

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 estadísticas de los datos de calificaciones de hoteles usando librerías tales como: Numpy, pandas y Matplotlib con las cuales se analizarán los datos de calificación de 20490 hoteles, teniendo en cuenta que este análisis es basado en las opiniones de los usuarios, según lo que sienten a la hora de usar el servicio o después de usarlo.

Metodo y visualizacion

Tras haber realizado las pertinentes visualizaciones de los datos hemos notado que las opiniones de los huéspedes frente a la atención de los hoteles tiene calificación en un rango de 1 a 5 estrellas, según la experiencia de uso del servicio.

para lo cual mediante el uso de herramientas y métodos de estadística, se tomó la decisión de realizar un análisis de los datos, los cuales constan de hallar la media o media aritmética, la mediana, la moda, y la varianza, con las respectivas ecuaciones.

Analisis de sentimientos

Para el análisis de sentimientos se usó las librerías Pandas, textblob y seaborn. Las librerías 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 librería está contruida por otras librerías muy conocidas para el análisis de texto NLTK y pattern.

Seaborn es solo un paquete de matplotlib la cual ayuda a que las gráficas sean más estéticas

1. Analisis de tweets

En el análisis de sentimientos sirve para identificar la polaridad y subjetividad los cuales nos ayudan a identificar y extraer información, lo que se quiere hacer con este análisis es determinar la actitud del usuario con respecto a algún 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 hillo...	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

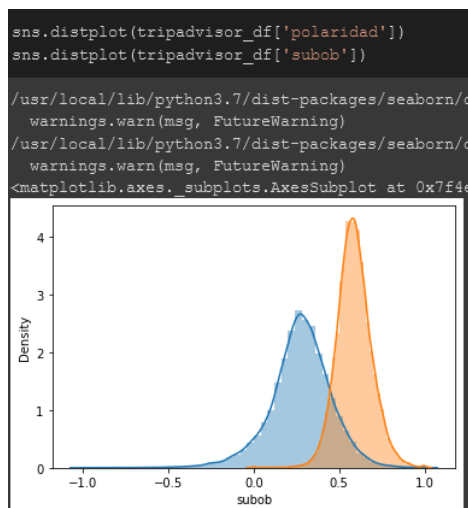
(a) valor de la polaridad en los tweets

```
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



(a) grafica de polaridad y subobjetyivi-
dad de los datos analizados

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.

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

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.

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

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

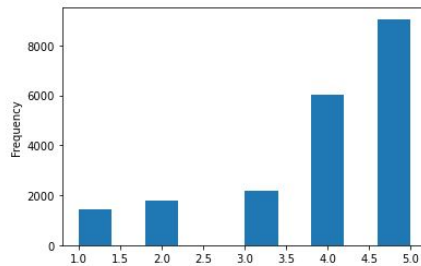
$$Mediana(X) = Media(X_{\frac{N}{2}}, X_{\frac{N}{2}+1}) = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2} \quad (3)$$

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

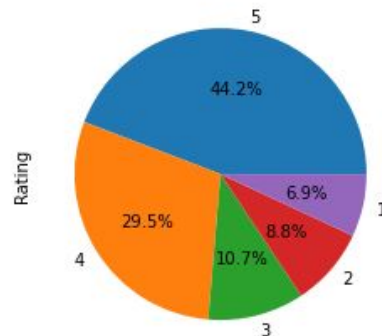
$$M_o = Li + \left(\frac{\Delta 1}{\Delta 1 + \Delta 2} \right) * C \quad (4)$$

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

$$Var(X) = \frac{\sum_1^n (x_1 - \bar{X})^2}{n} \quad (5)$$



(a) Hist: hoteles segun la calificacion de usuario



(a) Porcentaje de calificación según experiencia del usuario

3. Analisis

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

4. Conclusiones

Como se puede observar en la investigación 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 calificación de los huéspedes que mas se repite para los hoteles es de 5.0 estrellas

5. Links video

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

6. 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>