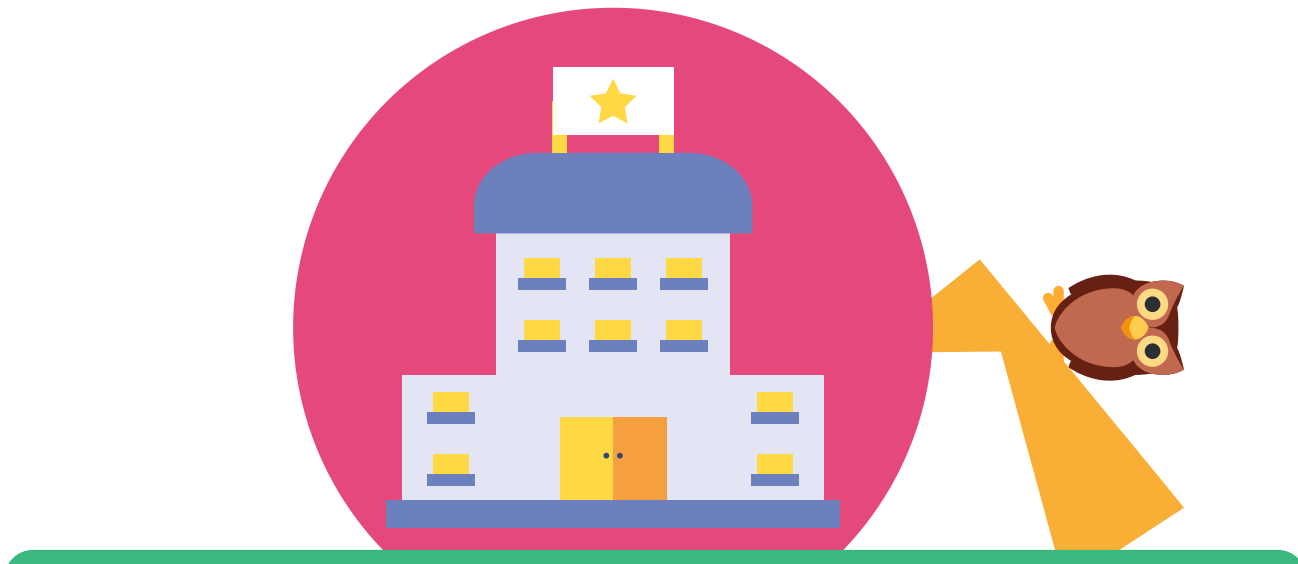


RESEÑAS EN TRIPADVISOR

Poster

Investigación de Datos: Reseñas de Tripadvisor

CONTEXTO

En la actualidad la industria hotelera está atravesando un capítulo particular en nuestra era moderna: la pandemia. A pesar de ello es una industria en crecimiento y en momento de éxito de negocios, dado el comportamiento presentado en 2019.

TRIPADVISOR Y LA TECNOLOGÍA EN EL TURISMO

Según un estudio, mencionado en el periódico The Guardian, en la plataforma digital un incremento de una estrella en la calificación de un hotel representa un incremento de ingresos entre un 5% a un 9%.

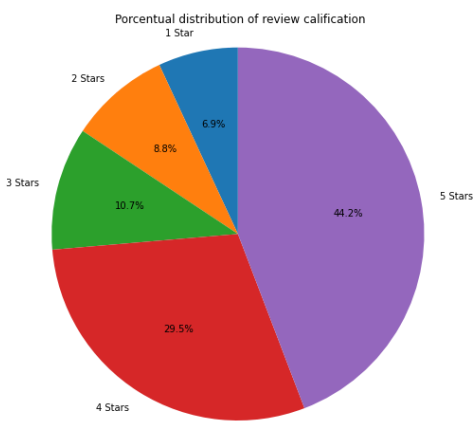
INVESTIGACIÓN

A partir de ello se decidió investigar los datos acerca de las reseñas que se encuentran en TripAdvisor para ver que palabras eran comunes y para plantear algún modelo que describiera la relación entre las palabras usadas y la calificación que se le otorgaba al hotel. El se de datos es: Trip Advisor Hotel Reviews. 20k Hotel reviews extracted from Tripadvisor.

EXPLORACIÓN

En la exploración de datos se obtuvo información interesante.

DISTRIBUCIÓN DE LAS CALIFICACIONES



PALABRAS MÁS REPETIDAS EN LAS RESEÑAS DE 5 ESTRELLAS

PROCESAMIENTO

Para el procesamiento de los datos se recurrió a al método de vectorización. Es decir trasladar un párrafo a un vector de números que representan cada uno una palabra. Estas palabras se catalogan en un diccionario como se ve en la figura.

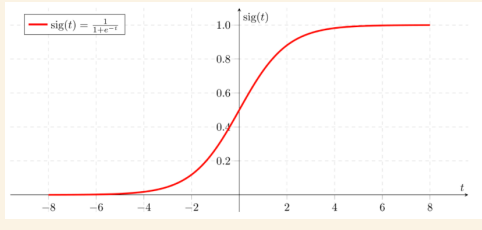
Diccionario

Aa Palabra	# Llave
'nice'	9324
'hotel'	6767
'expensive'	5030
...	

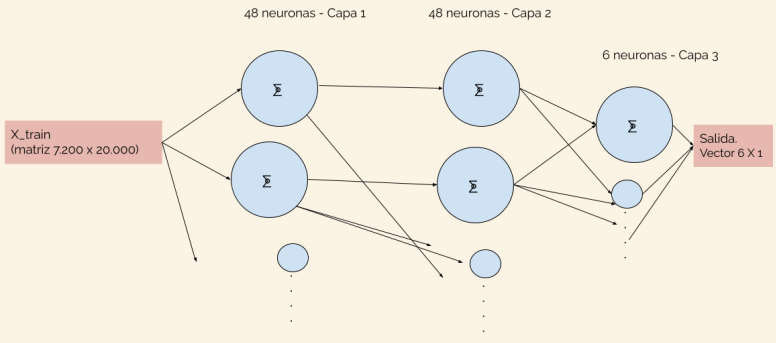
MODELO

Se plantean dos modelos: Regresión Logística y una Red Neuronal.

REGRESIÓN LOGÍSTICA



RED NEURONAL



RESULTADOS

En la exploración de datos se obtuvo información interesante.

RESULTADOS PARTE I

Modelo de ML: Regresión logística
Accuracy

```
[23] lr.score(X_test, y_test)

     0.5761111111111111
```

Resultados de *score*

```
[26] print(classification_report(y_pred,y_test))

                   precision    recall  f1-score   support

                1       0.60      0.69      0.64       122
                2       0.23      0.38      0.29       102
                3       0.12      0.38      0.18        64
                4       0.53      0.48      0.51       589
                5       0.80      0.66      0.72       923

         accuracy                           0.58      1800
        macro avg       0.46      0.52      0.47      1800
     weighted avg       0.64      0.58      0.60      1800
```

RESULTADOS PARTE II

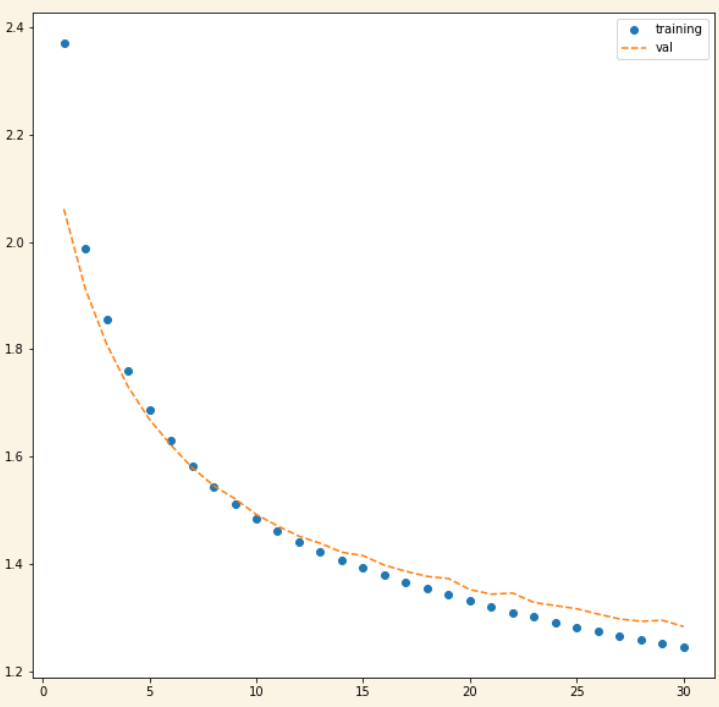
Modelo de Red Neuronal. En el caso de la red neuronal fue necesario un paso adicional en los datos *target*_u objetivo. La salida de la red neuronal es una funciónon sigmoide la cual te da una probabilidad entre todas las opciones. Tenemos una calificación de 0 a 5, es decir 6 opciones. El procesamiento se ve así, arriba como es y_train originalmente, abajo categorzado en un vector.

```
[30] print(y_train[4302])
     print('----------')
     print(y_train_categorical[0])

     5
     ----------
     [0. 0. 0. 0. 0. 1.]
```

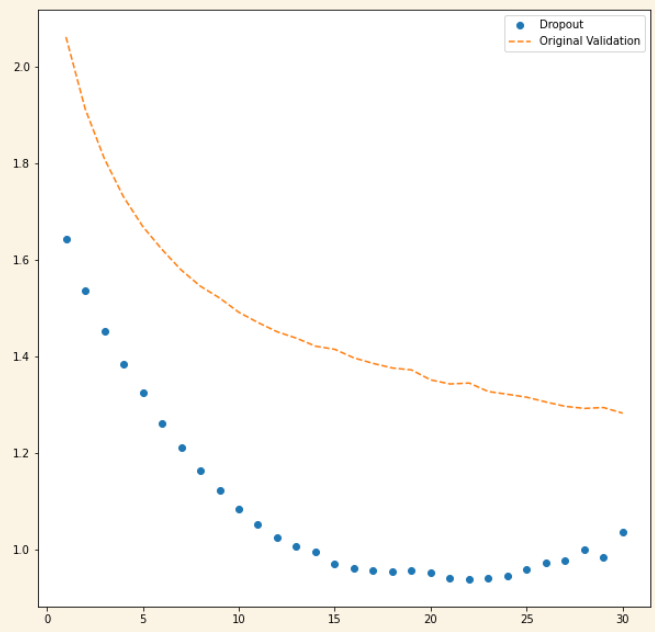
RED NEURONAL I

En la primera red neuronal hecha tenemos el siguiente resultado, un valor de *accuracy* del 52%, además el valor training tiende al overfitting:



RED NEURONAL II

Por ende se creó otra red neuronal, esta vez con un proceso adicional llamado *Dropout*, consiste *en apagar algunas neuronas de manera aleatoria* para evitar el proceso que causa overfitting. Se puede ibservar la diferencia entre el valor de accuracy de la red origina (red I) y la red con dropout. Además se puede ver que el mejor valor conseguido está en la época 21 o 22.



Resultados de *score*

```
[46] print(classification_report(y_pred,y_t)) # Y predecido

                   precision    recall  f1-score   support

                1       0.39      0.71      0.50        78
                2       0.41      0.38      0.40       180
                3       0.22      0.36      0.27       118
                4       0.47      0.49      0.48       512
                5       0.80      0.66      0.72       912

         accuracy                           0.57      1800
        macro avg       0.46      0.52      0.48      1800
     weighted avg       0.61      0.57      0.58      1800
```

Poster
Investigación de Datos: Reseñas de Tripadvisor

Grupo
Juan Sebastián Vargas Castañeda
David Enrique Eslava
Juan Sebastián Montoya Combita

Introducción a los Sistemas Inteligentes
2021-I
Profesor
Fabio A. González
Carrera de Ingeniería de Sistemas y Computación
Universidad Nacional de Colombia