PROYECTO FINAL - PRIMER ENTREGA

JUAN SEBASTIÁN ALVAREZ ERAZO
LEONARDO ANGEL SANCHEZ
EDINSON CACERES PARRA
SARA FERNANDA SANCHEZ SANCHEZ

MACHINE LEARNING TECHNIQUES

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LOS ANDES

TABLA DE CONTENIDO

Objetivos del proyecto

Objetivo principal

Construir un modelo basado en Procesamiento de Lenguaje Natural (Natural Language Processing - NLP) que mediante un API permita al usuario interactuar con él, y que a partir de un texto y preguntas relativas al mismo sea capaz de responderlas de forma coherente.

Objetivos específicos

1. Construir y entrenar el modelo de NLP.
2. Evaluar los resultados obtenidos del modelo.
3. Ajustar el modelo con técnicas de regularización, reducción dimensional, etc. de ser necesario.
4. Crear un API que consuma el modelo y sus resultados.
5. Desplegar el API en una plataforma web.
6. Garantizar la disponibilidad 24/7 de la plataforma y el funcionamiento de la API.

Justificación

Con el fin de desarrollar un modelo que permita conocer las necesidades de un usuario y poder guiarlo dadas ciertas interacciones con el mismo, a manera de ChatBot, se pretende configurar el espacio de desarrollo dentro de Collab y el API correspondiente que permita desplegar este desarrollo, siendo proporcionado por el canal adecuado al usuario final.

Propuestas analizadas

Para escoger el proyecto expuesto anteriormente, que se trabajará a lo largo del semestre, se presentaron y evaluaron distintas alternativas, todas enfocadas en el algoritmo de NPL, pues es una herramienta con gran potencial tanto para identificación de intenciones, sentimientos, solicitudes, quejas, etc:

1. **Resúmenes de texto**: Se cuenta con un texto inicial, luego se entrena un modelo que por medio de palabras clave recorre todo el documento, cuya salida es el resumen del texto: Aquí se encuentra la fuente: https://huggingface.co/datasets/scientific_papers.

2. **Ropa de mujer:** Contiene información, proporcionada por mujeres, referente a sus compras y opiniones sobre diverso tipo de prendas de vestir, se incluyen la edad de la persona, la review en texto, el rating dado y el positive feedback. Lo que se busca es entrenar un modelo supervisado, el cual, al ingresar una nueva *review* sobre una prenda, sea capaz de asignar un puntaje, dadas las palabras clave usadas por las mujeres al calificar prendas anteriormente.(Adjunto en Data, *women_clothing_data.csv*)

3. **Noticias de la BBC:** El dataset se compone de diferentes txt que contienen noticias, clasificadas dada la carpeta en la que se encuentran, al hacer este etiquetado principal, se establecen palabras clave para cada tipo de noticia. Con esto, usando NLP se espera introducir un nuevo texto y que devuelva el tipo de noticia en que se etiquetara. (Adjunto en Data, *Noticias BBC.zip*)

4. **Preguntas y respuestas:** Se pretende realizar un modelo que con base a una pregunta realizada por un usuario, se entregue una respuestas basada en la información de Wikipedia: Fuente: https://rajpurkar.github.io/SQuAD-explorer/

5. **Preguntas y respuestas:** Emulando el funcionamiento de un ChatBot, se pretender elaborar un modelo capaz de entablar conversación con el usuario, mediante preguntas prediseñadas un de un pasaje de texto aleatorio: Fuente: https://stanfordnlp.github.io/coqa/ (Revisar punto 4).

La propuesta ganadora es la 5, ya que teniendo en cuenta el alcance del proyecto, su alineación con la temática de la materia y la capacidad del equipo de trabajo, se encuentra como la mejor opción para desarrollar un mínimo producto viable.

Diccionario de datos

Los datos los tenemos disponibles en un JSON que tiene la siguiente estructura:

Atributo	Atributo padre	Descripción
source	No aplica	Origen del archivo a analizar
id	No aplica	Código asignado al texto
filename	No aplica	Nombre de archivo a analizar
story	No aplica	Texto a analizar
questions	No aplica	Array de preguntas

input_text	questions	Pregunta
turn_id	questions	ID de la pregunta
answers	No aplica	Array de respuestas obtenidas
span_start	answers	Posición donde inicia el texto relacionado a la pregunta
span_end	answers	Posición donde finaliza el texto relacionado a la pregunta
span_text	answers	Texto de donde se obtiene la respuesta
Input_text	answers	Texto con la respuesta a la pregunta
turn_id	answers	Consecutivo de la pregunta

Tabla 1: Diccionario de datos. Elaboración propia.

Anexos

Paper 1: Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/pdf/1810.04805.pdf

Paper 2: Siva Reddy, Danqi Chen, Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. https://arxiv.org/pdf/1808.07042.pdf

Paper 3: Chengyu Wang, Minghui Qiu, Jun Huang, Xiaofeng He. 2020. Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining. https://arxiv.org/pdf/2003.13003v2.pdf

Referencias

Hugging Face (2018). *Dataset: Scientific Papers.* Recuperado el 11 de Octubre de 2022 de: https://huggingface.co/datasets/scientific_papers.

The Stanford NLP Group. (2020). *CoQA: A Conversational Question Answering Challenge*. Recuperado el 11 de Octubre de 2022 de: https://stanfordnlp.github.io/coqa/

The Stanford NLP Group. (2022). *SQuAD 2.0: The Stanford Question Answering Dataset*. Recuperado el 11 de Octubre de 2022 de: https://rajpurkar.github.io/SQuAD-explorer/

Wang C., Qiu M., Huang J., y He X. (2020). *Meta Fine-Tuning Neural Language Models for Multi-Domain Text Mining.* Recuperado el 11 de Octubre de 2022 de: https://arxiv.org/pdf/2003.13003v2.pdf

Devlin J., Chang M., Lee K., y Toutanova K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Recuperado el 11 de Octubre de 2022 de: https://arxiv.org/pdf/1810.04805.pdf

Reddy S., Chen D., Manning C. (2019). *CoQA: A Conversational Question Answering Challenge.* Recuperado el 11 de Octubre de 2022 de: https://arxiv.org/pdf/1808.07042.pdf