Tecnológico de Monterrey

CAMPUS ESTADO DE MÉXICO

TC3002B

Desarrollo de aplicaciones avanzadas de ciencias computacionales

Miguel González Mendoza

Raúl Monroy Borja

Ariel Ortíz Ramírez

Jorge Adolfo Ramírez Uresti

Fase 3 Evidencia

INTEGRANTES

Jorge Rojas Rivas	A01745334
Nadia Paola Ferro Gallegos	A01752013
Melissa Aurora Fadanelli Ordaz	A01749483

6 de mayo de 2024 (Versión 1.0)

Descripción

Se comienza registrando los archivos de plagio. De los archivos de un directorio que contiene los archivos con plagio, se genera una lista que los contiene. Luego, se ordena la lista en orden alfabético. Este proceso se repite para los archivos que se usarán para construcción y aquellos que se usarán para pruebas.

Como siguiente paso de preparación, se leen todos los archivos de plagio, construcción y prueba. Se definen las *stopwords* con nltk y se limpian los conjuntos de datos de plagio, construcción y prueba. La limpieza conlleva preparar un arreglo vacío para el texto limpio a generar. Luego, se preprocesa por medio de transformación a letras minúsculas y descartando las *stopwords* que se hallen en los textos. Finalmente, se ingresan todas estas palabras en un arreglo.

Para determinar si un texto que se revise en un futuro es plagiado o no, se preparan el modelo y el tokenizador para esta tarea. Se utiliza 'sentence-transformers/bert-base-nli-mean-tokens'. Esto contiene una librería para tareas de procesamiento de oraciones, con el modelo base de BERT y otro para lenguaje natural.

Se almacenan y tokenizan los conjuntos de datos para plagio, construcción y prueba. Esto se realiza por medio de un diccionario para conjunto. Luego, se tokeniza cada oración para ser almacenada en el diccionario. Después, de estos tokens generados, se reformatean para que sean de un solo tensor y se procesan a través del modelo. Finalmente, de estos se consiguen los embeddings.

Se aplican las máscaras a los embeddings de plagio, construcción y prueba. Luego, se obtiene el promedio de los embeddings de estos conjuntos. Finalmente, se calcula la similitud de coseno. La similitud es calculada entre dos conjuntos. El primero son 20 archivos que entre ellos existen algunos que son plagiados. El segundo conjunto son 100 archivos que son originales y han sido usados para entrenar el modelo. Se decidió que cualquier valor de similitud de coseno que superara el umbral de 40% sería considerado plagio.

Para la parte del modelo de *Machine Learning* se empieza por elegir los documentos que se van a utilizar, en este caso se realizaron 100 documentos plagiados, 20 de cada tipo de plagio, y se tomaron los originales correspondientes a cada texto. Estos se agregaron junto con el tipo de plagio que se cometió en cada texto.

Se realizó la división del dataset en entrenamiento y prueba, en este caso se omitió la parte de validación, ya que aunque es prudente tener un set de validación para los modelos, por la escasa cantidad de datos que tenemos elegimos tomar el riesgo de solo entrenarlo y probarlo.

Seguido de esto se realizaron los *embedding* de cada texto con su original utilizando la opción que nos provee BERT para esta acción. Estos *embeddings* se guardaron en el *dataframe* junto con el resto de la información. Después se realiza una tokenización de los textos plagiados y originales que nos da como resultado tensores de *PyTorch* con los IDs, las máscaras de atención y las etiquetas, estos tensores se utilizan para poder entrenar al modelo.

Una vez terminado el proceso de tokenización y *encoding* se realizan *dataloaders* para poder cargar el modelo para su entrenamiento. Debido a la alta complejidad del modelo y los recursos limitados no se pudo entrenar el modelo.

Descripción Mérito Relativo

Debido a la alta complejidad del modelo y los recursos limitados solo se pudo realizar una descripción de mérito relativo.

Para la primera herramienta, se evaluó con los documentos provistos por los profesores junto con la tabla de resultados esperados. El umbral de plagio que se decidió establecer por medio de varias pruebas fue de 40%. El objeto a evaluar fue la salida de resultados que fueron generados comparados con los resultados esperados. De ella, se identificaron 10 resultados verdaderos negativos (TN), 7 verdaderos positivos (TP), 1 falso negativo (FN) y 2 falsos positivos (FP). Estos números resultan en una tasa de verdadero positivo (TPR) de 87.5% y una tasa de falso positivo (FPR) de 16.6%. Finalmente, utilizando el área bajo la curva operacional receptora (AUC) se puede concluir que la AUC de nuestra herramienta es del 85.41%.

Cabe destacar algunas observaciones registradas durante la evaluación de la herramienta. La primera observación es que, en uno de los resultados (el texto FID-010), se detectó un texto plagiado adicional a pesar de que este no debía ser detectado. Esto se debe a la tasa de plagio que fue declarada para declarar un documento como plagio, sin embargo, esta tasa también fue beneficiosa para identificar un caso de plagio (el texto FID-005). La segunda observación es que se identificó un caso de plagio de un texto completamente diferente. Se detectó plagio en el texto FID-020 del texto org-066, sin embargo se tenía que detectar el org-014. Nunca se identificó parte del texto org-014 y el 066 alcanzó una tasa alta de plagio.

Documento Sospechoso	**Copia Detectada**	**Copia Esperada**	**Documento Plagiado**	**Documento Plagiado Esperado**	**% plagio**
FID-001	No	No			
FID-002	No	No			

FID-003	No	No			
FID-004	No	No			
FID-005	Si	Si	org-023	org-023	79.12%
			org-059	org-059	40.39%
FID-006	No	No			
FID-007	No	No			
FID-008	No	No			
FID-009	No	No			
FID-010	Si	Si	org-005	-	40.21%
			org-091	org-091	97.41%
FID-011	No	No			
FID-012	No	No			
FID-013	No	Si	-	org-001	-
				org-009	-
FID-014	Si	Si	org-011	org-011	47.46%
			org-019	org-019	77.84%
FID-015	Si	Si	org-034	org-034	97.32%
FID-016	Si	Si	org-046	org-046	96.66%
FID-017	Si	Si	org-062	org-062	100.00%
FID-018	Si	Si	org-057	org-057	97.44%
FID-019	Si	Si	org-066	org-066	78.31%
FID-020	Si	Si	org-066	org-014	78.31%

Conclusión

Por complejidades del modelo y recursos limitados, no se pudo terminar la segunda herramienta. Sin embargo, para casos más sencillos, la primera herramienta es satisfactoria para identificar plagio en la mayoría de los casos. Es una herramienta que mejorará su desempeño en cuanto más información tenga para entrenar.