RETO 3

Procesamiento del Lenguaje Natural

Análisis de sentimientos en Amazon

Contexto del caso práctico

En esta práctica, imaginaremos que te has incorporado a uno de los equipos de análisis de datos de Amazon. En Amazon, están interesados en hacer un seguimiento de las opiniones que los clientes dejan en la web en relación a productos que han adquirido. En este caso, les gustaría tener información acerca del grado de satisfacción con el altavoz inteligente conocido como **Alexa**. Sin la ayuda de modelos de IA, esta tarea de procesamiento de volúmenes de datos masivos sería seguramente inabarcable, así que en esta práctica deberás computerizar el proceso de análisis de datos de forma que Amazon pueda realizar predicciones lo más fiables posibles sobre el sentir de la clientela en torno a este producto.

Objetivo: partiendo de la información recopilada por la compañía en torno al producto, se te ha encomendado **preparar un modelo que pueda predecir el grado de satisfacción de los clientes con el mismo.**

Duración

La práctica puede realizarse en unas 2-3 horas aproximadamente.

Recursos

En "Aules" encontrarás el *dataset* a usar en la práctica.

Fases para la elaboración de la solución

Optar por un modelo de tipo *Naive Bayes* que tenga en cuenta las impresiones (sentimientos) de los clientes en torno al producto puede ser una buena forma de abordar este problema. En este apartado tienes una descripción del conjunto de fases que tienes que cubrir para poder dar una solución al problema planteado. La solución requerirá modelizar un *pipeline*, que debe contemplar las siguientes fases:

1) exploración inicial de los datos

Para empezar, una vez cargado el *dataset* en un Pandas, necesitamos entender el conjunto de datos que vamos a manejar[1]. Por ello, incluye en tu cuaderno la representación gráfica[2] más adecuada que muestre los siguientes aspectos (en el pie de página tienes orientaciones para realizar cada parte):

a) existencia de datos vacíos.
b) puntuaciones de los clientes (columna *rating*).
c) la extensión (es decir, el número de palabras) de las opiniones.
d) un par de opiniones cuya extensión ronde la media numérica de palabras tomando como referencia el *corpus* de de opiniones[3].
e) palabras con mayor presencia en el *corpus* de opiniones[4].

2) preprocesado del texto

En esta fase se hace necesario preprocesar el texto con las opiniones como paso previo a la construcción del modelo. Codifica a tal efecto una función que elimine tanto los signos de puntuación como las palabras irrelevantes, de forma que pueda ser aplicada posteriormente a la columna correspondiente del Pandas.

[1] Para una impresión general de los datos, puedes apoyarte en los métodos "describe" e "info" de Pandas.
[2] Puedes usar para ello la librería **Matplotlib** o **Seaborn**, eligiendo entre las diferentes formas de visualizar datos (gráficos de barras, histogramas, de tipo zona de calor, etc.)
[3] Ayuda: en este caso puede ser útil añadir una columna adicional al Pandas (por ejemplo, puedes llamarla "length") con el número de palabras de cada opinión.
[4] Sugerencia: se recomienda en este caso usar algo como la nube de palabras (wordcloud). Consulta en internet cómo implementarla.

3) paso previo a la preparación del modelo

Abordar este problema desde *Naive Bayes* hace del mismo un caso de aprendizaje automático supervisado. Sin embargo, si revisamos el *dataset* del que partimos, comprobamos que no incluye expresamente una columna que haga las veces de etiquetado como sí ocurría en los casos que hemos realizado previamente en torno a la detección de SPAM o la identificación del remitente de cierto mensaje anónimo.

Para poder resolver este caso desde *Naive Bayes*, necesitamos, por tanto, generar valores que expresen las impresiones (positivas o negativas) del cliente hacia el producto a partir de la información con la que contamos de origen.

Se pide:

Añadir una columna llamada "feedback" con los valores **1** (que significará tener una opinión positiva) y **0** (que significará tener una opinión positiva negativa) a partir de las puntuaciones dadas por los clientes de Amazon. Para ello, a fin de reducir las puntuaciones a dos únicos casos (clases) consideraremos como "1" aquellas opiniones de los clientes que hayan tenido una puntuación igual o superior 3, y "0" a las que hayan sido menores a 3.

4) creación y análisis de resultados de un modelo *Naive Bayes*

En esta última fase, deberás crear un modelo multinomial de tipo *Naive Bayes* que sirva para hacer predicciones sobre las opiniones de los compradores de "Alexa Echo".

Se pide ahora que muestres:

a) la matriz de confusión resultante.
b) el informe con las métricas.

Contesta finalmente a las siguientes preguntas (también en el cuaderno):

1. **¿Qué podemos decir sobre la fiabilidad de la predicción del modelo?**

   Justifica tu respuesta en el cuaderno que entregues con la solución del ejercicio.

2. **¿Qué similitudes y diferencias comparte el planteamiento, desarrollo y resolución de este caso aplicando Naïve Bayes con respecto a los dos casos tratados en clase (es decir, de las prácticas relativas al analizador de SPAM, por un lado, y la identificación de remitente anónimo de una carta, por otro)?**

   Justifica igualmente la comparación que hagas de los 3 casos incluyendo tu respuesta en el cuaderno que entregues con la solución del ejercicio.

Entrega y evaluación

- La práctica deberá entregarse en moodle en formato **cuaderno Jupyter** (con extensión "ipynb").
- El cuaderno será el informe que presentarías al jefe de departamento, **en el que debes dar respuesta de una forma argumentada a las preguntas y apartados previos** (es decir, debe combinar código con la interpretación que haces de resultados).
- La fecha límite de entrega será el **28 de Enero de 2024**.
- La evaluación del ejercicio se hará conforme a los criterios recogidos en la siguiente rúbrica, pudiendo alcanzar un total de **10 puntos** como máximo (de los 100 de que consta todo el curso).

U3 :: Procesamiento del Lenguaje Natural				
Proyecto 1: Análisis sobre opiniones en Amazon				
Criterios	**Puntuación**			
	0	**1**	**2**	**3**
Exploración inicial de los datos	No se realiza ninguna tarea relacionada con la exploración de datos	Se dejan varios de los 5 apartados indicados sin cubrir, o bien se cubren todos pero de forma incorrecta en más de un caso	Se deja alguno de los 5 apartados indicados sin cubrir, o bien se cubren todo pero de forma incorrecta en alguno de los casos (por ejemplo, no eligiendo las formas de representación gráfico de los datos más	Se cubren correctamente los 5 apartados indicados, eligiendo las formas de representación gráfico de los datos más adecuada

			adecuada)	
Preparación del modelo	No se realiza ninguna tarea de preparación del modelo	Se deja sin completar o no se realiza correctamente alguna de las tareas indicadas de preparación del modelo (preprocesado y generación de columna con impresión del cliente del cliente)	Se realizan completamente las tareas indicadas de preparación del modelo (preprocesado y generación de columna con impresión del cliente del cliente)	---
Creación del modelo e interpretación de resultados	No se crea el modelo, ni se realiza de forma correcta la vectorización previa	Se realiza la parte de vectorización, pero no se crea el modelo correctamente, ni se llega a mostrar la matriz de confusión y métricas del modelo	Se muestra la matriz de confusión y métricas del modelo, sin interpretar del todo correctamente los resultados	Se muestra la matriz de confusión y métricas del modelo, interpretando correctamente los resultados
Análisis de casos prácticos	No se hace mención a los casos previos	Se deja algún casos sin comparar, o bien se hace, sin ofrecer argumentos concluyentes que demuestren una comprensión de los mismos	Se compara correctamente los 3 casos, ofreciendo argumentos que demuestran una comprensión de Naïve Bayes y su utilidad	---