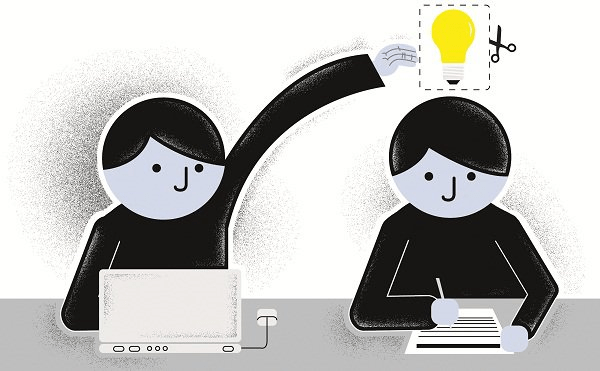
Trabajo Práctico

Alumno: Oviedo Facundo

Materia: Procesamiento del Lenguaje Natural

Prof: Borré Hernán

Tema: Plagio



Contenido

[Funcionamiento 3](#_Toc45904828)

[Desarrollo del algoritmo 3](#_Toc45904829)

[Citas 4](#_Toc45904830)

# Funcionamiento

El algoritmo es capaz de encontrar un porcentaje de oraciones plagiadas con un porcentaje de confianza seteado por el usuario, a partir de un docx o pptx dado, buscando en todos los archivos que se coloquen en una carpeta dada, que también deben estar en formato .docx o .pptx

# Desarrollo del algoritmo

El algoritmo sigue una serie de pasos que se ejecutan en orden

1\_Preparamos el archivo a buscar plagio, extrayendo todo el texto de él, usando la herramienta textract, filtrando en el proceso todos los caracteres de control, tokenizando en oraciones con el tokenizer de nltk y filtrando las oraciones que sean cortas (<30) y los links

2\_Pasado esto, nuestro primer paso es encontrar el nombre del alumno que escribió el docx, se usa el Stanford Post Tagger[1] en español, buscando los nombres propios y filtrando con heurística las palabras que si bien son nombres propios, no son del alumno (cosas como el nombre de la materia o del TP en cuestión)

3\_El siguiente paso es preparar el corpus, es un procedimiento análogo al paso (1) pero con todos los archivos que están en la carpeta que busco el corpus (pudiendo enlistarlas con la library glob), guardando el path donde están ubicados.

4\_ Con una serie de loops for, extraigo oración por oración de mi texto y lo comparo oración por oración con el texto, para hacer este procedimiento se siguen varios pasos

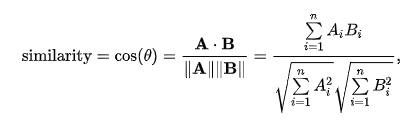
4.1\_Teniendo ya las dos oraciones, las preparo a ambas de la misma manera, primero elimino las stop words en español con la library sklearn

4.2\_Tras ello, tengo que lematizar o sea, llevar las palabras a su palabra raiz, para ello, usé Spacy[2][3] que soporta el idioma español y fue entrenado con Wikipedia

4.3\_Con la comodidad de tener las palabras lematizadas, busco los sinónimos de estas palabras usando LAS-WordNet[4][7], los cuales los pongo en una lista, añadiendo también la palabra original en caso de que no se hayan encontrado sinónimos

4.4\_Convierto la lista de palabras en un set, para eliminar repetidas

4.5\_Una vez pasado a set, las vuelvo a unir en un único string, les aplico la función coseno[5][6] para encontrar similaridad, cuya fórmula es:



5\_El resultado, es un número entre 0 y 1, que lo podemos interpretar como un porcentaje, entonces, a partir del porcentaje de confianza que nos pidió el usuario, podemos definir si en una oración hubo plagio o no

6\_En caso afirmativo, se muestra en pantalla ambas oraciones y usando los índices de los for loop, cuál es el archivo de mi dataset donde se encontró el plagio y la posición de la oración en este documento y se marca la oración examinada como plagiada

7\_Contando la cantidad de oraciones marcadas como plagiadas en relación a la cantidad de oraciones que tiene el documento examinado, se determina el porcentaje de plagio hallado

# Citas

[1] harrison4 (Octubre 2018) Why Python NLTK does not tag correctly in spanish language? <https://stackoverflow.com/questions/52920179/why-python-nltk-does-not-tag-correctly-in-spanish-language> Recuperado 10/07/2020

[2] Benjamín Serra (Marzo 2020) How to solve Spanish lemmatization problems with SpaCy? <https://stackoverflow.com/questions/60534999/how-to-solve-spanish-lemmatization-problems-with-spacy> Recuperado 15/7/2020

[3] José Manuel Martínez Martínez (Agosto 2018) Spanish lemmatization <https://github.com/explosion/spaCy/issues/2710> Recuperado 15/7/2020

[4] Instituto Caro y Cuervo - Grupo de Investigación en Lingüística (Noviembre 2017) LAS-WordNet: una WordNet para el español obtenida con traducción automática <https://dev.socrata.com/foundry/www.datos.gov.co/ghdm-ragt> Recuperado 6/7/2020

[5] Rohan Amrute (Enero 2018) Matching two strings together using NLTK? <https://stackoverflow.com/questions/34847554/matching-two-strings-together-using-nltk> Recuperado 4/7/2020

[6] vpekar (marzo 2013) y Taylor Monacelli (Marzo 2020) Calculate cosine similarity given 2 sentence strings <https://stackoverflow.com/questions/15173225/calculate-cosine-similarity-given-2-sentence-strings> Recuperado 4/7/2020

[7] Stéphanie C (Julio 2015) How to use the language option in synsets (nltk) if you load a wordnet manually? <https://stackoverflow.com/questions/31478152/how-to-use-the-language-option-in-synsets-nltk-if-you-load-a-wordnet-manually> Recuperado 6/7/2020