Propuesta para la creación del Modelo de Limpieza de Datos en la plataforma Hugging fase.

Autores

Wilmar José Martínez
Wilson Ramírez Vega
Edwin Peñarredonda Novoa

Asesores tutor:

Alexander Toscano Ricardo

Universidad de Córdoba
Facultad de Educación y Ciencias Humanas
Licenciatura en Informática con énfasis en Medios Audiovisuales

2024

Propuesta para la Creación de un Modelo de Limpieza de Datos en la Plataforma Hugging Face

1. Introducción

La limpieza de datos es una etapa fundamental en el flujo de trabajo de ciencia de datos. Este proceso asegura que los datos empleados en el entrenamiento y evaluación de modelos sean precisos, consistentes y de alta calidad, factores esenciales para optimizar el rendimiento de aplicaciones de procesamiento de lenguaje natural (NLP). Este proyecto propone desarrollar un modelo en la plataforma Hugging Face, orientado a realizar la limpieza automática de conjuntos de datos textuales, mejorando así la calidad y confiabilidad de los datos en diversas aplicaciones de NLP.

2. Objetivo General

Desarrollar un modelo automatizado de limpieza de datos enfocado en mejorar la calidad de datasets textuales, mediante la eliminación de errores, duplicados, inconsistencias y otros problemas comunes presentes en los datos sin procesar.

3. Objetivos Específicos

- Implementar funciones de preprocesamiento: Estas funciones se centrarán en tareas de eliminación de duplicados, corrección de errores ortográficos, normalización de texto y eliminación de ruido (caracteres no deseados, espacios innecesarios, etc.).

- Diseñar un modelo que identifique y corrija inconsistencias en el etiquetado de datos.

- Desarrollar una API accesible para la integración del modelo con otros proyectos de NLP en Hugging Face.

- Realizar pruebas exhaustivas utilizando diferentes datasets de NLP en Hugging Face, evaluando la eficacia y robustez del modelo.

4. Justificación

El volumen creciente de datos no estructurados disponibles en la web y otras fuentes requiere herramientas avanzadas de limpieza de datos. Un modelo de limpieza automatizado facilitará el preprocesamiento de datos, ahorrando tiempo y mejorando la calidad de los modelos de aprendizaje automático y profundo que dependen de estos datos. Este proyecto pretende contribuir al campo del NLP con una herramienta robusta de limpieza de datos que será de gran valor para la comunidad de desarrolladores y profesionales en Hugging Face.

5. Metodología

5.1 Análisis del Problema

Se identificarán las principales deficiencias que suelen observarse en datasets textuales, como errores ortográficos, duplicados y formatos inconsistentes. Este análisis permitirá definir los pasos de limpieza específicos necesarios para el proyecto.

5.2 Recolección de Datos

Se emplearán datasets disponibles en la plataforma Hugging Face para el desarrollo y prueba del modelo. Esta selección garantizará que los datos usados representen los desafíos comunes en NLP.

5.3 Diseño del Modelo

Se desarrollará un pipeline de procesamiento de datos que incluya módulos para:
- Tokenización y normalización de texto.
- Eliminación de duplicados.
- Corrección de errores ortográficos (utilizando herramientas como SymSpell o Hunspell).
- Identificación y limpieza de ruido textual, eliminando elementos irrelevantes como enlaces web y símbolos no deseados.

5.4 Implementación Técnica

Se utilizarán frameworks de desarrollo como PyTorch o TensorFlow para construir el backend del modelo. Además, se diseñará una API que permita a otros desarrolladores integrar el modelo de limpieza en sus proyectos.

5.5 Pruebas y Evaluación

Se validará el modelo empleando diversas métricas (precisión, recall, F1-score) para evaluar la efectividad del proceso de limpieza en diferentes datasets. La selección de estas métricas permitirá medir la eficiencia y robustez del modelo en escenarios reales.

6. Tecnologías a Utilizar

- Plataforma Hugging Face: Transformers, datasets.
- Lenguaje de Programación Python (con bibliotecas de soporte: pandas, nltk, spacy).
- Frameworks de Desarrollo: PyTorch/TensorFlow.

- Herramientas de Corrección Ortográfica: SymSpell, Hunspell.
- Integración con Hugging Face Hub para compartir el modelo y su documentación con la comunidad.

7. Resultados Esperados

- Un modelo robusto de limpieza de datos alojado en Hugging Face, accesible y útil para la comunidad.
- Documentación completa que permita a otros desarrolladores integrar y adaptar fácilmente el modelo en proyectos de NLP.
- Mejoras sustanciales en la calidad de datasets utilizados por otros modelos, facilitando el desarrollo de aplicaciones más precisas y efectivas.

8. Conclusiones y Futuras Aplicaciones

Se espera que este modelo de limpieza de datos se convierta en una herramienta esencial para el preprocesamiento de textos en proyectos de NLP. A futuro, se prevé la adaptación del modelo a otros idiomas y aplicaciones más específicas, extendiendo su impacto y utilidad en proyectos de NLP.