

# 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 se 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





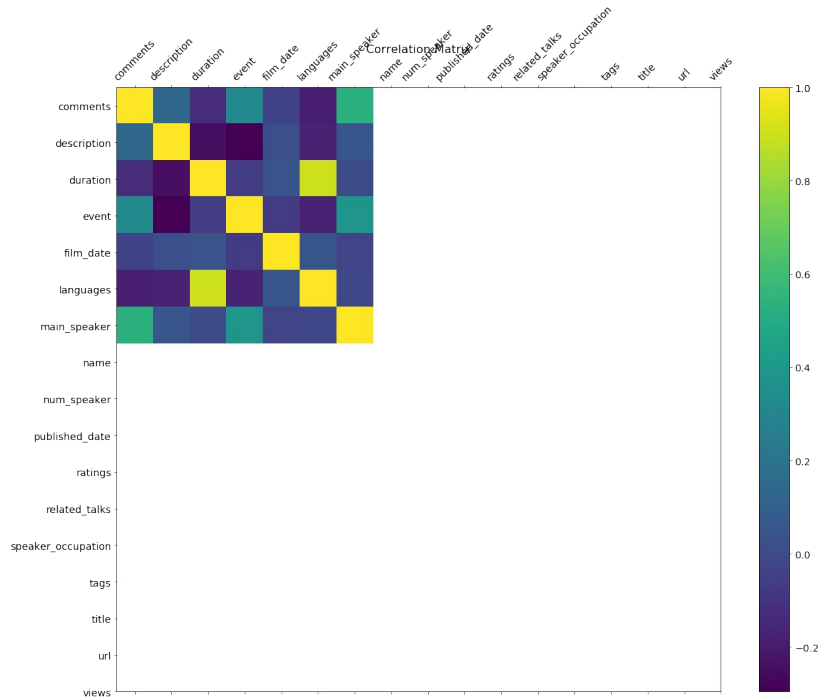
# Campos eliminados (drops)

- film\_date
- published\_date
- description\_title
- event
- main\_speaker
- name
- related\_talks
- ratings
- tags
- url
- speaker\_occupation
- film\_yearpublished\_year



# Análisis de correlaciones preliminares

# Correlaciones



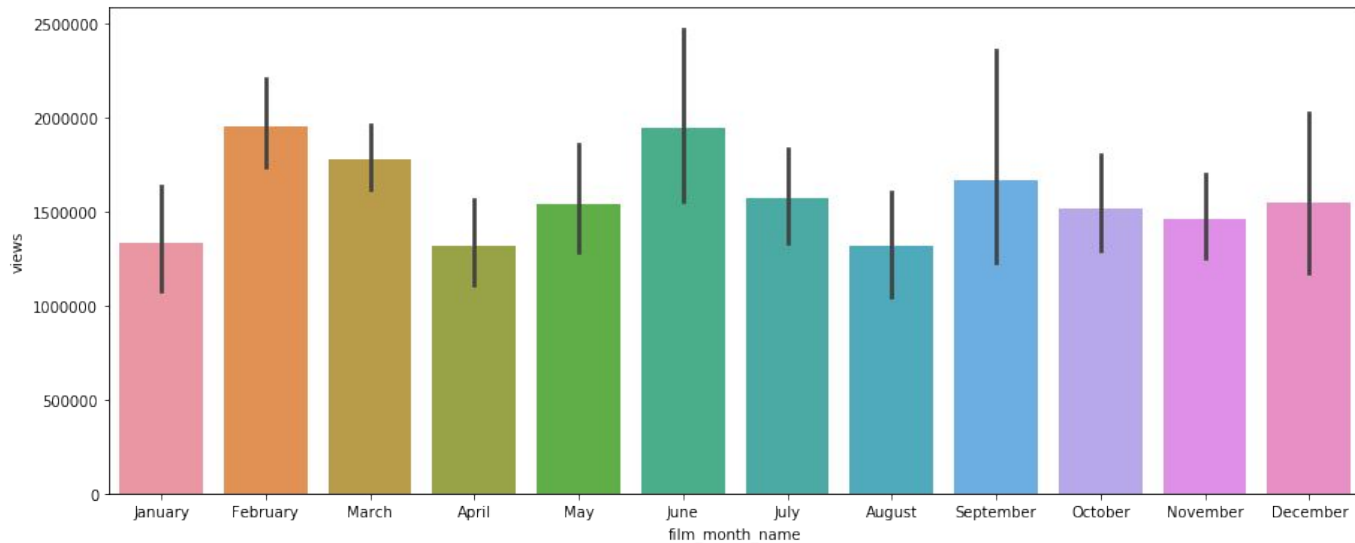
# Correlaciones (cont.)

	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

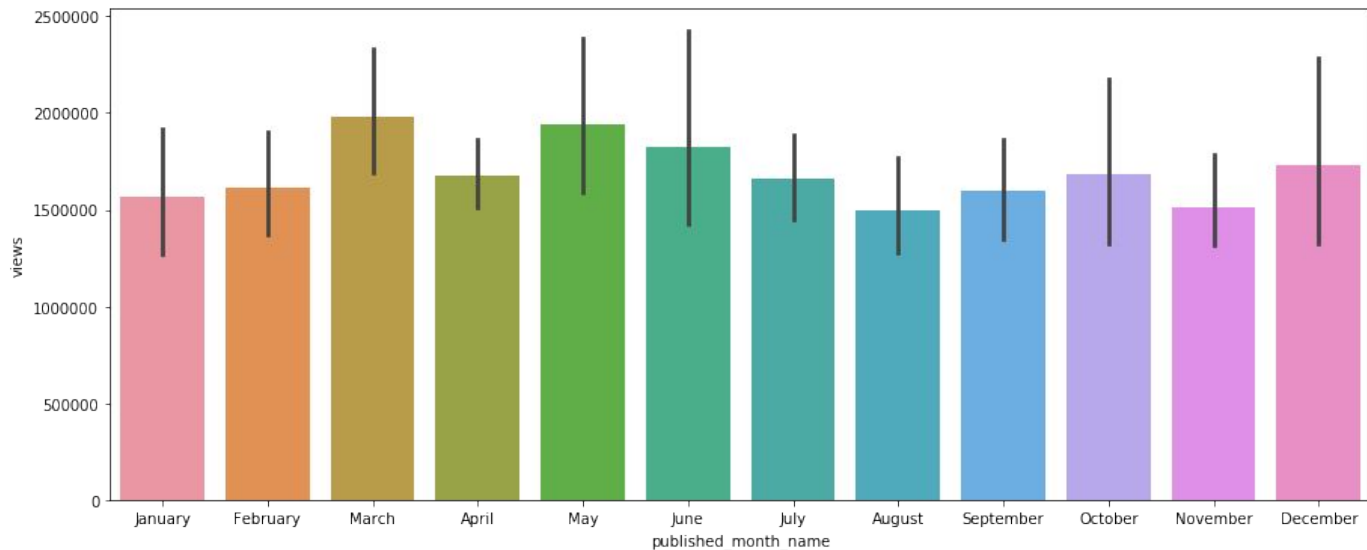


# Visualizaciones preliminares

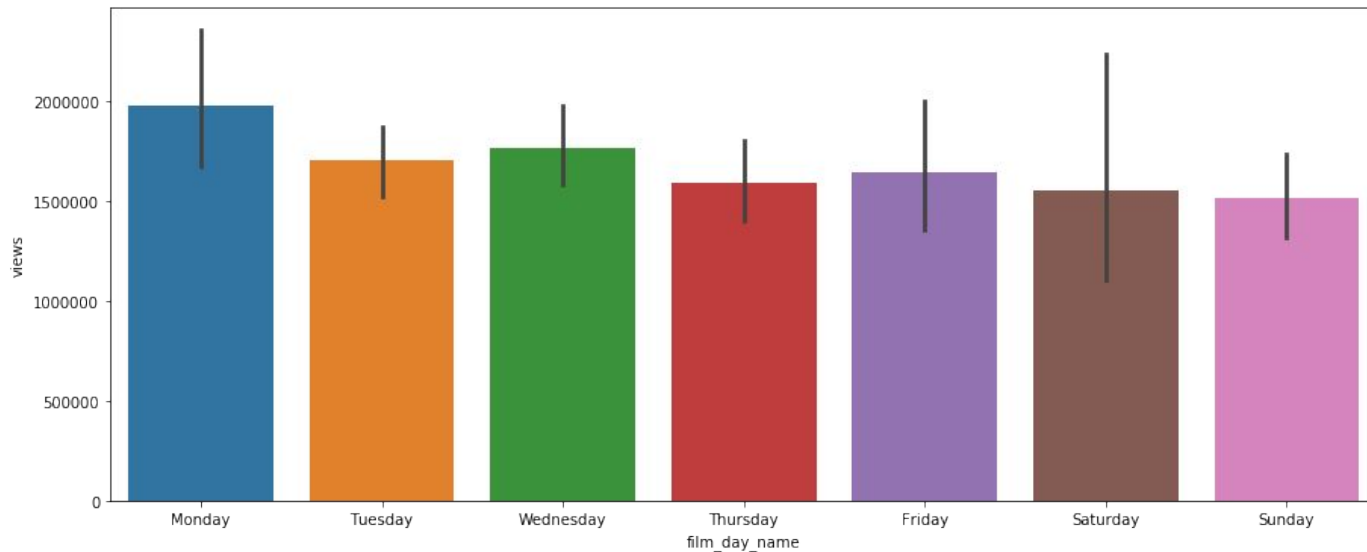
# Visualizaciones por mes



# Visualizaciones por mes de publicación

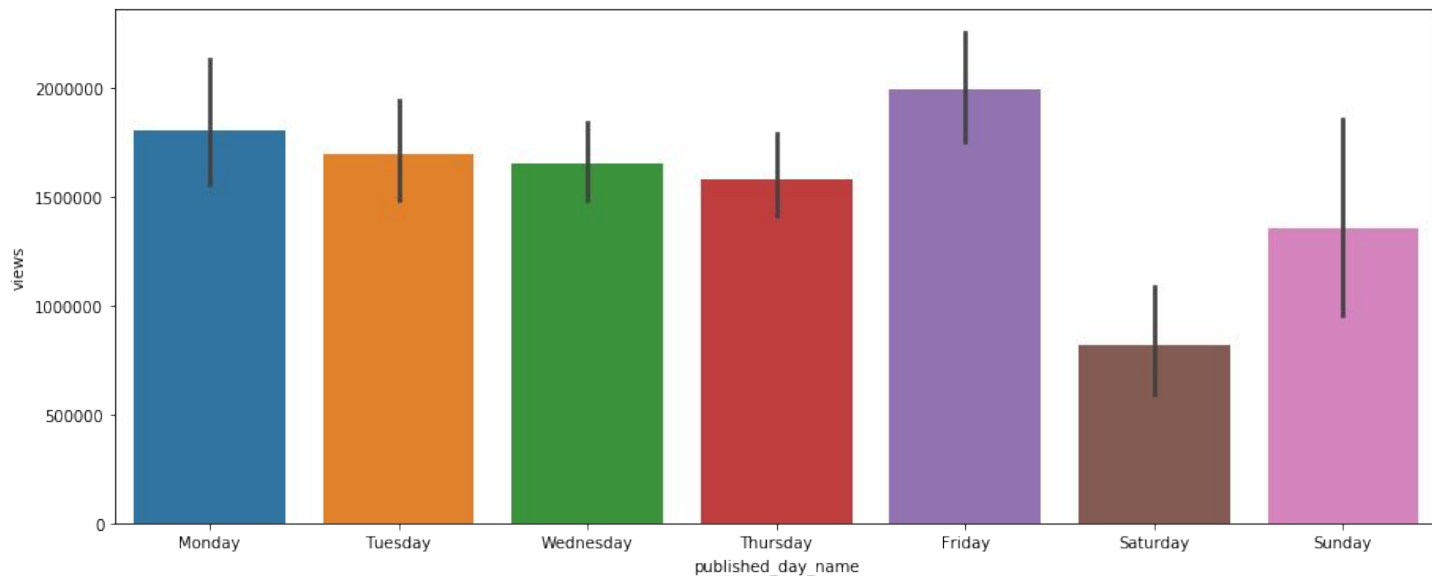


# Visualizaciones por día

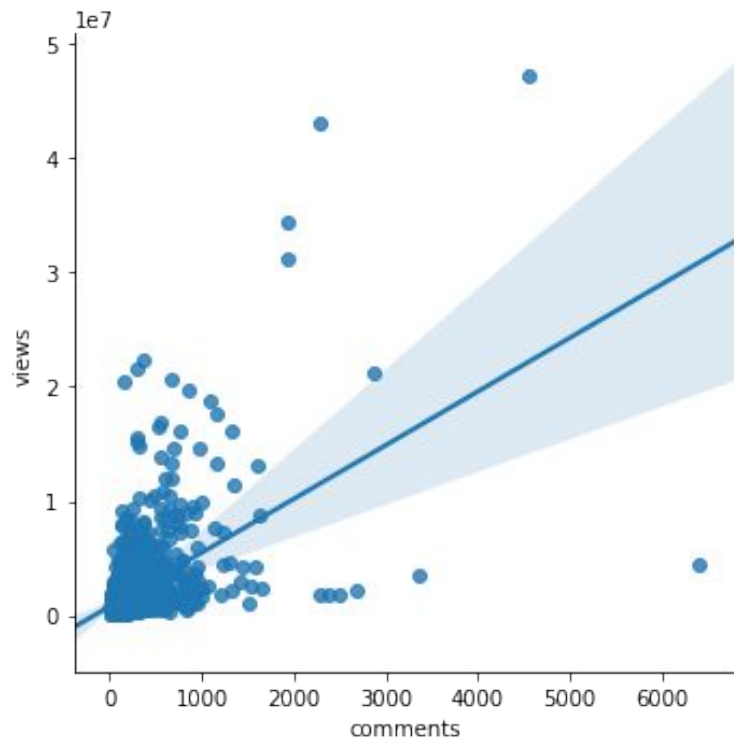




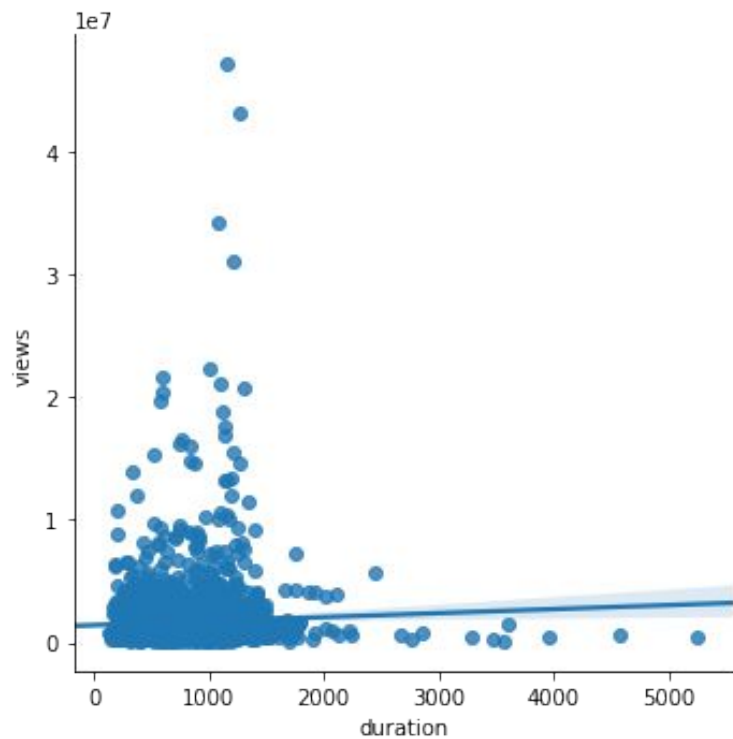
# Visualizaciones por día de publicación



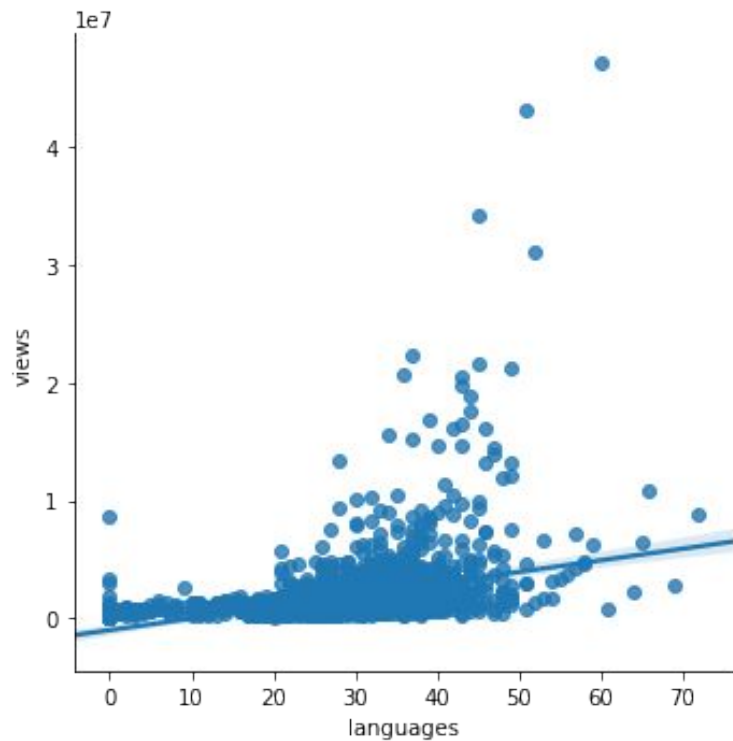
# Visualizaciones por comentario



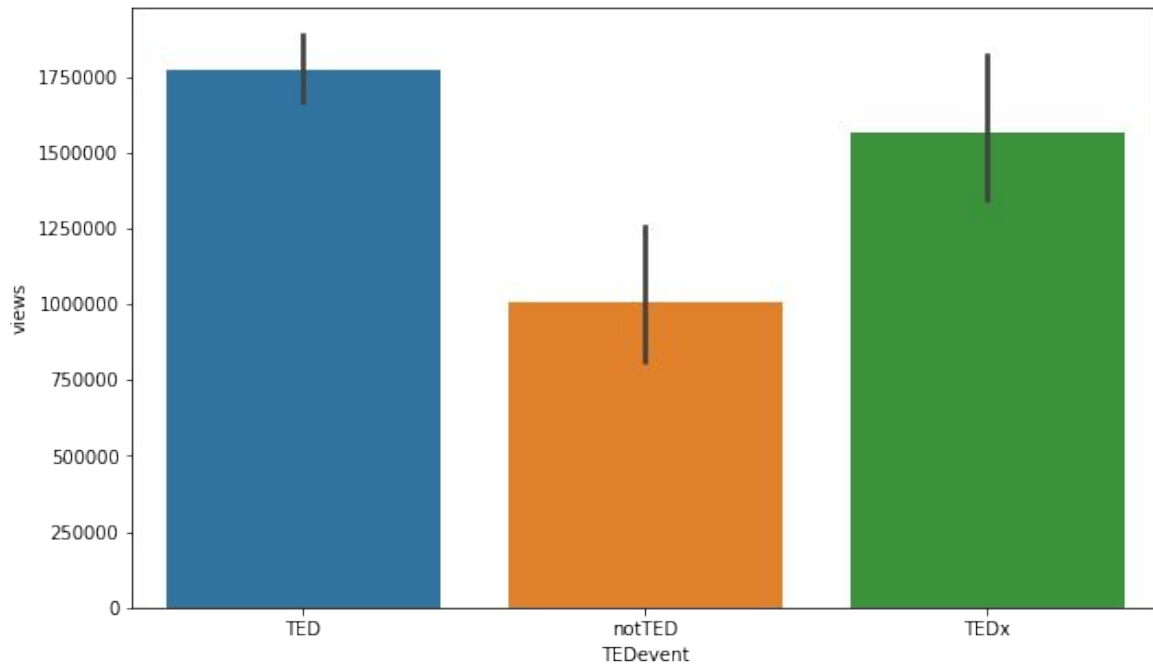
# Visualizaciones por duración



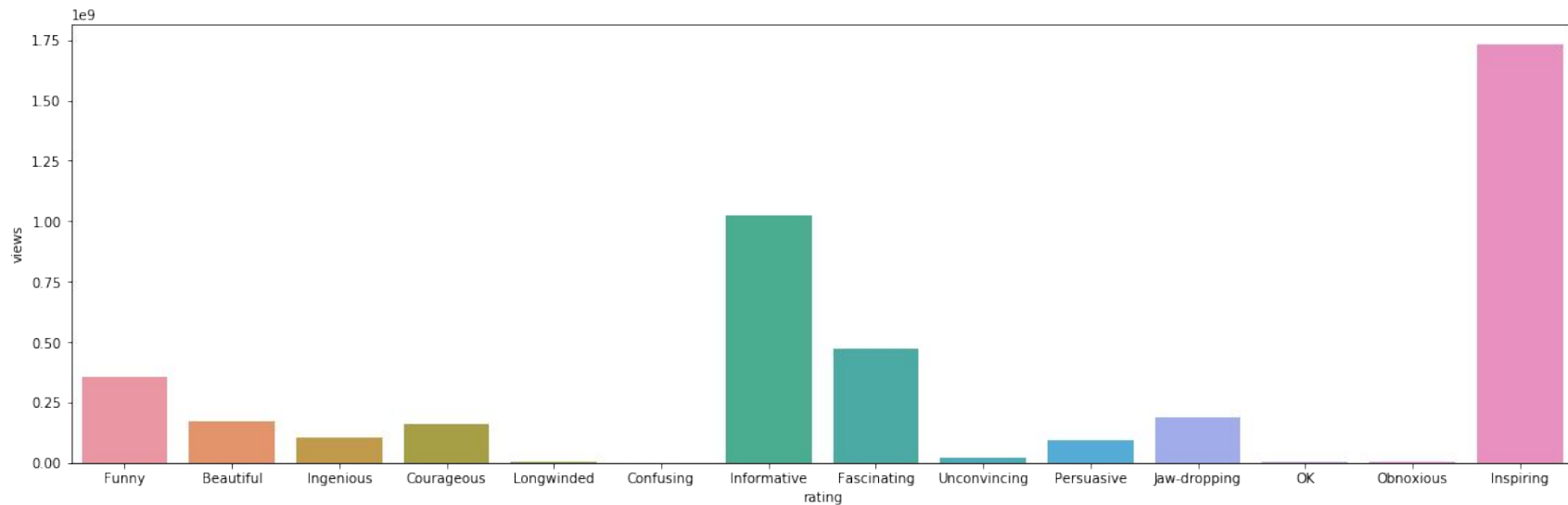
# Visualizaciones por lenguaje



# Visualizaciones por tipo de evento



# Visualizaciones por rating



# Arquitectura del modelo

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

# Detalle del modelo

## Capa 1

- 1024 nodos
- Función de activación RELU
- Regularización DropOut 10%

## Capa 2

- 256 nodos
- Función de activación RELU
- Regularización DropOut 10%

## Capa 3

- 16 nodos
- Función de activación RELU
- Regularización DropOut 10%

## Capa de salida

- 1 nodo
- Sin función de activación

## Función de pérdida

- Mean squared error

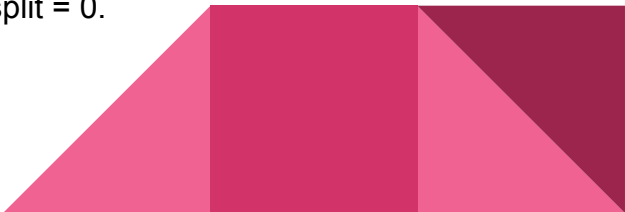
## Optimizador

- Adam
- Learning rate = 0.001

## Métricas

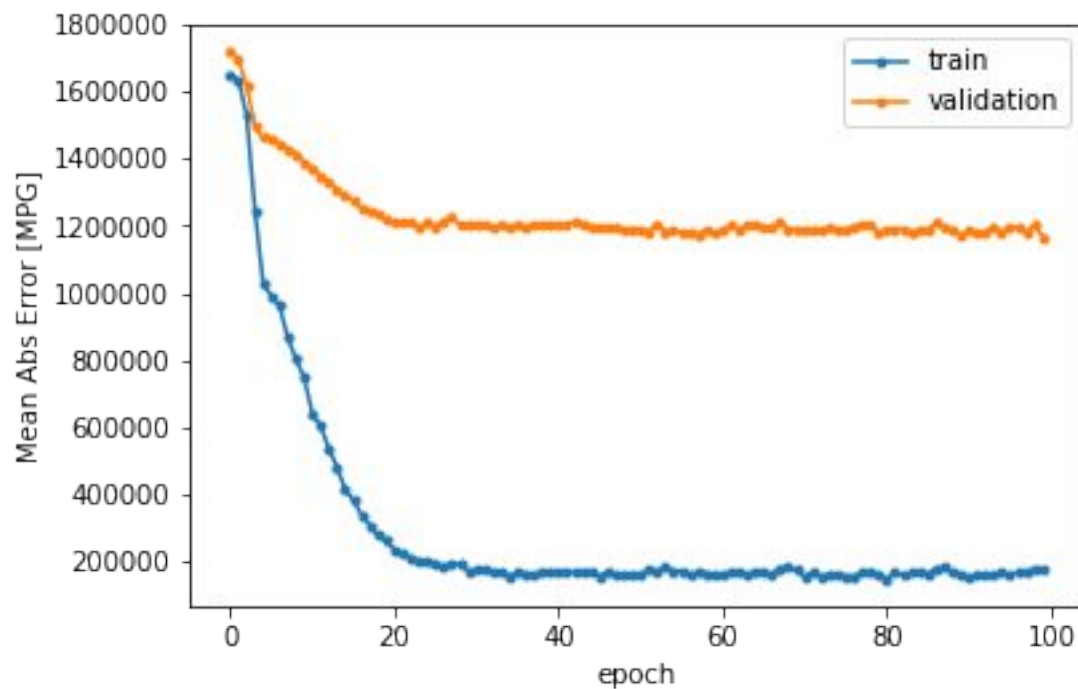
- Mean\_absolute\_error
- mean\_squared\_error

## Entrenamiento

- batch\_size = 32
  - epochs = 500
  - validation\_split = 0.
- 



# 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 cientos de miles) 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 optó por una arquitectura de cuatro capas (densas) ocultas y una capa de salida

# Detalle del modelo

## Capa 1 (1024 nodos)

- Función de activación RELU
- Regularización DropOut 10%

## Capa 2 (512 nodos)

- Función de activación RELU
- Regularización DropOut 50%

## Capa 3 (256 nodos)

- Función de activación RELU
- Regularización DropOut 50%

## Capa 3 (64 nodos)

- Función de activación RELU
- Regularización DropOut 50%

## Capa de salida (nodos = cantidad de clases)

- Función de activación Sigmoid

## Función de pérdida

- `binary_crossentropy`

## Optimizador

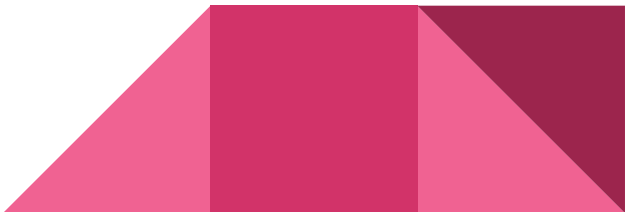
- Adam
- Learning rate = 0.001

## Métricas

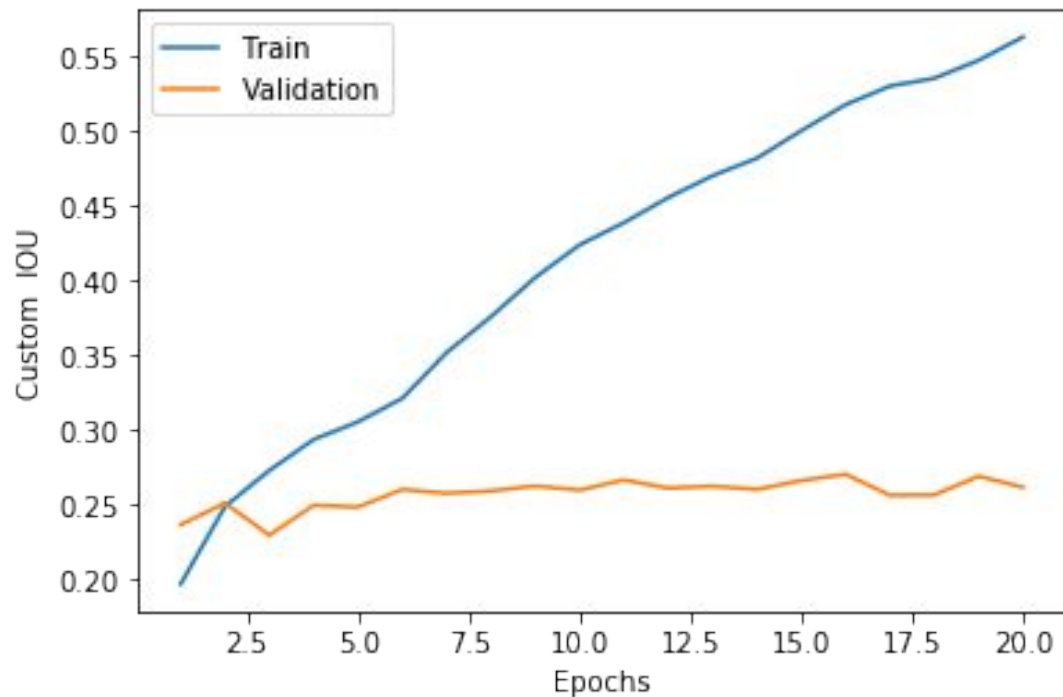
- `custom_iou`

## Entrenamiento

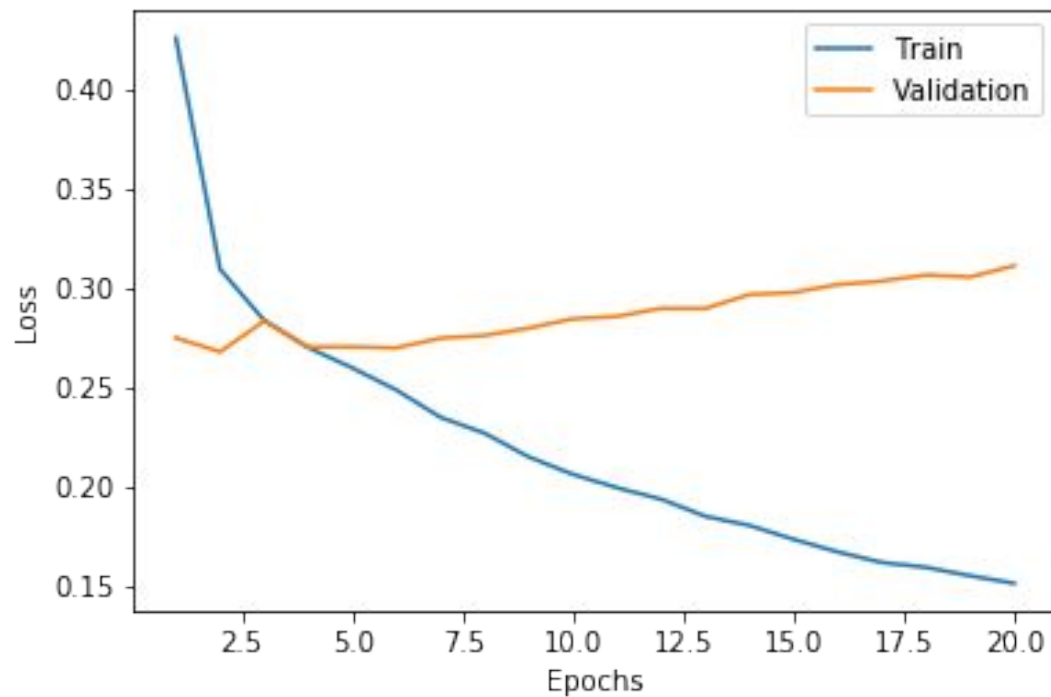
- `batch_size` = 32
- `epochs` = 20
- `validation_split` = 0.



# Resultado de entrenamiento



# Resultado de entrenamiento





# Conclusiones

---