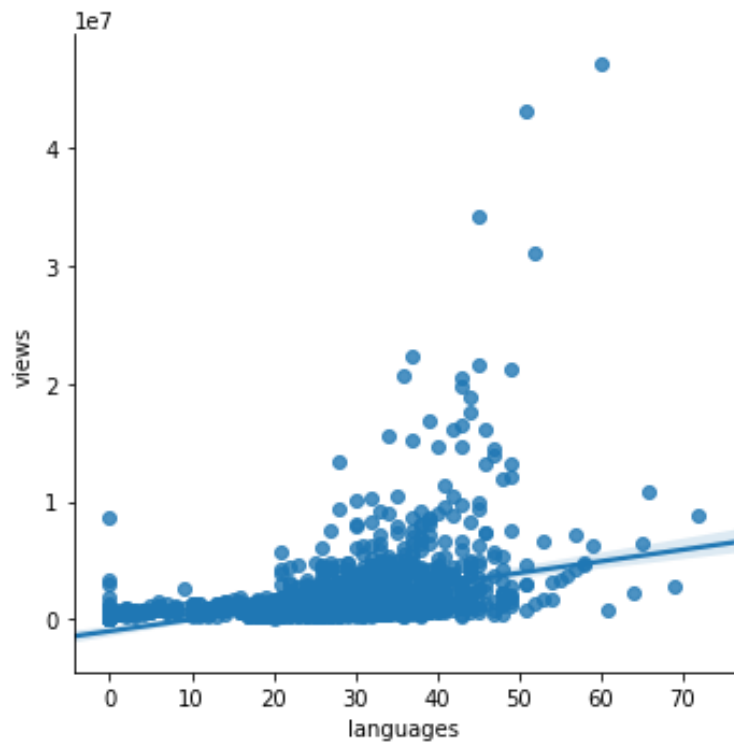
Visualizaciones por comentario



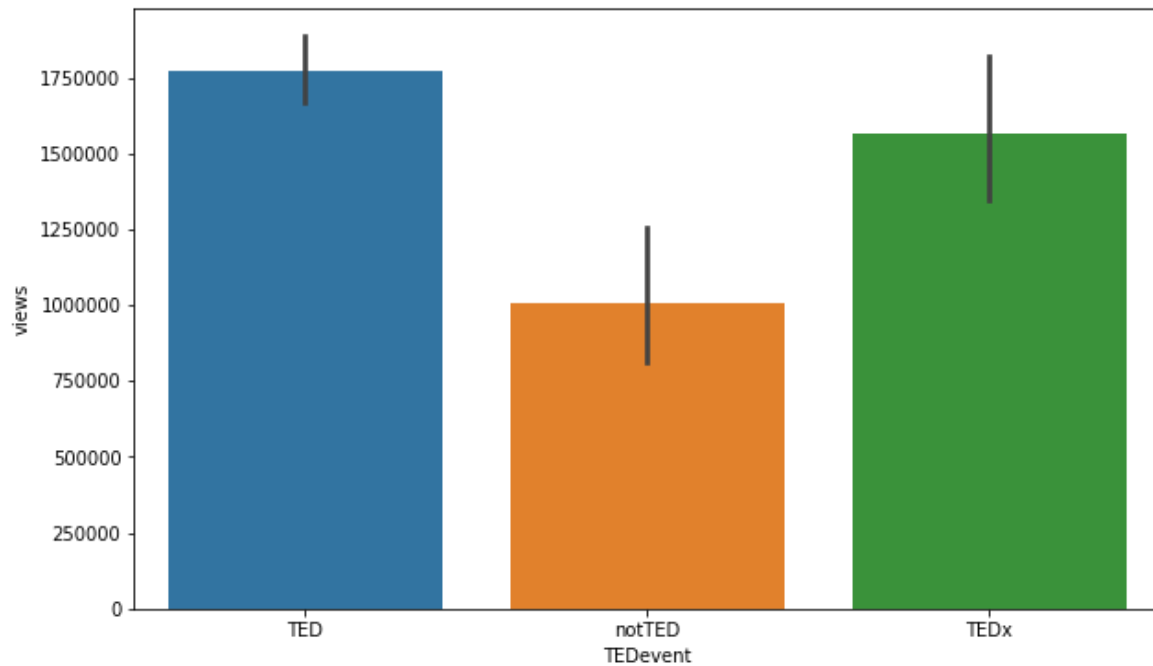
Visualizaciones por duración



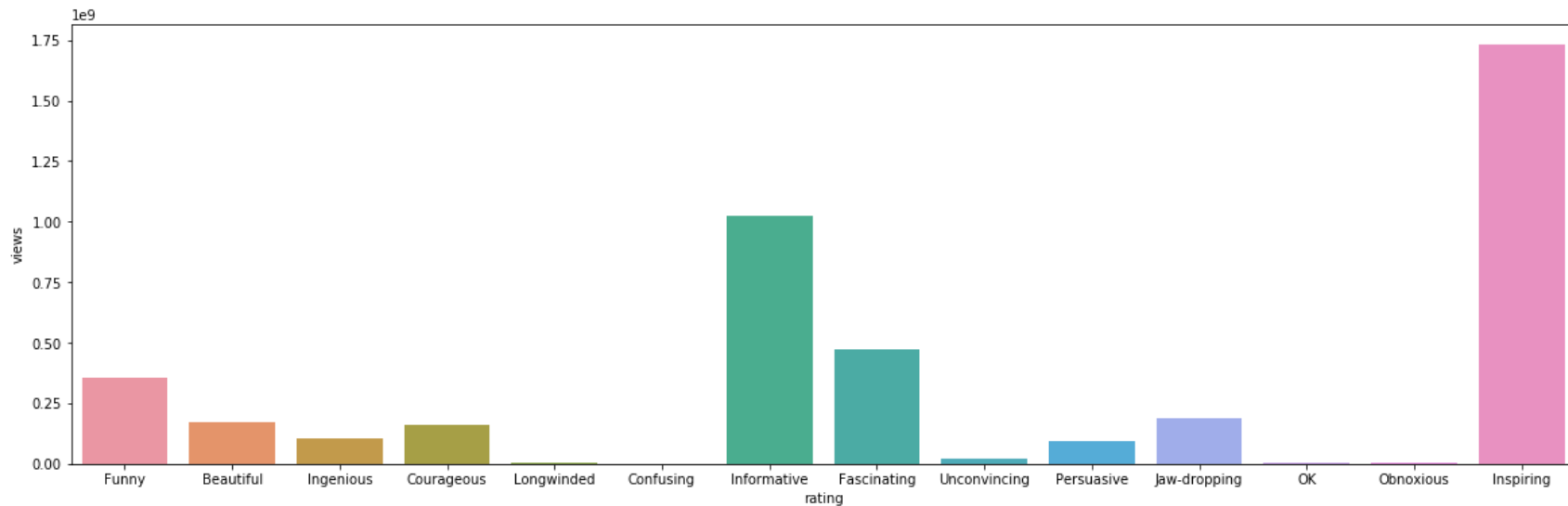
Visualizaciones por lenguaje



Visualizaciones por tipo de evento



Visualizaciones por rating



Arquitectura del modelo

- Se optó por una arquitectura de seis capas (densas) ocultas y una capa de salida.

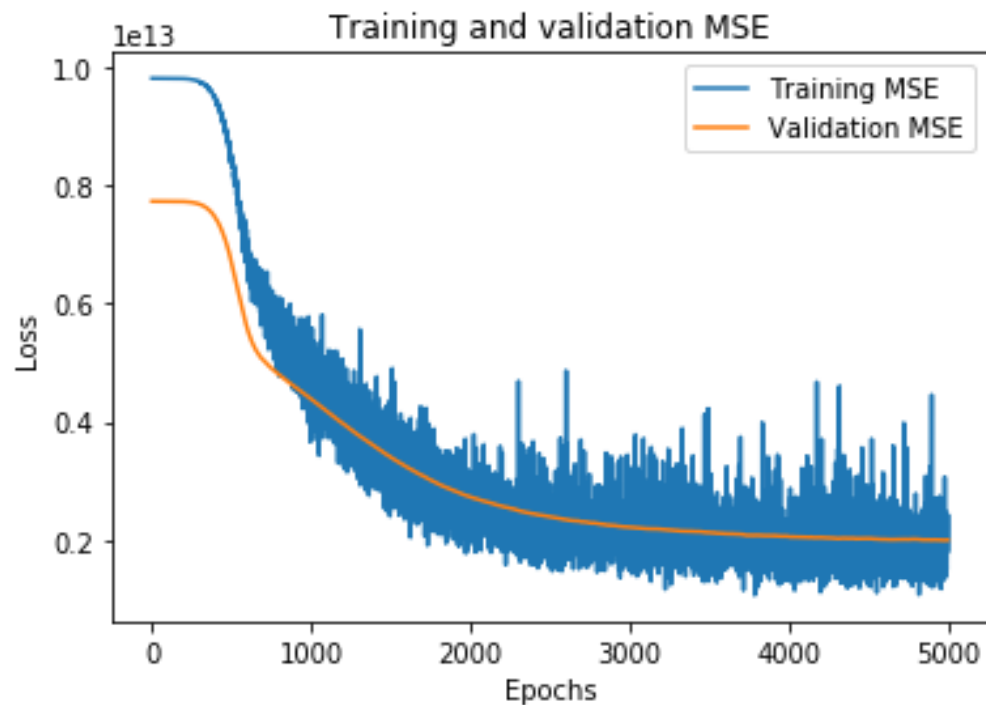
Detalle del modelo

Capa 1	512 Nodos	Activación RELU	
Capa 2	256 Nodos	Activación RELU	
Capa 3	256 Nodos	Activación RELU	DropOut 10%
Capa 4	128 Nodos	Activación RELU	DropOut 20%
Capa 5	128 Nodos	Activación RELU	DropOut 20%
Capa 6	64 Nodos	Activación RELU	DropOut 30%
Capa 7	32 Nodos	Activación RELU	DropOut 60%
Ultima Capa	1 Nodo	Sin activacion	

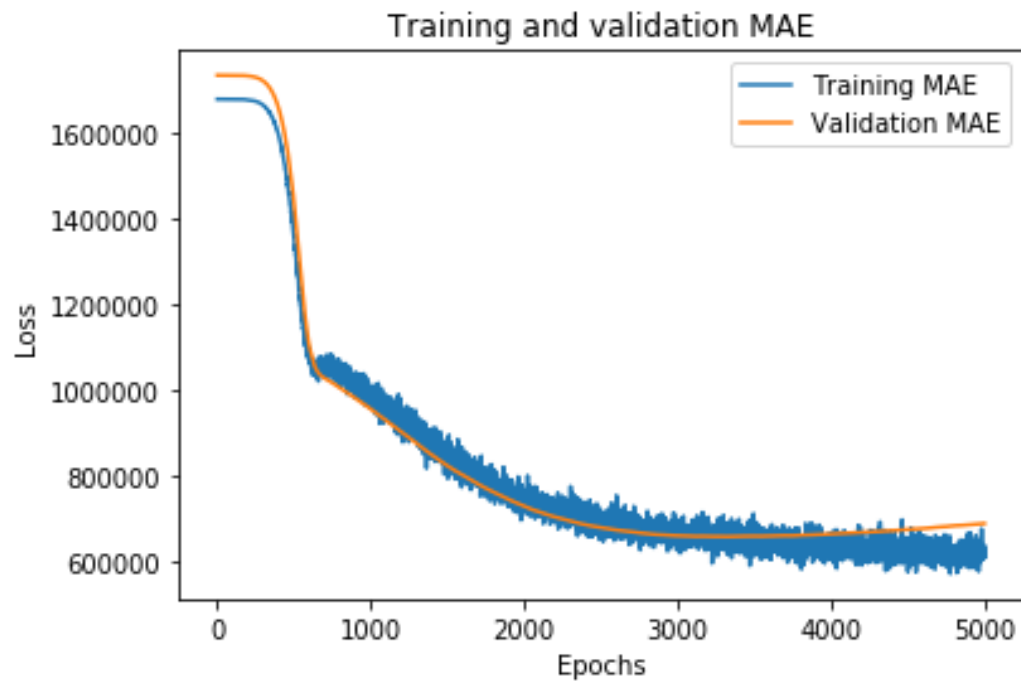
Función de pérdida	Mean squared error
Optimizador	RMSprop Learning rate = 1e-6
Métricas	Mean_squared_error r2_for_keras
Entrenamiento	batch_size = 8 epochs = 5000 validation_split = 0.3



Resultado de entrenamiento



Resultado de entrenamiento



Conclusiones

Después de haber realizado varios entrenamientos y probado con distintas configuraciones de hiper parámetros, en todos los casos obtenemos una diferencia considerable (en el rango de las 600000 visualizaciones) entre los datos de entrenamiento y validación.

Por tal motivo, concluimos que con el conjunto de datos y el análisis descriptivo realizado **no podemos predecir con una precisión aceptable** la cantidad de visualizaciones de una charla TED.

Planteo del problema

Clasificación (Caso 2)

Entrenar y optimizar una red neuronal que se capaz de predecir la categoría de una charla en base a la información provista en el dataset.

En base a análisis descriptivo anterior.

- Se eliminan a las columnas tags de los datos.
- Se toman los tags como resultados.
- Se estandarizan los datos.



Arquitectura del modelo

- Se utilizó la herramienta **Talos** para obtener la mejor combinación de hiperparámetros según la información dada.

Talos - opciones de parámetros

first_neuron	[512, 256]
hidden_layers	[2, 4, 6]
dropout	[0, 0.1, 0.25]
shapes	['brick', 'triangle']
batch_size	[64, 128]
epochs	[20, 50, 80]
activation	['relu', 'tanh']
optimizer	['Adam', 'SGD']

lr	[0.001]
last_activation	['sigmoid']
metrics	[custom_iou]
loss	['binary_crossentropy']

864 Combinaciones

Tiempo: 2h 05m



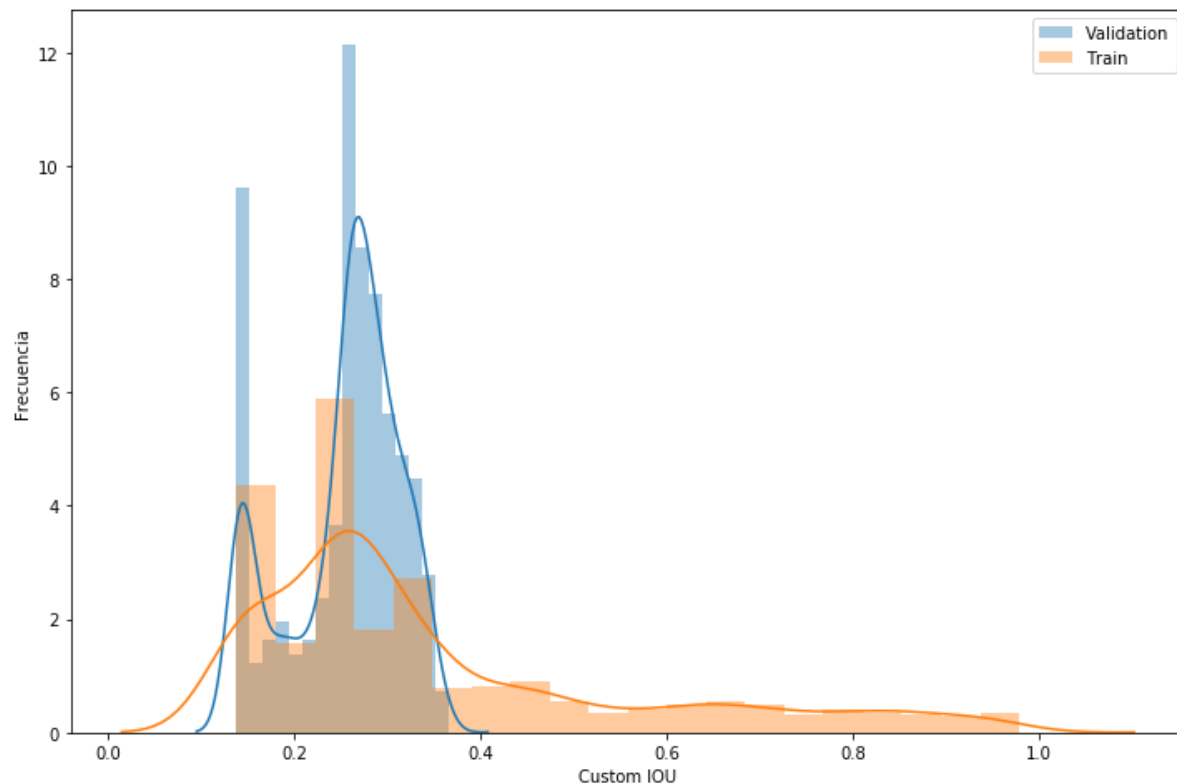
Selección del modelo según Talos

Para seleccionar el “mejor” modelo de utilizó la métrica **val_custom_iou**

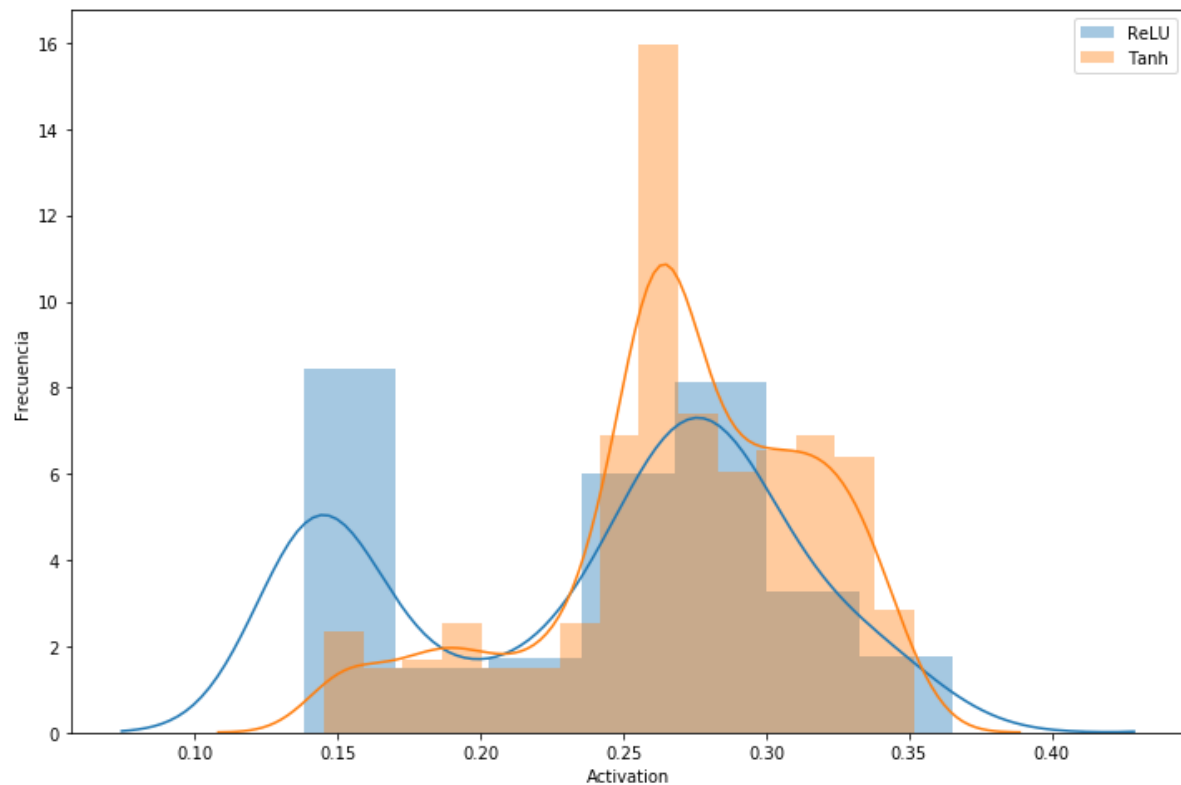
Esta una métrica suma los verdaderos positivos tanto en la unión (or) como en la intersección (and) de los conjuntos y luego los divide entre si.



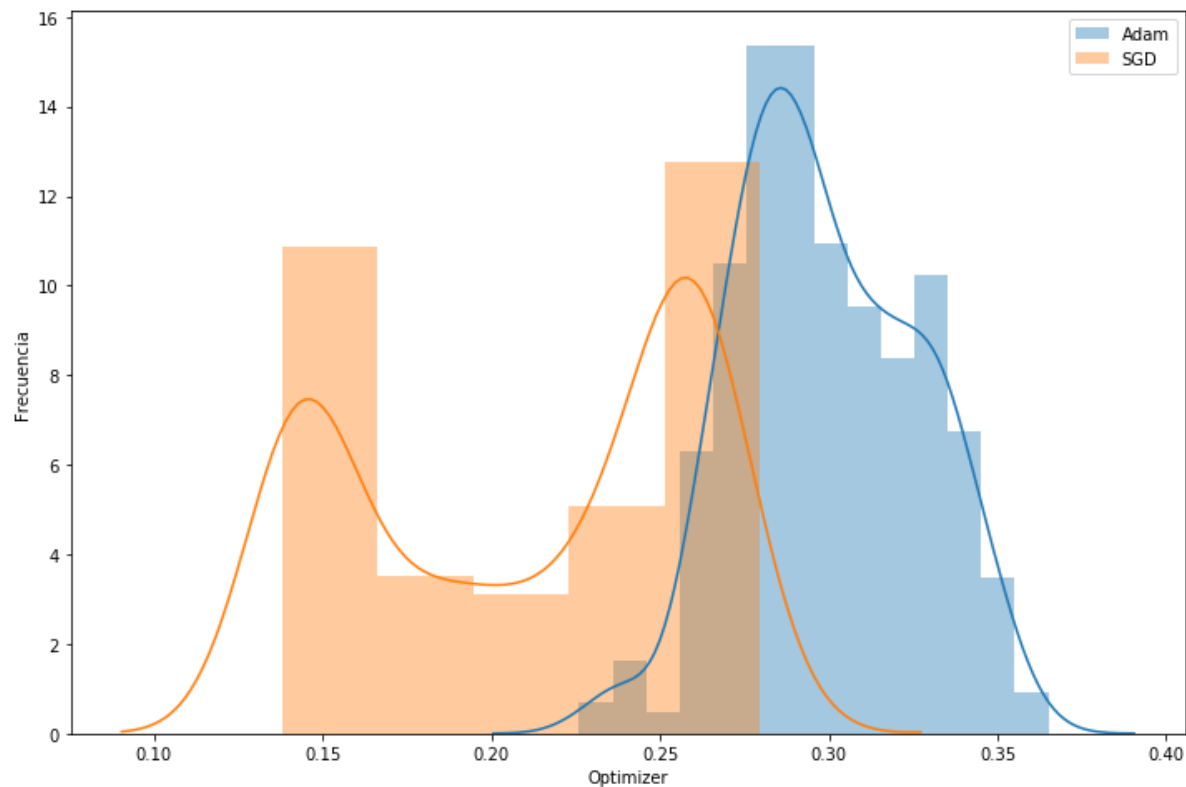
Talos - metricas en base val_custom_iou



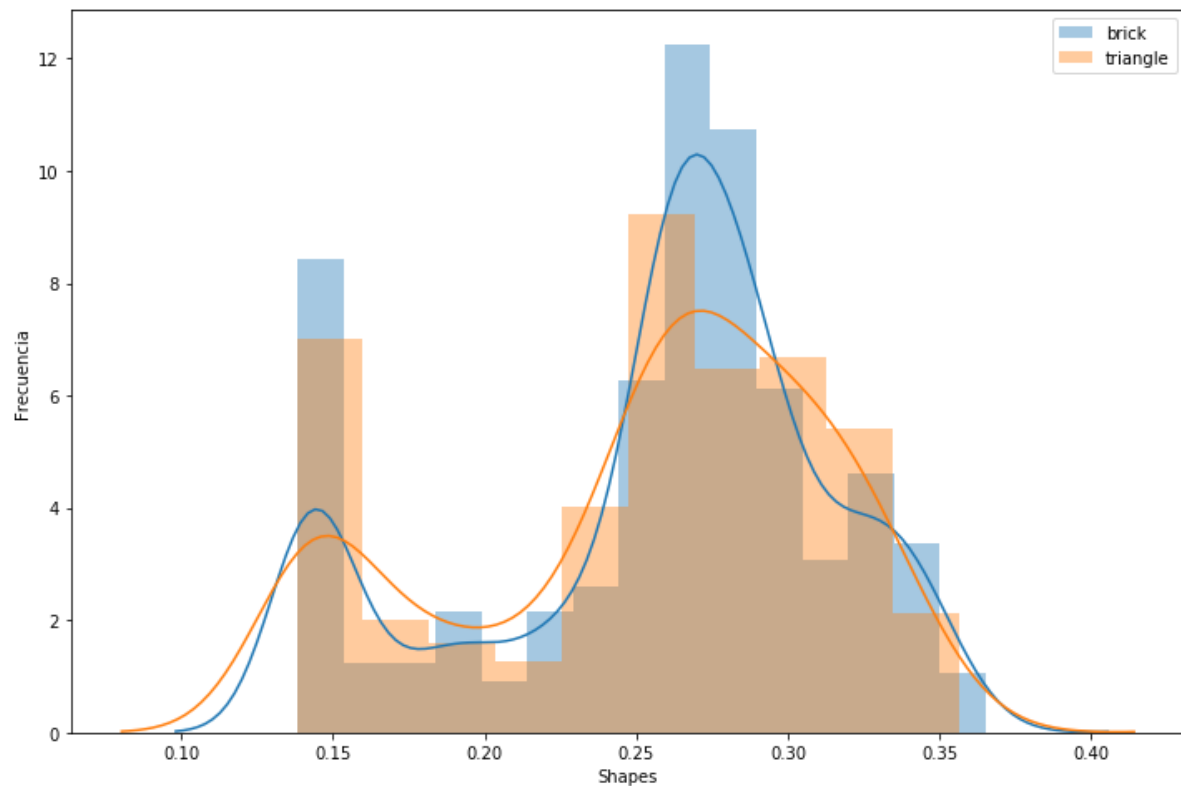
Talos - metricas en base val_custom_iou



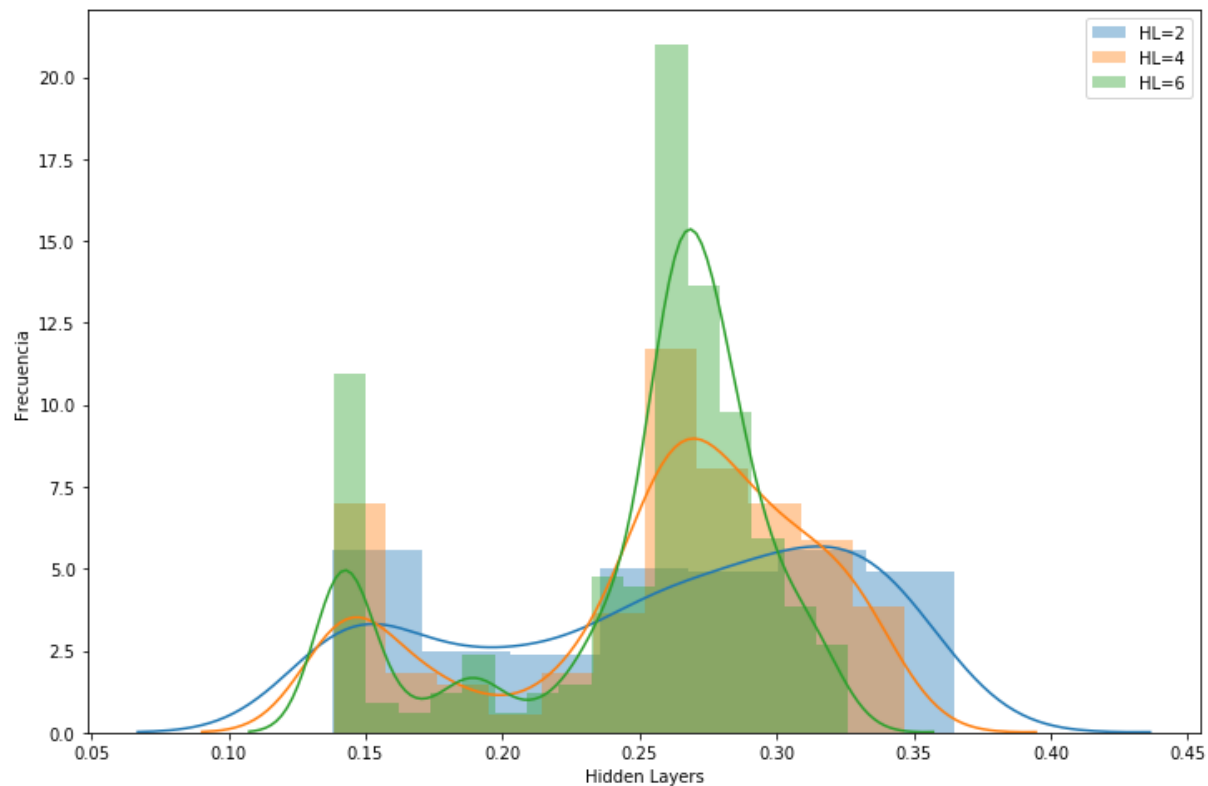
Talos - metricas en base val_custom_iou



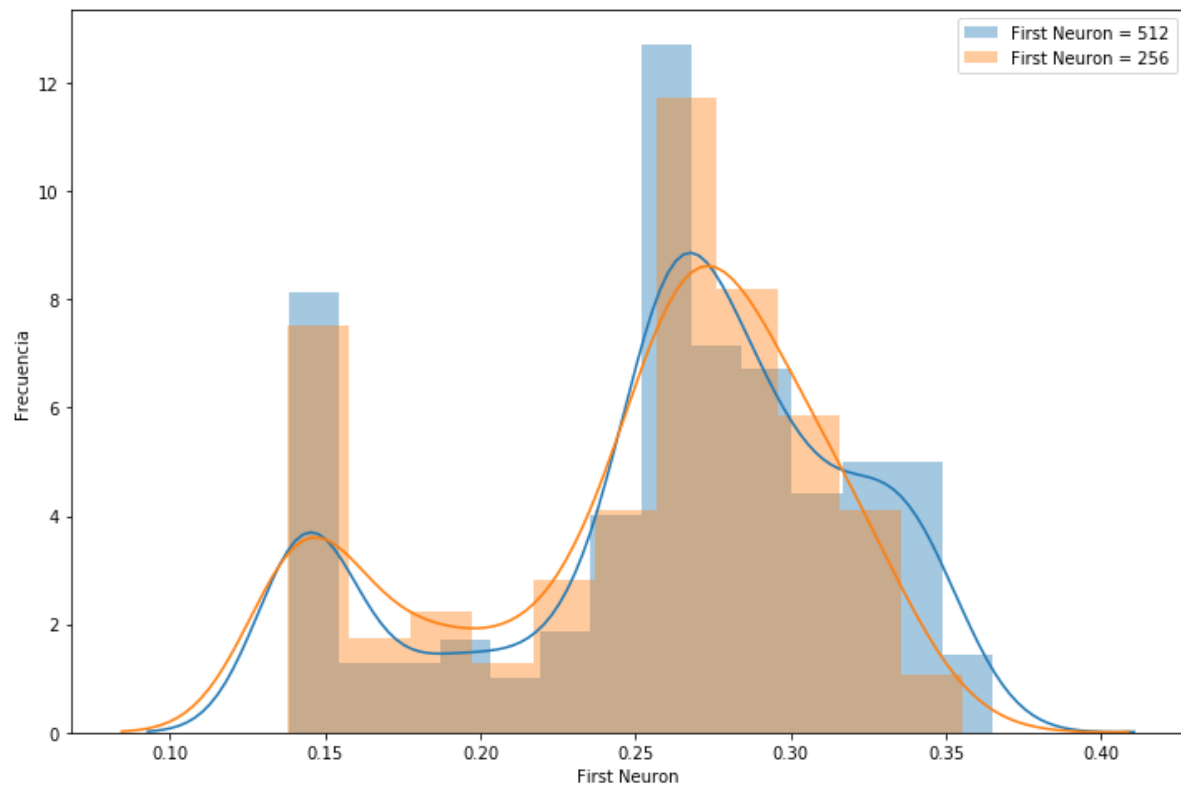
Talos - metricas en base val_custom_iou



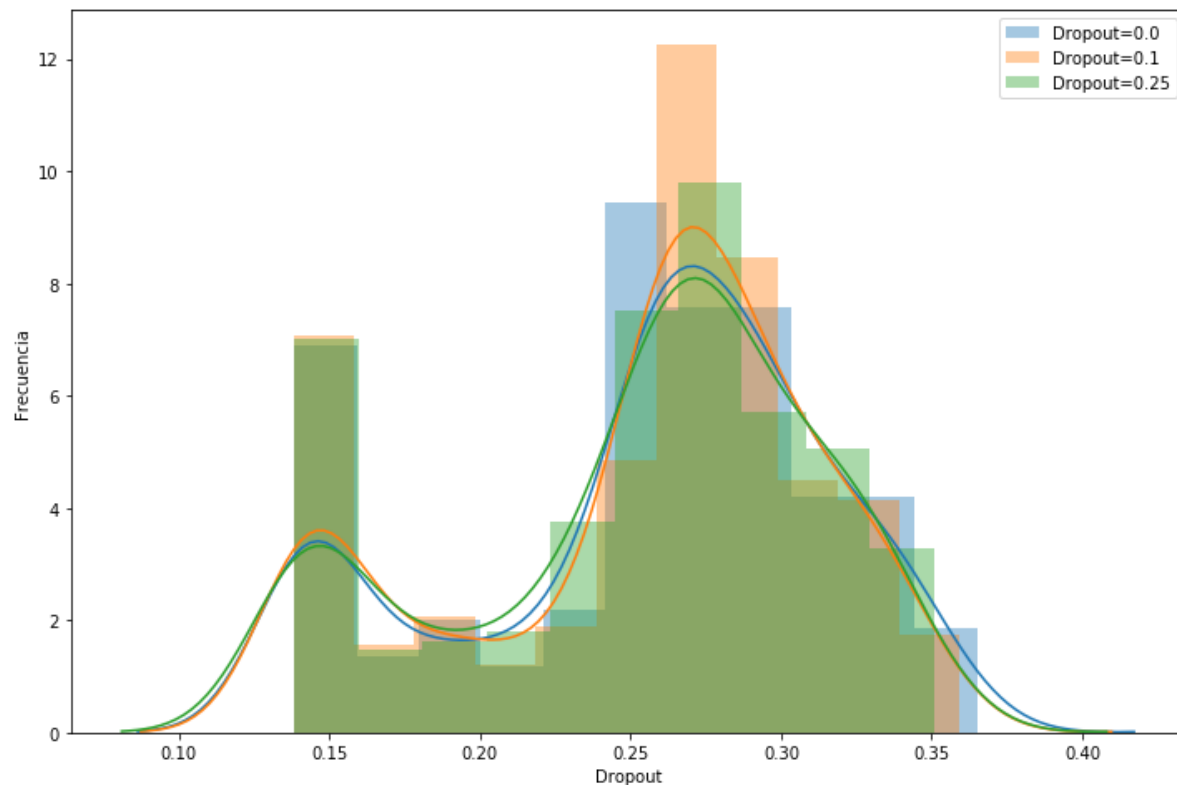
Talos - metricas en base val_custom_iou



Talos - metricas en base val_custom_iou



Talos - metricas en base val_custom_iou



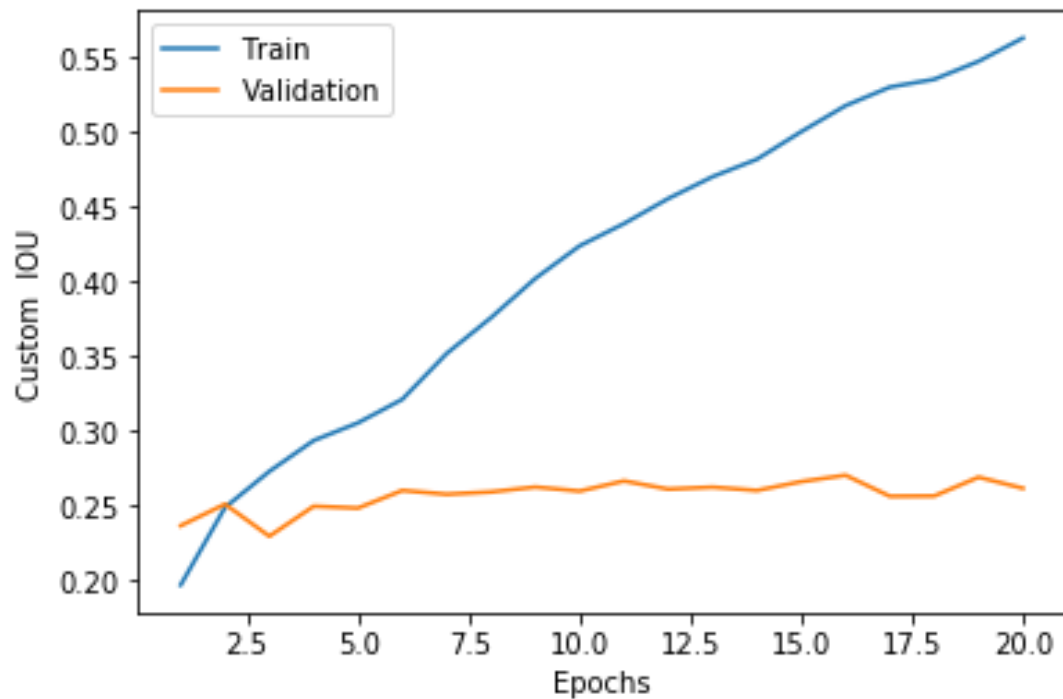
Detalle del modelo seleccionado

Capa 1	512 Nodos	Activación RELU	
Capa 2	512 Nodos	Activación RELU	Dropout 0%
Capa 3	512 Nodos	Activación RELU	Dropout 0%
Capa 4	26 Nodos	Activación sigmoid	

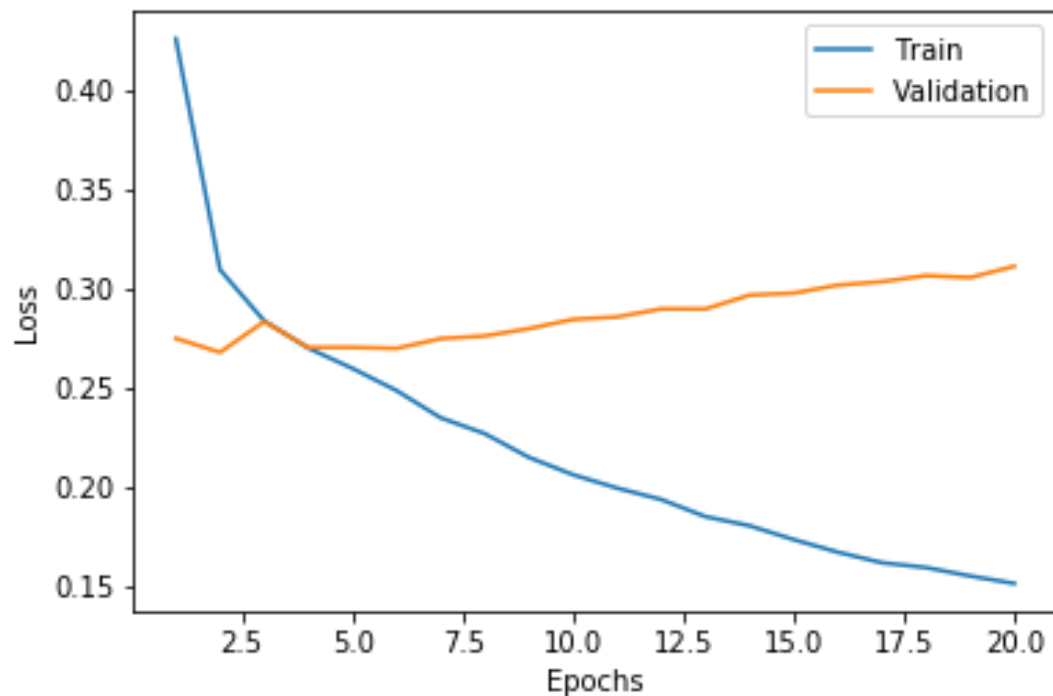
Loss	binary_crossentropy
Optimizador	Adam Learning rate = 0.001
Entrenamiento	batch_size = 64 epochs = 80, round_epochs = 10



Resultado de entrenamiento



Resultado de entrenamiento

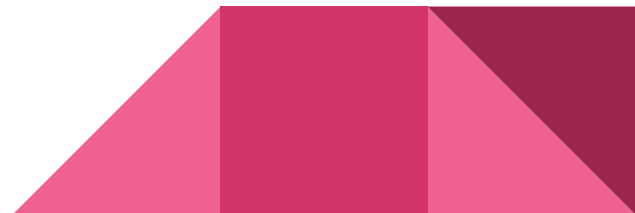


Resultado IOU por Clase

	Tag	Best_Threshold	IOU
0	tag_technology	0.1	17.299578
1	tag_culture	0.1	14.000000
2	tag_science	0.1	12.820513
3	tag_innovation	0.1	9.677419
4	tag_economics	0.1	7.692308
5	tag_politics	0.2	5.263158
6	tag_art	0.3	5.263158
7	tag_design	0.6	5.128205
8	tag_business	0.2	5.050505
9	tag_humanity	0.1	3.921569
10	tag_future	0.1	3.921569

	Tag	Best_Threshold	IOU
10	tag_future	0.1	3.921569
11	tag_entertainment	0.3	3.797468
12	tag_health	0.1	2.857143
13	tag_society	0.1	2.816901
14	tag_children	0.1	2.702703
15	tag_activism	0.1	2.272727
16	tag_community	0.4	2.272727
17	tag_personal_growth	0.1	0.000000
18	tag_identity	0.1	0.000000
19	tag_communication	0.1	0.000000
20	tag_global_issues	0.1	0.000000

	Tag	Best_Threshold	IOU
21	tag_environment	0.1	0.0
22	tag_nature	0.1	0.0
23	tag_social_change	0.1	0.0
24	tag_collaboration	0.1	0.0
25	tag_religion	0.1	0.0



Conclusiones

Después de haber utilizado la herramienta **Talos** para obtener la mejor configuración de hiperparámetros en base a las opciones dadas y el análisis descriptivo realizado, concluimos que **no podemos predecir con una precisión aceptable** los Tags asociados a una charla TED.

Pensamos que la forma para obtener mejores resultados sería utilizar **embeddings** con el texto transcriptor de las charlas, de esta manera sumariamos más información relacionada al tema de la charla y nos permitiría clasificarlas con mayor precisión.
