Optimización de la Metodología de Detección de Plagio en Informes mediante Modelización Matemática y Text Mining

IGNACIO SANTAMARÍA

PADRÓN: 101613

Email: isantamaria@fi.uba.ar

ARIEL SCHWARTZ

PADRÓN: 101611

Email: aschwartz@fi.uba.ar

Objetivos Específicos: 2 y 4

Objetivos Específicos: 1 y 3

TUTOR: MARIANO BONOLI ESCOBAR

COTUTOR: XAVIER GONZÁLEZ

Fecha de entrega:

ÍNDICE

1. Presentación del problema

1.1. Introducción

1.1.1. Evolución en la Industria de la Educación

Vivimos en un mundo que está en constante cambio. La aparición de nuevas tecnologías disruptivas hace que la evolución no suceda de forma lineal, sino de forma exponencial. Esto puede apreciarse fácilmente si analizamos los lapsos de tiempo entre las distintas revoluciones industriales y tecnológicas, las cuales cada vez se dan más rápidamente y con menor cantidad de años entre ellas.

La educación como industria no queda exenta de estos cambios, y muchas veces la necesidad de adaptarse es inevitable. El hecho histórico más reciente fue la pandemia del Coronavirus, la cual obligó a la población a recluirse en sus hogares. Esto hizo que la educación tenga que adaptarse para poder continuar en formato a distancia. A su vez, el avance tecnológico en el ámbito educativo conlleva ciertos cambios de paradigma, y es fundamental comprender hacia dónde vamos para acompañar estas transformaciones sociales de la mejor manera.

Si bien hoy en día el retorno a la presencialidad es una realidad, indudablemente muchas de las herramientas y recursos utilizados en la modalidad virtual seguirán vigentes. Especialmente las plataformas educativas como Moodle, el Campus virtual de la universidad, la entrega de tareas en formato digital, el correo electrónico, entre otras. Estas herramientas enriquecen el proceso de enseñanza y facilitan la tarea tanto del docente como del alumno, optimizando los tiempos de ambos.

1.1.2. Inconvenientes en la Evaluación Académica a través de Informes

Dada la evolución en la Industria Educativa presente hoy en día, cada vez existe un mayor uso de estas plataformas y por ende muchas materias universitarias se ven dispuestas a generar actividades académicas que consisten en la entrega de informes en formato digital. Sin embargo, en estas actividades realizadas por los estudiantes de manera remota y asincrónica, siguen existiendo situaciones donde se detectan copias entre informes académicos de este estilo.

La detección de estos posibles plagios académicos se dificulta aún más en aquellas materias masivas donde se entregan para corregir cientos de informes de una misma temática. Considerando también que en estas materias son varios los profesores que se reparten para corregir estas tareas, esto dificulta aún más la detección de posibles copias en los análisis que deben ser realizados por los estudiantes.

Ante esta situación muchas materias optan por no tomar evaluaciones a través de informes, dado que no consideran poder tener una certeza de que las tareas no están sujetas a copias y por desconocer si verdaderamente los alumnos pudieron

adquirir los conocimientos necesarios de la materia a través de estas evaluaciones, generando una limitación en la elección de metodologías de enseñanza y evaluación confiables.

Esto conduce a una necesidad de optimizar la metodología de detección de plagios en informes académicos, y aprovechando la conveniencia de la entrega digital de estas tareas, es posible basarnos en herramientas informáticas para el desarrollo de modelos matemáticos y algoritmos que nos ayuden en nuestra metodología de detección de copias.

Este trabajo tiene como objetivo final que muchas materias puedan tener una mayor confiabilidad en la adquisición de conocimientos por parte del estudiante, a través de una evaluación tomada mediante la presentación de informes académicos.

1.1.2. Ciencia de Datos y Procesamiento de Lenguaje Natural

La aplicación de tecnologías basadas en Ciencia de Datos en la Industria 4.0 (la llamada 4ta Revolución Industrial) es uno de los principales hitos tecnológicos en nuestros tiempos. En particular, el uso de la minería de textos y del Procesamiento de Lenguaje Natural, ramas de la Ciencia de Datos, son cada vez más necesarios y recurrentes para el apoyo en la resolución de problemas en cualquier industria.

Para una industria como la educativa, donde los principales recursos y desarrollos de aprendizaje se realizan a través del lenguaje escrito, se presenta una importante oportunidad de utilizar la Minería de Textos (*Text Mining*) como herramienta para la optimización de los trabajos y la toma de decisiones.

El fin de esta práctica es obtener información a partir de datos no estructurados como lo es el lenguaje natural (forma más usual de comunicación entre seres humanos) a través de diferentes técnicas y con diversos objetivos. Esto lo logramos planteando diferentes algoritmos que definen cómo se deben interpretar los textos, qué consideraciones se toman, y qué se tiene en cuenta; con el fin de optimizar nuestra metodología de trabajo a la hora de analizar textos.

El Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés) permite reducir tiempos y aumentar la eficiencia de tareas tales como: conteo de palabras comunes y repetitivas, análisis de los sentimientos y de la carga emocional que lleva el texto, comparación entre dos o varios textos para detectar similitudes, entre otras tantas aplicaciones.

Dado que en la industria educativa existe la necesidad de comparar informes académicos a la hora de evitar casos de plagio en las entregas, podemos plantear el uso del NLP como una herramienta de apoyo para la metodología de detección de similitudes entre estos textos. Es decir, a partir de modelos matemáticos podremos realizar una clasificación que identifique a los pares o grupos de informes académicos que posean mayor similitud en su contenido y redacción.

Además, asociado a la Minería de Textos, se encuentra otra rama de la Ciencia de Datos, el Aprendizaje Automático (*Machine Learning*). El cual, mediante diferentes modelos entrenados con datos, permite hallar patrones, predecir, clasificar, y finalmente optimizar la toma de decisiones y los tiempos de las tareas.

Por ende, mediante un estudio de modelos de NLP y Aprendizaje Automático, se lograría crear una herramienta y una metodología de análisis altamente optimizada. Reduciendo esfuerzos, errores y tiempos a la hora de la detección de plagios en informes académicos.

1.2. Objetivos Específicos

1.2.1. Objetivo Específico 1: *(Ariel Schwartz)*

Realizar un análisis sobre la necesidad de una metodología eficiente en calidad y tiempos para la detección de similitudes entre informes y trabajos académicos dentro de la industria de la educación.

1.2.2. Objetivo Específico 2: *(Ignacio Santamaria)*

Elaborar un algoritmo que contribuya a la detección de similitudes entre textos. Mediante modelos matemáticos basados en distintos tipos de distancias entre textos, apalancados por técnicas de Procesamiento de Lenguaje Natural.

1.2.3. Objetivo Específico 3: *(Ariel Schwartz)*

Elaborar un algoritmo que contribuya a la detección de similitudes entre textos. Mediante modelos matemáticos de Aprendizaje Automático, a partir de los resultados obtenidos en el Objetivo Específico 2.

1.2.4. Objetivo Específico 4: *(Ignacio Santamaria)*

Desarrollar una metodología, junto con las herramientas necesarias, para la estandarización del trabajo paso a paso a realizar por el docente y la interpretación de los resultados para la toma de decisiones.

1.3. Alcance

Lograr desarrollar una metodología y herramienta de comparación de informes académicos eficiente, en calidad y tiempos, que sea de uso sencillo para el docente analista.

Se planea utilizar la primera versión de la metodología dentro del Departamento de Tecnología Industrial y el Departamento de Gestión de la Facultad de Ingeniería de la UBA, para luego extender su aplicación en otras facultades de la UBA, y finalmente en otras universidades.

2. Metodología de trabajo

2.1. Análisis de la situación actual de la Industria de la Educación

Como primer paso se realizará una encuesta a docentes de la FIUBA para conocer sobre la necesidad de esta metodología para la comparación de informes académicos. Buscamos conocer sus opiniones y experiencias en la corrección de trabajos entregados en la facultad de manera digital.

2.2. Procesamiento de Lenguaje Natural

Como datos de base para el desarrollo de los modelos y algoritmos, se utilizarán informes presentados en la materia Estadística Aplicada II. Los mismos serán manipulados y procesados en las siguientes etapas.

A) Selección de textos: seleccionar los textos correspondientes a utilizar para su comparación y análisis.

B) Pre-procesamiento de datos no estructurados: limpiar y depurar la base de datos de textos, dejando aquellos datos que sirvan para obtener información relevante.

C) Transformación de Datos No Estructurados: transformar la base de datos a un corpus apto para entrenar los modelos a desarrollar.

D) Modelización y Minería de datos: búsqueda de similitudes entre textos, agrupamiento y clasificación de los mismos a través de diferentes modelos matemáticos.

E) Interpretación: interpretar los patrones obtenidos para la extracción de información útil en la toma de decisiones.

Durante estas etapas, y particularmente en la Transformación de Datos No Estructurados, se hará uso de las diversas técnicas y conceptos que aportarán al modelizado.
Entre estas se encuentran la tokenización, siendo un caso particular los n-gramas, los cuales consisten en una transformación del texto completo en pequeñas cadenas de texto (tokens) de n palabras, con el fin de lograr un input más acorde para futuros modelos. Otra de estas técnicas es el Word Embedding, que consiste en modelar las palabras de los textos convirtiéndolas en un vector numérico, permitiendo compararlas en un plano n-dimensional.
A su vez también se utilizarán transformaciones como la lematización, que consiste en el reemplazo de palabras que representan lo mismo por un lema. Y técnicas de limpieza del texto como la remoción de Stopwords, es decir, palabras comunes del idioma que podrían generar un importante sesgo en el output del modelo si no se tratan y/o eliminan.

2.3. Modelización mediante distancias entre textos

El uso de distancias entre textos permite desarrollar un modelo, que en la teoría, es altamente explicativo pero por ello cuenta con ciertos límites en la precisión de los resultados obtenidos en la clasificación de los textos.
Existen muchas distancias matemáticas para modelar la similitud o diferencia entre textos. Todas ellas poseen ventajas y desventajas, llevando a que, con un conjunto de distancias acordes podríamos ver la similaridad entre textos en muchas dimensiones (variables). Y así aprovechar las ventajas de los distintos modelos que plantea cada distancia, ponderando sus ventajas y minimizando el impacto de sus desventajas.
Algunas de las distancias más comunes que se tendrán en cuenta serán: Jaccard, Coseno, Hamming, Levenshtein y Word Mover 's Distance (mediante Word Embeddings), entre otras.

2.4. Modelización mediante aprendizaje automático

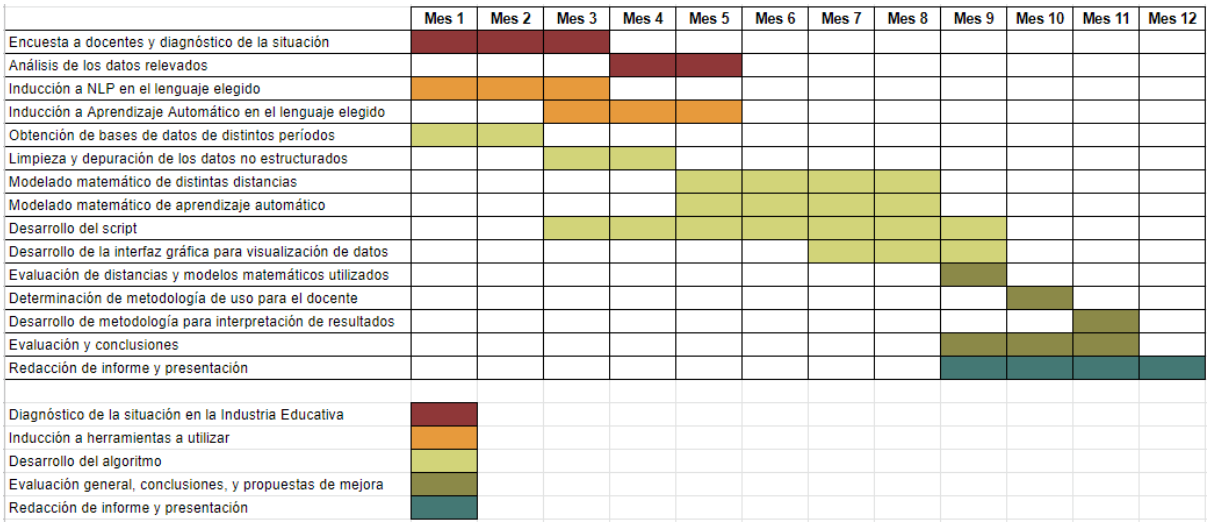
El uso de aprendizaje automático para la comparación de textos permite desarrollar un modelo matemático de mayor precisión y flexibilidad, a costa de una menor explicabilidad, comparado con el modelo de distancias antes planteado.
Este modelo de aprendizaje automático puede recibir como input de datos una o más de las distancias calculadas en el anterior modelo, además de otros datos, cómo por ejemplo metadatos del archivo analizado (ejemplo, el propietario del mismo) o imágenes presentes en el informe. A partir de esto, el algoritmo clasificará los pares o grupos de textos indicando si corresponden a un plagio o no. Aunque en otros casos, también permitirá calcular una probabilidad de que estos sean plagio o no.
Para esto se desarrollarán y testearán distintos modelos. Algunos con técnicas de Aprendizaje Supervisado, donde se entrenará al modelo a partir de una variable de respuesta conocida, que le informará cuáles casos son plagio y cuáles no. Para esto se podrá optar por utilizar casos reales de copias que se hayan podido detectar anteriormente, así como casos artificiales creados por nosotros mismos para enseñarle al modelo qué es lo que debe observar y detectar.
Por otro lado también se plantearán otros modelos que harán uso de técnicas de Aprendizaje No Supervisado, donde el propio algoritmo agrupará y decidirá qué pares o grupos de texto son posibles plagios, a partir de las especificaciones que nosotros demos a la hora de armar el modelo.

2.5. Herramienta y metodología para el docente

En base de los resultados de los pasos anteriores y teniendo en cuenta lo obtenido en el análisis de necesidad anteriormente realizado, se decidirá la combinación más acorde de distancias y modelos de Aprendizaje Automático.
A partir de esto se desarrollará una herramienta y metodología de trabajo estandarizada para el análisis de grandes cantidades de informes académicos.

Se establecerá la metodología en la que el docente analista deberá cargar los datos y principalmente como deberá interpretar los resultados de la minería de textos. Indicando la forma en la que deberá analizar ciertas situaciones según posibles precisiones o sesgos de los modelos.
Para llevar adelante esta metodología, se desarrollará una herramienta informática con una interfaz acorde al análisis de datos requerido.

3. Cronograma



4. Bibliografía

- **“An Introduction to Statistical Learning: With Applications in R”**
- **“The Elements of Statistical Learning”**
- **“Text Mining Package” - https://cran.r-project.org/web/packages/tm/tm.pdf**
- **“Detect Text Reuse and Document Similarity” - https://cran.r-project.org/web/packages/textreuse/vignettes/textreuse-introduction.html & https://cran.r-project.org/web/packages/textreuse/textreuse.pdf**
- **“3 basic Distance Measurement in Text Mining” - https://towardsdatascience.com/3-basic-distance-measurement-in-text-mining-5852becff1d7**
- **“3 text distances that every data scientist should know” - https://towardsdatascience.com/3-text-distances-that-every-data-scientist-should-know-7fcdf850e510**
- **“Word Distance between Word Embeddings” - https://towardsdatascience.com/word-distance-between-word-embeddings-cc3e9cf1d632#:~:text=Word%20Mover's%20Distance%20(WMD)%20is,introduced%20by%20Kusner%20et%20al**