Procesamiento de Información 2023

Unidad 5 - Tarea

Modelado de espacio vectorial

David Aarón Ramírez Olmeda

Introducción

Para abordar esta tarea, podemos aprovechar gran parte del trabajo realizado en la actividad anterior y realizar algunas modificaciones para incorporar la creación de vectores y el uso de la similitud de coseno.

Detectar el plagio manualmente puede ser una tarea tediosa y poco efectiva, por lo que es importante contar con herramientas automatizadas para realizar esta tarea. El procesamiento del lenguaje natural y los algoritmos de similitud coseno son técnicas que pueden ser útiles para resolver este problema.

Teniendo en cuenta estos conceptos, podemos aplicar lo aprendido en la lección sobre la medida coseno. Esta medida se utiliza comúnmente con vectores que contienen valores reales que definen diferentes ponderaciones de las características del objeto. Principalmente se utiliza para comparar objetos durante los procesos de agrupamiento o clasificación de datos. La fórmula para calcular la medida coseno es la siguiente:

$$Coseno(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} = \frac{\sum_{i=1}^{N} x_i \times y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \sqrt{\sum_{i=1}^{N} y_i^2}}$$

Desarrollo

A grandes rasgos:

- Se carga la muestra de archivos fuente y los archivos sospechosos.
- Se preprocesan los archivos fuente y los archivos sospechosos.
- Se crea una matriz de similitud coseno entre la muestra de archivos fuente y los archivos sospechosos.
- Se encuentran los 3 archivos más parecidos para cada uno de los 20 archivos de la muestra de archivos fuente.

In [1]:
```
import os
import random
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
import numpy as np
from collections import defaultdict
from itertools import combinations
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

In [2]:
```
#nltk.download("stopwords")
stop_words = set(stopwords.words("english"))
stemmer = SnowballStemmer("english")
#nltk.download('punkt')
```

In [3]:
```
def load_files(directory):
    files = []
    for filename in os.listdir(directory):
        if filename.endswith(".txt"):
            filepath = os.path.join(directory, filename)
            with open(filepath, "r", encoding="utf-8") as f:
                text = f.read()
                files.append((filename, text))
    return files
```

In [4]:
```
source_dir = "source-documents"
suspicious_dir = "suspicious-documents"

muestra = 20
source_files = load_files("source-documents")
source_sample = random.sample(source_files, muestra)
suspicious_files = load_files("suspicious-documents")
```

In [5]:
```
print("Los 20 archivos muestra seleccionados serán:")
for filename, text in source_sample:
    print(filename)
```

```
Los 20 archivos muestra seleccionados serán:
source-document0056.txt
source-document0034.txt
source-document0050.txt
source-document0014.txt
source-document0041.txt
source-document0074.txt
source-document0151.txt
source-document0148.txt
source-document0230.txt
source-document0236.txt
source-document0168.txt
source-document0208.txt
source-document0212.txt
source-document0207.txt
source-document0054.txt
source-document0189.txt
source-document0109.txt
source-document0192.txt
source-document0198.txt
source-document0166.txt
```

In [6]:
```
# Preprocesamiento
def preprocess(text):
    text = re.sub(r"[^a-zA-Z\s]", "", text)
    text = text.lower()
    tokens = nltk.word_tokenize(text)
    tokens = [token for token in tokens if token not in stop_words]
    tokens = [stemmer.stem(token) for token in tokens]
    return " ".join(tokens)
```

In [7]:
```
# Creación de vectores TF-IDF
corpus = [preprocess(text) for filename, text in source_sample + suspicio
vectorizer = TfidfVectorizer()
tfidf = vectorizer.fit_transform(corpus)

# Cálculo de similitud coseno
similarity_matrix = cosine_similarity(tfidf[:muestra], tfidf[muestra:])
```

TfidfVectorizer proporciona una forma fácil y eficiente de convertir los textos en vectores TF-IDF y de calcular la similitud coseno entre los vectores.

In [8]:
```
# Encontrar los 3 archivos más parecidos para la muestra (para cada archi
for i, file in enumerate(source_sample):
    similarities = similarity_matrix[i]
    top_similar_indices = np.argsort(similarities)[-3:][::-1]
    print("({}) Archivo de la muestra:".format(i+1), file[0])
    print("Texto original del archivo de muestra:\n", file[1][:100], "...
    for j, index in enumerate(top_similar_indices):
        suspicious_file = suspicious_files[index][0]
        similarity = similarities[index]
        print("- - - Archivo sospechoso #{}:".format(j+1), suspicious_fil
        print("Texto original del archivo sospechoso:\n", suspicious_file
    print()
```

(1) Archivo de la muestra: source-document0056.txt
Texto original del archivo de muestra:
 But the lesson that should not be lost is the transcendent one: Claren
ce Thomas made it in America b ...
- - - Archivo sospechoso #1: suspicious-document0560.txt - Similitud:
0.16717423611014134
Texto original del archivo sospechoso:
 It's time, once again, for the screaming. Evan O'Neal , 4, ambles int
o an examination room at the   ...
- - - Archivo sospechoso #2: suspicious-document2107.txt - Similitud:
0.14989602829196094
Texto original del archivo sospechoso:
  WASHINGTON _ A federal judge Monday found President Clinton in c
ivil contempt of court for lyi ...
- - - Archivo sospechoso #3: suspicious-document0559.txt - Similitud:
0.1469912884752169
Texto original del archivo sospechoso:
 Was TJ Maxx trying to promote a more youthful corporate culture when
a young manager at a New Jer ...

(2) Archivo de la muestra: source-document0034.txt
Texto original del archivo de muestra:
 Firefighters got the upper hand Friday on an 8,200-acre brush fire in
Cleveland National Forest that ...
- - - Archivo sospechoso #1: suspicious-document0339.txt - Similitud:
0.38344071662047235
Texto original del archivo sospechoso:
 Investigators from the Riverside County and California will review par
t of the flight control system ...
- - - Archivo sospechoso #2: suspicious-document0340.txt - Similitud:
0.34689509902664295
Texto original del archivo sospechoso:
 SEATTLE _ Headlines report a $ 400 million effort to save salmon on
the Cleveland National Forest ...
- - - Archivo sospechoso #3: suspicious-document0259.txt - Similitud:
0.2806886522235405
Texto original del archivo sospechoso:
  PARIS _ " Concorde 4590, you have flames, you have flames behind yo
u," the control tower warned th ...

(3) Archivo de la muestra: source-document0050.txt
Texto original del archivo de muestra:
 A controversial videotape being shown among activists nationwide shows
Los Angeles police officers i ...
- - - Archivo sospechoso #1: suspicious-document0470.txt - Similitud:
0.3128111379356221
Texto original del archivo sospechoso:
 BEIJING _ In late 1997 , President Jiang Zemin announced that Chin
a would cut the People's Lib ...
- - - Archivo sospechoso #2: suspicious-document0452.txt - Similitud:
0.2859313843233019
Texto original del archivo sospechoso:
 BEIJING _ In late 1997 , President Jiang Zemin announced that Chin
a would cut the People's Lib ...
- - - Archivo sospechoso #3: suspicious-document0423.txt - Similitud:
0.2850549415663416
Texto original del archivo sospechoso:

BEIJING _ In late 1997 , President Jiang Zemin announced that Chin
a would cut the People's Lib ...

(4) Archivo de la muestra: source-document0014.txt
Texto original del archivo de muestra:
Just last week, a federal judge hearing a civil rights suit issued an
unusual order that employees f ...
- - - Archivo sospechoso #1: suspicious-document2048.txt - Similitud:
0.27797102978242527
Texto original del archivo sospechoso:
BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion o
f having a stolen car. They st ...
- - - Archivo sospechoso #2: suspicious-document2121.txt - Similitud:
0.2762117662926215
Texto original del archivo sospechoso:
BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion o
f having a stolen car. They st ...
- - - Archivo sospechoso #3: suspicious-document2214.txt - Similitud:
0.276123262153711
Texto original del archivo sospechoso:
BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion o
f having a stolen car. They st ...

(5) Archivo de la muestra: source-document0041.txt
Texto original del archivo de muestra:
Language: Spanish Article Type:BFN [Text] The Shining Path does not in
tend to seek reconciliation in ...
- - - Archivo sospechoso #1: suspicious-document0409.txt - Similitud:
0.32601583757380365
Texto original del archivo sospechoso:
MEXICO CITY _ While members of Mexico 's latest generation of stude
nt revolutionaries are known t ...
- - - Archivo sospechoso #2: suspicious-document0410.txt - Similitud:
0.2796319292426042
Texto original del archivo sospechoso:
DETROIT (AP)-- Despite months of labor peace since last summer 's d
evastating strikes, United Au ...
- - - Archivo sospechoso #3: suspicious-document0407.txt - Similitud:
0.20362762783363586
Texto original del archivo sospechoso:
WASHINGTON _ Democratic and Republican lawmakers predicted Tuesday
that the House will reject ...

(6) Archivo de la muestra: source-document0074.txt
Texto original del archivo de muestra:
Cloudy weather Saturday threatened to mar the show for thousands of Fi
nnish and foreign skygazers ho ...
- - - Archivo sospechoso #1: suspicious-document0740.txt - Similitud:
0.24501858580307975
Texto original del archivo sospechoso:
DETROIT _ Crippling strikes at General Motors parts factories have l
eft the United Automobile Wor ...
- - - Archivo sospechoso #2: suspicious-document0739.txt - Similitud:
0.22311605811245816
Texto original del archivo sospechoso:
SAN FRANCISCO _ The Force will soon be arriving at a toy store and
Taco Bell near you. In May , ...

- - - Archivo sospechoso #3: suspicious-document0710.txt - Similitud: 0.2101620562425283
Texto original del archivo sospechoso:
MIAMI _ As the city's streets settled into an uneasy calm, the battle over six-year-old Cuban rafter ...

(7) Archivo de la muestra: source-document0151.txt
Texto original del archivo de muestra:
U.S. military helicopters on Sunday located the wreckage of a plane th at crashed last Monday with Te ...
- - - Archivo sospechoso #1: suspicious-document1510.txt - Similitud: 0.31658843688910177
Texto original del archivo sospechoso:
TOPPENISH , Wash. (AP)-- At Tiny's Tavern , bartender Nena Garc ia wonders how much longer sh ...
- - - Archivo sospechoso #2: suspicious-document1509.txt - Similitud: 0.22148464280466748
Texto original del archivo sospechoso:
Photo and graphics information and editors' names can be found at the end of this budget. INTERNATIO ...
- - - Archivo sospechoso #3: suspicious-document0169.txt - Similitud: 0.20293861599681137
Texto original del archivo sospechoso:
Investigators from the Atlantic and Hatteras will review part of the f light control system in the ta ...

(8) Archivo de la muestra: source-document0148.txt
Texto original del archivo de muestra:
"Mad cow disease" has killed 10,000 cattle, restricted the export mark et for Britain's cattle indust ...
- - - Archivo sospechoso #1: suspicious-document0698.txt - Similitud: 0.2344169565817479
Texto original del archivo sospechoso:
HENDERSONVILLE , N.C. _ The best defense against West Nile virus, the Third World pathogen th ...
- - - Archivo sospechoso #2: suspicious-document0700.txt - Similitud: 0.20221335438234544
Texto original del archivo sospechoso:
QUEBEC CITY _ Five years ago, Canada seemed a country about to rip i tself apart. In a 1995 refer ...
- - - Archivo sospechoso #3: suspicious-document0855.txt - Similitud: 0.20136524338860817
Texto original del archivo sospechoso:
HENDERSONVILLE , N.C. _ The best defense against West Nile virus, the Third World pathogen th ...

(9) Archivo de la muestra: source-document0230.txt
Texto original del archivo de muestra:
India's military and paramilitary forces were put on "red alert" as ga ngs took to the streets of New ...
- - - Archivo sospechoso #1: suspicious-document2300.txt - Similitud: 0.3695530734638179
Texto original del archivo sospechoso:
WASHINGTON _ A federal judge Monday found President Clinton in c ivil contempt of court for lyi ...
- - - Archivo sospechoso #2: suspicious-document2299.txt - Similitud: 0.33842858193232905

Texto original del archivo sospechoso:
 SANTA FE, N.M. _ Bucking recent changes in Kansas and other states that allow public schools to ...
- - - Archivo sospechoso #3: suspicious-document2292.txt - Similitud: 0.3196936476019484
Texto original del archivo sospechoso:
 Unlike past death penalty cases, where Rajiv Gandhi waited until all appeals has been exhausted befo ...

(10) Archivo de la muestra: source-document0236.txt
Texto original del archivo de muestra:
 The accidental shooting death of a young stockbroker by an officer looking for a burglar is one more ...
- - - Archivo sospechoso #1: suspicious-document2048.txt - Similitud: 0.25328234655504117
Texto original del archivo sospechoso:
  BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion of having a stolen car. They st ...
- - - Archivo sospechoso #2: suspicious-document2214.txt - Similitud: 0.25120554926784433
Texto original del archivo sospechoso:
  BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion of having a stolen car. They st ...
- - - Archivo sospechoso #3: suspicious-document2157.txt - Similitud: 0.2511982655694678
Texto original del archivo sospechoso:
  BIRMINGHAM , England _ Police stopped Carl Josephs on suspicion of having a stolen car. They st ...

(11) Archivo de la muestra: source-document0168.txt
Texto original del archivo de muestra:
 Sir, Mr James Skinner (Letters, April 11) charges me with misuse of statistics and understanding the ...
- - - Archivo sospechoso #1: suspicious-document1679.txt - Similitud: 0.3434934749604141
Texto original del archivo sospechoso:
  WASHINGTON (AP)-- House of Lords Chairman Lord Bauer emphatically cautioned lawmakers today agai ...
- - - Archivo sospechoso #2: suspicious-document1650.txt - Similitud: 0.18542283028463424
Texto original del archivo sospechoso:
 Slovenia (AP)-- European Community Chairman Alan Greenspan emphatically cautioned lawmakers toda ...
- - - Archivo sospechoso #3: suspicious-document1366.txt - Similitud: 0.1827050162207383
Texto original del archivo sospechoso:
 Slovenia (AP)-- World Bank Chairman Alan Greenspan emphatically cautioned lawmakers today agains ...

(12) Archivo de la muestra: source-document0208.txt
Texto original del archivo de muestra:
 With the winds of Hurricane Gilbert clocked at 175 miles per hour, U.S. weather officials called Gil ...
- - - Archivo sospechoso #1: suspicious-document2079.txt - Similitud: 0.3213985456531076
Texto original del archivo sospechoso:
 At this stage of research, scientists can not firmly link global warmi

ng to changes in circulation p ...
- - - Archivo sospechoso #2: suspicious-document2059.txt - Similitud: 0.277488265293066
Texto original del archivo sospechoso:
  MOSCOW _ With a hefty lead in the polls, Vladimir Putin has perfec ted his campaign strategy _ ac ...
- - - Archivo sospechoso #3: suspicious-document2080.txt - Similitud: 0.26608055274410713
Texto original del archivo sospechoso:
 Yucatan Penninsula _ French investigators looking into the crash 1935 of an Air France Concorde sa ...

(13) Archivo de la muestra: source-document0212.txt
Texto original del archivo de muestra:
 As of last month, live pines in the Mount Palomar area near San Diego had less moisture than boards ...
- - - Archivo sospechoso #1: suspicious-document0259.txt - Similitud: 0.2462934525405549
Texto original del archivo sospechoso:
  PARIS _ " Concorde 4590, you have flames, you have flames behind yo u," the control tower warned th ...
- - - Archivo sospechoso #2: suspicious-document1389.txt - Similitud: 0.21915363032135993
Texto original del archivo sospechoso:
  LONDON (AP)-- Passengers describe taking off on the Concorde as " a kick in the back" and compla ...
- - - Archivo sospechoso #3: suspicious-document2269.txt - Similitud: 0.21181396128535718
Texto original del archivo sospechoso:
 Yellowstone National Park _ French investigators looking into the cras h last September of an Secreta ...

(14) Archivo de la muestra: source-document0207.txt
Texto original del archivo de muestra:
 Language: Serbo-Croatian Article Type:BFN [Article by Verica Rudar: "W hat Is Ljubljana's Message for ...
- - - Archivo sospechoso #1: suspicious-document1629.txt - Similitud: 0.3250532153626831
Texto original del archivo sospechoso:
 The Slovenia relatives of Elian Gonzales Friday lost what may be t heir final battle to prevent t ...
- - - Archivo sospechoso #2: suspicious-document1774.txt - Similitud: 0.29695716804042027
Texto original del archivo sospechoso:
  NAIROBI , Kenya _ Tanzania charged two men on Monday with 11 co unts of murder in connection w ...
- - - Archivo sospechoso #3: suspicious-document2070.txt - Similitud: 0.29361117713658613
Texto original del archivo sospechoso:
  CARACAS , Venezuela (AP) _ Former Lt. Col. Hugo Chavez , who sta ged a bloody coup attempt six ...

(15) Archivo de la muestra: source-document0054.txt
Texto original del archivo de muestra:
 Thomas and his brother made it in the white world. Their sister, reare d by an aunt, had four childre ...
- - - Archivo sospechoso #1: suspicious-document1040.txt - Similitud:

0.22743208454537261
Texto original del archivo sospechoso:
  WASHINGTON _ The centerpiece of the Clinton administration's econo
mic policy this year has been ...
- - - Archivo sospechoso #2: suspicious-document1692.txt - Similitud:
0.22006080806064915
Texto original del archivo sospechoso:
  WASHINGTON _ A coalition of influential Christian groups has launche
d a national ad campaign attac ...
- - - Archivo sospechoso #3: suspicious-document0539.txt - Similitud:
0.20507899903869672
Texto original del archivo sospechoso:
 Was TJ Maxx trying to promote a more youthful corporate culture when
a young manager at a New Jer ...

(16) Archivo de la muestra: source-document0189.txt
Texto original del archivo de muestra:
 Firefighters in California, Michigan, Montana, Wyoming and Utah battle
d holiday weekend fires which ...
- - - Archivo sospechoso #1: suspicious-document0420.txt - Similitud:
0.32657996539951123
Texto original del archivo sospechoso:
 TROY, Utah (AP)-- Zion National Park, the auto-parts manufacturer so
on to be independent from Cust ...
- - - Archivo sospechoso #2: suspicious-document1890.txt - Similitud:
0.3070271320266636
Texto original del archivo sospechoso:
  BOSTON _ In the midst of an economic and real estate boom, many low-
and moderate-income residents ...
- - - Archivo sospechoso #3: suspicious-document0259.txt - Similitud:
0.30183336716184084
Texto original del archivo sospechoso:
  PARIS _ " Concorde 4590, you have flames, you have flames behind yo
u," the control tower warned th ...

(17) Archivo de la muestra: source-document0109.txt
Texto original del archivo de muestra:
 Central American and Caribbean governments are awaiting with more than
passing interest an imminent ...
- - - Archivo sospechoso #1: suspicious-document1090.txt - Similitud:
0.2591411725780666
Texto original del archivo sospechoso:
  WASHINGTON _ Chemists who examined soil, sludge and debris samples f
rom a Sudanese pharmaceutical ...
- - - Archivo sospechoso #2: suspicious-document1089.txt - Similitud:
0.215776172903507
Texto original del archivo sospechoso:
 SAN FRANCISCO _ The Force will soon be arriving at a toy store and
Taco Bell near you. In May , ...
- - - Archivo sospechoso #3: suspicious-document1087.txt - Similitud:
0.1956724857820051
Texto original del archivo sospechoso:
 NEW YORK (AP)-- NAFTA is with " Star Wars" fans, and it's telling th
em to hit the toy stores early ...

(18) Archivo de la muestra: source-document0192.txt
Texto original del archivo de muestra:

Two years ago, it looked as if a vast part of the nation's farm empire was burning up as drought and ...
- - - Archivo sospechoso #1: suspicious-document0132.txt - Similitud: 0.33146007889340345
Texto original del archivo sospechoso:
WASHINGTON , Va. _ Sunrise at Sunnyside Farms is greeted by a rit ual that startles the uninitia ...
- - - Archivo sospechoso #2: suspicious-document0051.txt - Similitud: 0.3287909163680053
Texto original del archivo sospechoso:
WASHINGTON , Va. _ Sunrise at Sunnyside Farms is greeted by a rit ual that startles the uninitia ...
- - - Archivo sospechoso #3: suspicious-document0073.txt - Similitud: 0.3284107623746158
Texto original del archivo sospechoso:
WASHINGTON , Va. _ Sunrise at Sunnyside Farms is greeted by a rit ual that startles the uninitia ...

(19) Archivo de la muestra: source-document0198.txt
Texto original del archivo de muestra:
A tornado blasted through this North Florida town before dawn today, d estroying several homes and a ...
- - - Archivo sospechoso #1: suspicious-document1980.txt - Similitud: 0.2470325200126677
Texto original del archivo sospechoso:
WASHINGTON (AP)-- With federal regulators pondering whether they s hould start rating child safet ...
- - - Archivo sospechoso #2: suspicious-document1979.txt - Similitud: 0.23034318231875256
Texto original del archivo sospechoso:
NEW YORK (AP)-- Proving even more virulent than first believed, the computer virus Worm. Explore. ...
- - - Archivo sospechoso #3: suspicious-document0487.txt - Similitud: 0.1853524703180542
Texto original del archivo sospechoso:
SPARTA, Ga. _ By 10 o'clock on a typical summer 's morning, Hancoc k County is almost deserted. ...

(20) Archivo de la muestra: source-document0166.txt
Texto original del archivo de muestra:
Solar telescopes yielded views of flare-producing sunspots and silhoue tted mountains on the moon Tue ...
- - - Archivo sospechoso #1: suspicious-document1660.txt - Similitud: 0.23646177997193574
Texto original del archivo sospechoso:
BOSTON (AP) _ Suddenly hundreds of thousands, maybe even millions, o f Americans with arthritis are ...
- - - Archivo sospechoso #2: suspicious-document0710.txt - Similitud: 0.20143726887868776
Texto original del archivo sospechoso:
MIAMI _ As the city's streets settled into an uneasy calm, the battle over six-year-old Cuban rafter ...
- - - Archivo sospechoso #3: suspicious-document1330.txt - Similitud: 0.20125812934735648
Texto original del archivo sospechoso:
Earth (AP)-- Passengers describe taking off on the Concorde as " a k

```
ick in the back" and complain  ...
```

Conclusión

Hemos usado la técnica de similitud coseno para detectar plagio en textos. Al utilizar esta medida y la creación de vectores TF-IDF, se pueden encontrar similitudes entre diferentes documentos y determinar si existe plagio o no.

In [ ]: