082057 – Procesamiento del Lenguaje Natural

Trabajo Práctico Individual

CODIGO: 01 – Plagio

1 LINEAMIENTOS

- El trabajo práctico es 100% individual.

- El dataset se encuentra en: https://drive.google.com/drive/folders/1EBR8aSUM67LLpDeNrQk3TN5XOx3zUJCg?usp=sharing

- Todo el código fuente desarrollado (si hubiera) así como el documento de texto que lo acompaña con las justificaciones, técnicas, bibliografía, autores y demás correctamente citados, deben ser subidos a la cuenta de Github https://github.com/ **personal y privada** de cada estudiante y compartida con la cuenta de github del profesor (https://github.com/hernanborre). Sólo así se considera el trabajo entregado.

- El/La estudiante, deberá realizar una **defensa oral** de su código, técnicas utilizadas, citar autores utilizados si es necesario y poder expresar claramente tanto su desarrollo cognitivo a la solución, así como las conclusiones obtenidas.

- Este trabajo práctico constituye la única evaluación en primera instancia de esta materia, por lo cual la producción de la solución del mismo se espera que esté a la altura o supere el tiempo dedicado a un parcial.

- Si el TP no fuera entregado a tiempo o en forma, los recuperatorios serán un parcial escrito que constará con 5 preguntas sobre los temas vistos en la materia y/o detallados en el programa.

- Fecha/s de entrega: **24 de Junio y 8 de Julio de 2022**

2 CONSIGNA

Se pide desarrollar un sistema que pueda detectar **el nivel de plagio** de un Trabajo Práctico **T**, presentado por un alumno **X**, incluyendo la posibilidad de detectar **parafraseos** que no hayan sido debidamente **citados**.

Esta detección no se limita a los trabajos prácticos de otros años, presentados por otros alumnos x_i sino que además se deberá tener en cuenta la **web, libros, papers, conferencias**

-proceedings- y cualquier otra fuente de información de la cuál se pudiera estar cometiendo plagio.

Además de detectar el plagio, **se deberá indicar claramente, en qué líneas o palabras (ubicación relativa)** se encontró, a quien o quienes se está plagiando y dar la posibilidad de ver o revisar ambos textos (pudiendo ser referencias a links, trabajos de otros alumnos, libros, etc).

También se necesita mostrar **un intervalo de confianza de la predicción.**

La máxima nota se consigue cuando se pueda identificar parafraseo del texto a verificar sobre el texto un **sitio web** al cuál no se hace referencia explícita.

2 ENTREGA, CODIGO Y EJEMPLOS

El código deberá ser escrito en Python 3. Se pueden usar todas las librerías que se crean necesarias.
El dataset para entrenar y probar el modelo se podrá descargar desde el campus virtual o bien con un link a él. El mismo consiste de una serie de TPs reales de una materia real recopilada durante años y sucesivas entregas de estudiantes sobre el mismo o diversos temas en los cuales podría ocurrir plagio entre TPs del mismo año, de distinto año o bien sobre sitios webs o libros o artículos académicos.

Al ejecutar el código deberá recibir como entrada un documento de texto y devolverá en texto la salida:

- Nombre del archivo de texto procesado
- Nombre y Apellido del alumno
- Tópico o Tema del texto procesado
- Porcentaje de plagio del texto en general
- Listado de frases (y su ubicación dentro del texto) que podrían ser plagios de otros TPs previamente subido (fase de entrenamiento) o bien copiados de la web.

Además del código desarrollado, se debe entregar un documento de texto (puede ser un .doc o docx, o bien recomendamos usar la herramienta de generación de textos científicos LaTex), en la cual se explica la solución y se le da crédito a los autores consultados, las técnicas usadas, el código fuente o ejemplos tomados de otros blogs, videos, etc.

Las citas deben ser en formato **APA**