Primer entrega de proyecto

Por
MARIA ISABEL RUA VELEZ

Materia
Introducción a la ingeniería web

Profesor
Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN 2023

1. **Planteamiento del problema**

   Para este proyecto se escogió un dataset de propuestas de trabajo, para desarrollar un modelo de predicción que aprenda de las descripciones de trabajo que son fraudulentas. Utilizando un procesamiento de lenguaje natural.

2. **Dataset**

   Vamos a usar el dataset de kaggle **Real / Fake Job Posting Prediction (**https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction**),** que tiene 18.000 numero de muestras, con las siguientes características:

- **Job_id**
- **title -** Título de la publicación
- **location -** Ubicación del lugar de trabajo
- **department -** Departamento de trabajo
- **salary_range -** Rango de salario
- **company_profile -** Perfil de la compañía
- **description -** Descripción del trabajo
- **requirements -** Requerimientos para el trabajo
- **benefits -** Beneficios con el empleo
- **telecommuting -** 0 o 1 si se realiza teletrabajo
- **has_company_logo -** Si tiene logo de la empresa (0 o 1)
- **has_questions -** Si tiene preguntas (0 o 1)
- **employment_type -** Tipo de empleo (tiempo completo, medio tiempo, entre otros)
- **required_experience -** Experiencia requerida
- **required education -** Educación requerida
- **industry -** industria en donde se desarrolla el trabajo
- **function -** Función del trabajo
- **fraudulent -** Característica de salida (0 o 1)

3. **Avances**

   El código presentado realiza una serie de operaciones para la limpieza y preprocesamiento de un conjunto de datos que contiene información sobre publicaciones de trabajo. El objetivo principal es transformar los datos en un formato que sea adecuado para su posterior análisis y modelado.

   En primer lugar, se carga el conjunto de datos de trabajo en un objeto pandas DataFrame utilizando la función read_csv(). Luego, se crea una copia del conjunto de datos original para evitar modificar los datos originales.

A continuación, se realiza una exploración de los datos para identificar aquellas características que tengan más del 60% de valores nulos y se eliminan del conjunto de datos original. Esto se realiza para reducir el tamaño del conjunto de datos y centrarse en las características que tienen una cantidad significativa de datos disponibles.

Después, se rellenan los valores nulos con una cadena vacía utilizando la función fillna(). Esto se hace para evitar que los valores nulos afecten el análisis posterior del conjunto de datos.

Luego, se eliminan algunas columnas que no son útiles para el análisis posterior utilizando la función drop(). En este caso, las columnas que se eliminan son 'telecommuting', 'has_company_logo', 'has_questions' y 'job_id'.

A continuación, se combinan varias columnas del conjunto de datos en una sola columna llamada 'text'. Esto se hace para tener todas las características relevantes en un solo lugar y facilitar el preprocesamiento posterior del texto.

Después, se realizan una serie de operaciones de limpieza en el texto. En primer lugar, se reemplazan los saltos de línea, saltos de línea y espacios que son tabs con espacios en blanco. Luego, se eliminan los números y los caracteres especiales utilizando expresiones regulares y la función apply(). Finalmente, se convierte todo el texto en minúsculas utilizando la función apply().

En general, el código presentado intenta limpiar y preprocesar los datos antes de analizarlos y modelarlos. La eliminación de características con muchos valores nulos, la combinación de características relevantes en una sola columna y la eliminación de caracteres especiales y números son técnicas comunes para limpiar el texto antes del análisis posterior.

4. Conclusiones

En conclusión, el código presentado realiza una serie de operaciones de limpieza y preprocesamiento de un conjunto de datos que contiene información sobre publicaciones de trabajo. La limpieza se realiza de esta manera debido a que los datos son en lenguaje natural, lo que significa que pueden contener una gran cantidad de ruido y características irrelevantes que pueden afectar la calidad del análisis posterior.

La eliminación de características con muchos valores nulos, la combinación de características relevantes en una sola columna, la eliminación de caracteres especiales y números, y la conversión de todo el texto en minúsculas son técnicas comunes para limpiar el texto antes del análisis posterior. Estas técnicas permiten reducir el ruido en el conjunto de datos y asegurar que las características relevantes se identifiquen con precisión.

Es importante tener en cuenta que la limpieza y preprocesamiento de datos es una parte esencial del proceso de análisis de texto y modelado. Si no se realiza una limpieza adecuada de los datos, los modelos resultantes pueden ser imprecisos y no proporcionar resultados útiles. Por lo tanto, es importante seguir buenas prácticas de limpieza de datos para garantizar que los modelos de análisis de texto sean precisos y útiles para la toma de decisiones.