**Contexto**

Tengo la respuesta de los comentarios de los clientes sobre una encuesta en un flujo de ecommerce para una aerolínea. Dicha información está almacenada en un Excel, donde cada fila corresponde a la respuesta de un cliente único. Cabe considerar que no todas las respuestas son obligatorias, por lo que podemos tener celdas vacías.

Con esta información, necesito hacer una categorización de los comentarios y separar dichas categorías según el sentimiento del mensaje. Con esto me refiero a analizar cada comentario, clasificar dentro de este de 1 a 3 categorías que sean relacionadas a ecommerce y/o a una aerolínea. Estas categorías pueden ser del tipo: Sitio web, Cambios, Devoluciones, Precios, Packs, Vuelos disponibles, Entendimiento del sitio, entre otro, pero no cerrarse solo en este tipo de comentarios, si no que proponer los que hagan más sentido, manteniendo la lógica.

Respecto al tema de sentimiento, hacerlo al nivel de categorías, es decir, si un comentario tiene 2 categorías poner el sentimiento a cada una, esto puede pasar en el siguiente ejemplo: “El sitio funciona muy bien pero no tienen horarios de vuelo que me acomoden”

En este caso, el comentario contempla la categoría “Sitio Web” con un sentimiento positivo, y la categoría “Disponibilidad de horarios” con un sentimiento negativo.

Replicar estas lógicas para cada comentario y genera un archivo Excel con la categoría y el sentimiento para cada comentario de cliente. Si tiene más de uno, agrega las columnas Cat1, Sen1, Cat2, Sen2, Cat3, Sen3 según corresponda.

**Archivo de datos**

* Nombre del archivo: CSAT\_CLIENTES.xlsx
* La hoja de cálculo principal se llama: Detalle
* Los comentarios están en todas las columnas que tienen dentro del título la palabra “Otro”. Acá algunos ejemplos: “Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación con respecto a las Tarifas/Precios de los vuelos-Otro” / “Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación con respecto al Pack-Otro”
* Otras columnas relevantes incluyen: Cada pregunta de alternativas tiene vinculado una pregunta de CSAT, por ejemplo: La pregunta de CSAT “Ingreso datos pasajeros” tiene asociado todas estas preguntas de alternativas + la de comentarios: “Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación en cuanto a los Datos por pasajeros-No me gusta tener que dar mis datos personales”/” Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación en cuanto a los Datos por pasajeros-No es clara la información que me piden”/” Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación en cuanto a los Datos por pasajeros-No me dan la opción de asistencia para pasajero”/” Selecciona la(s) opción(es) que mejor exprese(n) el motivo de tu baja calificación en cuanto a los Datos por pasajeros-Otro”
* El archivo contiene aproximadamente 301 filas de datos

**Objetivos del análisis**

Necesito que me ayudes a:

1. Cargar y preprocesar los datos del archivo Excel
2. Realizar un análisis de sentimiento de los comentarios (positivo, negativo, neutro)
3. Identificar temas y categorías comunes mencionados en los comentarios
4. Encontrar palabras clave y frases recurrentes (considerando únicamente palabras asociadas a ecommerce y aerolíneas)
5. Visualizar los resultados de manera clara y comprensible
6. Proporcionar recomendaciones basadas en los hallazgos

**Pasos detallados que necesito**

Por favor, genera código en Python que:

**1. Cargue y prepare los datos**

* Código para importar el archivo Excel utilizando pandas
* Limpieza de datos: eliminar valores nulos, duplicados y caracteres especiales innecesarios
* Normalizar texto (minúsculas, eliminar acentos si es necesario)
* Preparar los comentarios para el análisis de texto

**2. Realice análisis de sentimiento**

* Utilizar NLTK, TextBlob o cualquier biblioteca adecuada para análisis de sentimiento
* Clasificar cada comentario como positivo, negativo o neutro
* Calcular puntuaciones de polaridad y subjetividad
* Crear nuevas columnas en el DataFrame con estos resultados

**3. Extraiga temas y categorías**

* Implementar técnicas de procesamiento de lenguaje natural para identificar temas comunes
* Utilizar métodos como LDA (Latent Dirichlet Allocation) o NMF para modelado de tópicos
* Agrupar comentarios similares
* Identificar las principales categorías de feedback

**3. Identifique palabras clave y frases**

* Extraer n-gramas más frecuentes (unigramas, bigramas, trigramas)
* Calcular métricas TF-IDF para identificar términos importantes
* Crear nubes de palabras u otras visualizaciones de frecuencia

**5. Visualice los resultados**

* Generar gráficos que muestren la distribución de sentimientos
* Crear visualizaciones para los temas principales
* Representar tendencias temporales si hay datos de fecha disponibles
* Utilizar matplotlib, seaborn o plotly para gráficos interactivos

**6. Genere un informe de hallazgos**

* Resumir los principales insights encontrados
* Proporcionar estadísticas clave sobre los comentarios
* Sugerir áreas de mejora basadas en los comentarios negativos
* Destacar puntos fuertes identificados en comentarios positivos

**Requisitos técnicos**

* Prefiero que uses Python 3.x
* Bibliotecas recomendadas: pandas, numpy, matplotlib/seaborn/plotly, nltk/spacy/textblob, scikit-learn
* El código debe incluir comentarios explicativos
* Incluye manejo de errores para casos problemáticos

**Formato de salida**

* Proporciona el código completo paso a paso
* Incluye instrucciones para instalar bibliotecas necesarias
* Explica brevemente qué hace cada sección de código
* Sugiere posibles interpretaciones de los resultados

**Adaptaciones específicas**

* El idioma principal de los comentarios es: [español/inglés/otro]
* [Cualquier otra característica específica de tus datos que pueda ser relevante]

Hola!  
Necesito que armes un proceso para poder categorizar comentarios de clientes. Este proceso debe considerar:  
-Categoría: Esto corresponde al tópico principal del comentario  
-Subcategoría: Esto corresponde al subtópico que se puede llegar a mencionar dentro del comentario. Este campo se debe llenar solo si aplica.  
-Tipo: Corresponde a la clasificación de la categoría. Esto busca poder detalle rápidamente a que hace referencia la categoría.  
Considera las siguientes categorías:  
-Datos\_Pasajero: Categoría corresponde a comentarios relacionados con el ingreso de los datos de los pasajeros  
-Website: Comentarios asociados al proceso de compra general en el sitio web  
-Proceso\_Pago: Comentarios asociados especificamente al proceso de pago  
-Discount\_Club: Dentro de la aerolínea tenemos una membresía llamada "Discount Club", cualquier comentario asociado a esta membresía debe quedar en esta categoría  
-Promociones: Cualquier comentario relacionado a descuentos, ofertas, promociones (esto es diferente al discount club)  
-Precios: Comentarios asociados a los precios (puede ser de vuelos, equipaje, etc)  
-Disponibilidad\_Vuelo: Comentarios asociados a temas de disponibilidad de horarios, destinos o conexiones de vuelos  
-Aeropuerto: Cualquier tema relacionado al aeropuerto y que sea externo a un proceso de compra en la web  
-Seats: Temas relacionados con los asientos del avión (compra, selección, tipo, comodidad, etc)  
-Equipaje: Todo lo asociado a los equipajes  
-Cambios\_Devoluciones (C&D): Cualquier tema relacionado con cancelaciones y/o devoluciones de vuelos  
-Otros: Si no queda en ninguna de las categorías anteriores, que lo deje en esta categoría