PRACTICA 05

Formación en NLP e interpretación de textos + extracción de contenido

Entregable 1: Análisis de tecnologías y costos asociados

El desarrollo de un sistema de detección de solicitudes de productos a partir de texto implica el uso de varias tecnologías clave:

1. **Lenguaje de Programación y Entorno de Desarrollo:**
   - Python: Con su amplia comunidad, flexibilidad y una gama robusta de bibliotecas, Python es una elección sólida. Además, puede emplearse en entornos como Jupyter Notebooks o IDEs como VSCode para la construcción del sistema.
2. **Librerías de Manipulación y Análisis de Datos:**
   - Pandas y NumPy: Estas librerías son fundamentales para la manipulación, limpieza y procesamiento de datos estructurados. Son de código abierto y ofrecen herramientas poderosas para trabajar con conjuntos de datos.
3. **Algoritmos de Similitud y Procesamiento de Texto:**
   - Cosine Similarity (scikit-learn): Utilizado para calcular la similitud entre vectores o matrices, esta función es esencial en el procesamiento de texto y la comparación de solicitudes con productos existentes.
4. **Vectorización de Texto:**
   - CountVectorizer (scikit-learn): Una herramienta clave para la conversión de texto en vectores numéricos, facilitando su procesamiento y análisis en algoritmos de aprendizaje automático.
5. **Costos Asociados:**
   La mayoría de estas tecnologías son de código abierto y no tienen costos asociados. Sin embargo, en proyectos que involucren almacenamiento masivo de datos o utilización intensiva de servicios en la nube para entrenamiento de modelos, plataformas como AWS pueden ofrecer niveles de servicio gratuitos con limitaciones o niveles de servicio de pago.

Considerando la preferencia por tecnologías de código abierto y sin costos asociados, el uso de bibliotecas como NumPy, Pandas, CountVectorizer y la funcionalidad de cosine_similarity en scikit-learn es una elección sólida para este proyecto.

Entregable 2: Desarrollo y entrega

Para realizar el codigo se han seguido los siguientes pasos:

1. Importar librerías:
- Importar las librerías necesarias (pandas, scikit-learn, etc.).

2. Cargar el dataset:
- Cargar el conjunto de datos que contiene la información de los productos.

3. Manejo de datos faltantes:
- Verificar si hay registros nulos o vacíos en el dataset.

4. Eliminación de registros nulos:
- Si se encuentran registros nulos, eliminarlos del conjunto de datos.

5. Identificación de registros duplicados:
- Revisar si existen registros duplicados en el dataset.

6. Normalización de texto:
- Eliminar tildes u otros caracteres especiales del texto de los productos para una comparación más precisa.

7. Cálculo de similitud coseno:
- Utilizar la similitud del coseno para comparar la frase del pedido del cliente con los registros del dataset.

8. Localización del producto y su ID:
- Identificar el producto más similar a la frase del pedido y obténer su ID.

9. Extracción de la cantidad:
- Extraer la cantidad mencionada en la frase del cliente.

10. Conversión de cantidad texto a numérica:
- Si la cantidad está expresada en texto, convertirla a su representación numérica.

11. Obtención de la cantidad final en formato numérico:
- Obténer la cantidad final en formato numérico para el producto.

12. Verificación del resultado final:
- Verificar y asegúrar que la cantidad y el producto identificado sean correctos con varios ejemplos.