Universidad Jorge Tadeo Lozano

Facultad de Ingenieria Sistemas

Proyecto Inteligencia Artificial

trip advisor-hotel-reviews

Autores: Nicolas Rojas Molina
Andres Felipe Barbosa Hernandez
Brandon Rambauth

Marzo 2021

Introduccion

¿Cuales son los mejores hoteles para los clientes segun su satisfacción o experiencia basada en como se sienten?

En el presente laboratorio se aplicarán las bases de python aprendidas, haciendo uso de ecuaciones estadisticas de los datos de calificaciones de hoteles usando librerias tales como: Numpy, pandas y Matplotlib con las cuales se analizaran los datos de calificacion de 20490 hoteles, teniendo en encuenta que este analisis es basado en las opiniones de los usuarios, segun lo que sienten a la hora de usar el servicio o despues de usarlo.

Metodo y visualizacion

Tras haber realizado las pertinentes visualizaciones de los datos hemos notado que las opiniones de los huespedes frente a la atencion de los hoteles tiene calificacion en un rango de 1 a 5 estrellas, segun la experiencia de uso del servicio.

para lo cual mediante el uso de herramientas y metodos de estadistica, se tomo la desicion de realizar un analisis de los datos, los cuales constan de hallar la media o media aritmetica,la mediana, la moda, y la varianza, con las respectivas ecuaciones.

Analisis de sentimientos

Para el analisis de sentimientos se uso las librerias Pandas, textblob y seaborn. Las librerias como textblob se usan para el procesamiento de texto el cual permite el procesamiento de lenguaje natural tales como el análisis morfológico, extracción de entidades, análisis de opinión, traducción automática, etc. Esta libreria esta contruida por otras librerias muy conocidad para el analisis de texto NLTK y pattern.

Seaborn es solo un paquete de matplotlib la cual ayuda a que las graficas sean mas esteticas

Machine learning

Se concentra en la construcción y estudio de sistemas que puedan aprender de los datos, uno de los problemas mas grandes del machine learning es encontrar patrones, relaciones y regularidades en los datos, con el fin de construir modelos predictivos.

1. Analisis de tweets

En el analisis de sentimientos sirve para identificar la polaridad y subjetividad los cuales nos ayuda a identificar y extraer informacion, lo que se quiere hacer con este analisis es determinar la actitud del usuario con respecto algun tema. la polaridad se analiza para saber si el contexto escrito por el usuario sea positivo o negativo

	Review	Rating
0	nice hotel expensive parking got good deal sta...	4
1	ok nothing special charge diamond member hilto...	2
2	nice rooms not 4* experience hotel monaco seat...	3
3	unique, great stay, wonderful time hotel monac...	5
4	great stay great stay, went seahawk game aweso...	5

(a) tabla de los primeros 5 tweets

```
print(tripadvisor_df['polaridad'].head())

0    0.208744
1    0.214923
2    0.294420
3    0.504825
4    0.384615
Name: polaridad, dtype: float64
```

(a) valor de la polaridad en los tweets

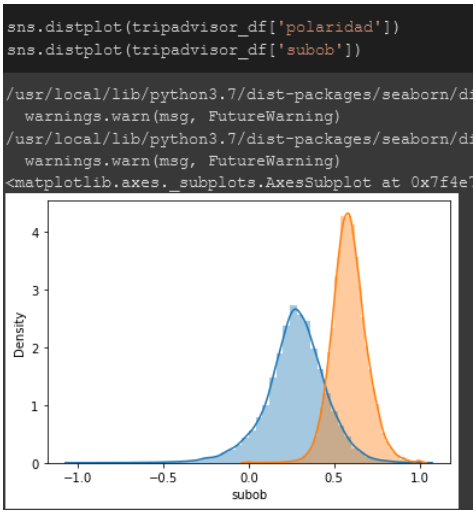
```
print(tripadvisor_df['subob'].head())

0    0.687000
1    0.495009
2    0.605208
3    0.691228
4    0.629396
Name: subob, dtype: float64
```

(a) valor de la subobjetividad de los tweets

2. Ecuaciones

La **Media** o **Media aritmetica** es el valor promedio de un conjunto el cual es calculado con la suma total de los valores del conjunto dividido entre el numero total de valores.

$$Media(X) = \bar{x} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{x_1 + x_2 + x_3 ... + x_n}{N} \tag{1}$$



(a) grafica de polaridad y subobjetyividad de los datos analizados

la **Mediana**, para calcularla lo primero que debemos saber es si N es par o impar.

- si N es impar, la mediana es el valor que esta al medio.

$$Mediana(X) = X_{\frac{N+1}{2}} \tag{2}$$

- si N es par, la mediana es la media de los valores del centro N/2 y N/2+1:

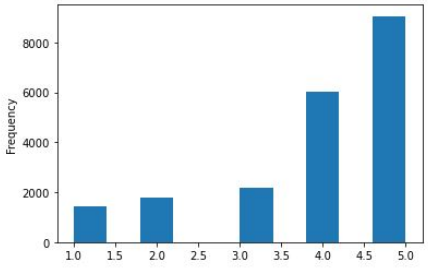
$$Mediana(X) = Media(X_{\frac{N}{2}}, X_{\frac{N}{2}+1}) = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2} \tag{3}$$

la **Moda** es el dato con mayor frecuencia en una distribucion de datos y se calcula de la siguiente forma:

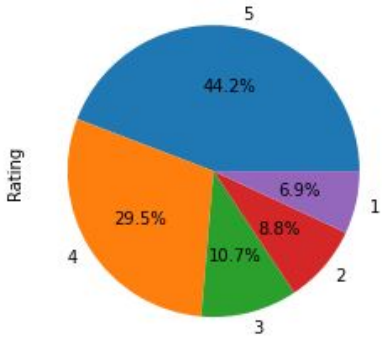
$$M_o = Li + (\frac{\triangle 1}{\triangle 1 + \triangle 2}) * C \tag{4}$$

la **Varianza** es la unidad de medida correspondiente a los datos pero elevada al cuadrado, y siempre es mayor o igual a cero.

$$Var(X) = \frac{\Sigma_1^n (x_1 - \bar{X})^2}{n} \tag{5}$$



(a) Hist: hoteles segun la calificacion de usuario



(a) Porcentaje de calificacion segun experiencia del usuario

3. Analisis

- al sumar la calificacion de estrellas dadas dividido las cantidades de la misma(20490), podemos sacar la media lo que equivale a el promedio de calificacion que en este caso seria de 3.95.
- al dividir la calificacion de estrellas podemos ver que la mediana es decir el 50 % de los datos es 4.0
- encontramos que el numero de la calificacion del hotel que mas se repite es 5 en la calificacion de los huespedes.
- podemos evidenciar en el histograma la separacicon de los datos al rededor de la media

4. Analisis de machine learning

-

```
from sklearn import preprocessing
from sklearn import datasets
from sklearn.linear_model import LogisticRegression
import pylab as pl
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from matplotlib.colors import Normalize
from time import time
```

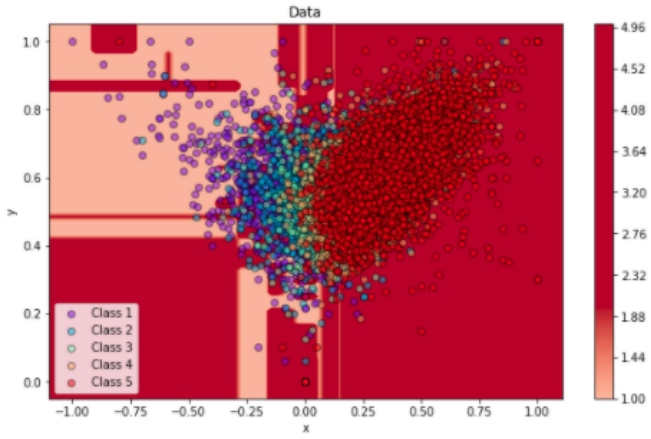
(a) importacion de datos

- extraccion de label en nuestro caso el Rating

```
[ ]  #extraer un label en nuestro caso el Rating

     Rating=tripadvisor_df['Rating']
     Rating
     tripadvisor_df.drop('Rating',axis=1,inplace=True)
```

(a) importacion de datos

```
In [404]: pl.figure(figsize = (10, 6))
          plot_decision_region(X, gen_pred_fun(classifier))
          plot_data(X_train, y_train)
```




```
In [405]: # Error de clasificación
          print("score con datos de entrenamiento")
          print(classifier.score(X_train,y_train))
          print("error con datos de entrenamiento")
          print(1-classifier.score(X_train,y_train))

          score con datos de entrenamiento
          0.7918845429826397
          error con datos de entrenamiento
          0.20811545701736034
```

(a) datos recopilados

5. Conclusiones

Como se puede observar en la investigacion realizada, los datos obtenidos al final nos da como conclusion:

1. Al analizar los datos nos encontramos con el promedio de las calificaciones de los hoteles visitados por los clientes, los cuales podemos evidenciar que la mitad de los datos son 4.0 estrellas y se puede demostrar que la calificacion de los huespedes que mas se repite para los hoteles es de 5.0 estrellas

6. Links video

- https://www.youtube.com/watch?v=KBqXbW3O1eQ
- https://github.com/IAutadeo/Primera_.Entrega.git

7. Glosario

- López, J. F. (2021, 13 febrero). Media. Economipedia. https://economipedia.com/definiciones/media.html
- Serra, B. R. (2020, 22 noviembre). Mediana. Universo Formulas. https://www.universoformulas.com/estadistica/descriptiva/mediana/
- Descriptiva, E. (2021, 10 marzo). moda, para datos agrupados y no agrupados. moda http://descriptiva2010.blogspot.com/2010/03/moda-para-datos-agrupados-y-no.html
- López, J. F. (2021, 27 enero). Varianza. Economipedia. https://economipedia.com/definiciones/varianza.html