4A. Práctica: Detección de Plagio

Nombre: **Luis Fernando Izquierdo Berdugo**

Materia: **Procesamiento de Información**

Fecha: **28 de Octubre de 2024**

Instrucciones:

En esta actividad, el ejercicio propuesto es: **detectar a nivel básico, por medio de una medida de similitud, el plagio que hay dentro de un conjunto de documentos**.

Los datos son documentos de noticias, adaptados a la tarea de detección de plagio. En el conjunto de datos del directorio `suspicious-documents`, hay documentos que tienen fragmentos plagiados de los documentos fuente `source-documents`, sin embargo, también hay otros documentos que simularon plagio, pero no se plagió ningún párrafo de los documentos fuente.

La tarea es encontrar el nivel de plagio entre los archivos del directorio `source-documents` y los archivos del directorio `suspicious-documents`.

Entregar un notebook con el código fuente para obtener la similitud entre dos documentos. Tomar 20 documentos de source-documents y por cada uno, encontrar los 5 archivos de suspicious-documents más parecidos. Indicar los nombres y la similitud obtenida.

Aplicar los preprocesamientos: usar solo letras minúsculas, remover stopwords (palabras que no son de contenido) y aplicar un proceso de stemming para reducir el vocabulario.

Nota: Usar la lista de stopwords para eliminarlas, recordar si se le aplicó stemming, aplicar stemming también a las stopwords, debido a que alguna palabra podría cambiar.

1. Se recomienda usar el stemmer de NLTK. Para facilidad, usar Porter (inglés).

http://www.nltk.org/howto/stem.html

Básicamente usar:

from nltk.stem.porter import *

stemmer = PorterStemmer()

word = stemmer.stem('pages')

Word contendrá la palabra normalizada por medio de stemming ('page') aplicar ese proceso a todos los textos.

2. Para procesar la medida de similitud, usar la medida Jaccard y Dice

Preprocesamiento

Lo primero que se hará será definitr la lista de stopwords y aplicarles stemming. Para esto se usará una instancia del Porter Stemmer de la biblioteca `nltk`.

In [15]:
```
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
# Descargar stopwords si no están descargadas
#nltk.download('stopwords')

# Crear una instancia del PorterStemmer
stemmer = PorterStemmer()

# Definir la lista de stopwords y aplicar stemming
stop_words = set(stopwords.words('english'))
stemmed_stop_words = {stemmer.stem(word) for word in stop_words}
```

Se definirá la función para leer los documentos. En esta se ordenan por nombre alfabético (para poder tomar los primeros 20 documentos) y se creará un diccionario con el nombre del archivo y el contenido de estos.

In [16]:
```
import os
def read_documents(directory, limit=None):
    documents = {}
    filenames = sorted([f for f in os.listdir(directory) if f.endswith(".txt")]
    for i, filename in enumerate(filenames):
        if limit is not None and i >= limit:
            break
        with open(os.path.join(directory, filename), 'r', encoding='utf-8') as
            documents[filename] = file.read()
    return documents
```

Lo siguiente será definir la función para el preprocesamiento del texto. En esta función se hará lo siguiente:

- Se convertirá a minúsculas el texto
- Se eliminarán caracteres especiales
- Se dividirá el texto en palabras
- Se aplicará el Porter stemming.
- Se eliminarán las stopwords

In [17]:
```
import re
def preprocess_text(text):
    """Preprocesa el texto: minúsculas, eliminar stopwords y aplicar stemming."
    # Convertir a minúsculas
    text = text.lower()
    # Eliminar caracteres especiales
    text = re.sub(r'[^a-z\s]', '', text)
    # Dividir en palabras
    words = text.split()
    # Aplicar stemming y eliminar stopwords
    stemmed_words = [stemmer.stem(word) for word in words if word not in stemme
    return ' '.join(stemmed_words)
```

Ya con las funciones se abren leen los archivos de las rutas y se preprocesa cada uno de ellos.

In [18]:
```
source_dir = "/Users/izluis/Documents/source-documents"
suspicious_dir = "/Users/izluis/Documents/suspicious-documents"

# Leer y preprocesar los documentos
source_documents = read_documents(source_dir, limit=20)
suspicious_documents = read_documents(suspicious_dir)

preprocessed_source_docs = {name: preprocess_text(text) for name, text in sourc
preprocessed_suspicious_docs = {name: preprocess_text(text) for name, text in s
```

Similitud de Jaccard

Se implementa con una función que efectúa lo siguiente:

- Divide los textos por palabras usando los espacios como delimitador
- Se convierte las listas de palabras en conjuntos para eliminar duplicados (así como permite que se hagan operaciones)
- Calcula la interseccion de los dos conjuntos (las palabras que están presentes en ambos textos)
- Calcula la unión de los dos conjuntos de palabras (todas las palabras que están mínimo en uno de los documentos)
- Se calcula la similitud de Jaccard dividiendo el tamaño de la intersección entre el tamaño de la unión.

In [19]:
```
def jaccard_similarity(doc1, doc2):
    words_doc1 = set(doc1.split())
    words_doc2 = set(doc2.split())
    intersection = words_doc1.intersection(words_doc2)
    union = words_doc1.union(words_doc2)
    return len(intersection) / len(union)
```

Similitud de Dice

La función para calcular la similitud de Dice efectúa lo siguiente:

- Divide los textos por palabras usando los espacios como delimitador.
- Se converten las listas de palabras en conjuntos para eliminar duplicados (así como permite que se hagan operaciones)
- Se calcula la intersección de los dos conjuntos (las palabras que están presentes en ambos textos)
- Se calcula la similitud de Dice con la fórmula, esta dice que es el doble del tamaño de la intersección dividido entre la suma de los tamaños de los dos conjuntos

In [20]:
```
def dice_similarity(doc1, doc2):
    words_doc1 = set(doc1.split())
    words_doc2 = set(doc2.split())
```

```
    intersection = words_doc1.intersection(words_doc2)
    return 2 * len(intersection) / (len(words_doc1) + len(words_doc2))
```

Cálculo de similutdes

Lo último será iterar por todos los `source documents` y encontrar los documentos más similares de `suspicious documents`. Primero se muestran los más similares por Jaccard y después los más similares por Dice.

In [23]:
```
# Calcular similitudes y encontrar los documentos más similares
results_jaccard = {}
results_dice = {}

for source_name, source_text in preprocessed_source_docs.items():
    similarities = []
    for suspicious_name, suspicious_text in preprocessed_suspicious_docs.items(
        jaccard_sim = jaccard_similarity(source_text, suspicious_text)
        dice_sim = dice_similarity(source_text, suspicious_text)
        similarities.append((suspicious_name, jaccard_sim, dice_sim))

    # Ordenar por similitud de Jaccard y tomar los 5 más similares
    similarities.sort(key=lambda x: x[1], reverse=True)
    results_jaccard[source_name] = similarities[:5]

    # Ordenar por similitud de Dice y tomar los 5 más similares
    similarities.sort(key=lambda x: x[2], reverse=True)
    results_dice[source_name] = similarities[:5]

# Mostrar resultados
print("Resultados por Jaccard:")
for source_name, similar_docs in results_jaccard.items():
    print(f"Documentos más similares a {source_name}:")
    for doc_name, jaccard_sim, dice_sim in similar_docs:
        print(f"  {doc_name}: Jaccard={jaccard_sim:.4f}")
    print()

print("-----------------------------------------------")
print("Resultados por Dice:")
for source_name, similar_docs in results_dice.items():
    print(f"Documentos más similares a {source_name}:")
    for doc_name, jaccard_sim, dice_sim in similar_docs:
        print(f"  {doc_name}: Dice={dice_sim:.4f}")
    print()
```

```
Resultados por Jaccard:
Documentos más similares a source-document0001.txt:
  suspicious-document0010.txt: Jaccard=0.1820
  suspicious-document2205.txt: Jaccard=0.1774
  suspicious-document2048.txt: Jaccard=0.1762
  suspicious-document2121.txt: Jaccard=0.1756
  suspicious-document1267.txt: Jaccard=0.1736

Documentos más similares a source-document0002.txt:
  suspicious-document0020.txt: Jaccard=0.1449
  suspicious-document0019.txt: Jaccard=0.1278
  suspicious-document1463.txt: Jaccard=0.1164
  suspicious-document0013.txt: Jaccard=0.1143
  suspicious-document1603.txt: Jaccard=0.1132

Documentos más similares a source-document0003.txt:
  suspicious-document0030.txt: Jaccard=0.2004
  suspicious-document0029.txt: Jaccard=0.1734
  suspicious-document1000.txt: Jaccard=0.1123
  suspicious-document0023.txt: Jaccard=0.1098
  suspicious-document0022.txt: Jaccard=0.1080

Documentos más similares a source-document0004.txt:
  suspicious-document0040.txt: Jaccard=0.1782
  suspicious-document0039.txt: Jaccard=0.1738
  suspicious-document0031.txt: Jaccard=0.1658
  suspicious-document0033.txt: Jaccard=0.1648
  suspicious-document0151.txt: Jaccard=0.1470

Documentos más similares a source-document0005.txt:
  suspicious-document0049.txt: Jaccard=0.1986
  suspicious-document0044.txt: Jaccard=0.1335
  suspicious-document2070.txt: Jaccard=0.1322
  suspicious-document0050.txt: Jaccard=0.1307
  suspicious-document1779.txt: Jaccard=0.1304

Documentos más similares a source-document0006.txt:
  suspicious-document0059.txt: Jaccard=0.1664
  suspicious-document0060.txt: Jaccard=0.1643
  suspicious-document0055.txt: Jaccard=0.1473
  suspicious-document0839.txt: Jaccard=0.1391
  suspicious-document0720.txt: Jaccard=0.1346

Documentos más similares a source-document0007.txt:
  suspicious-document0069.txt: Jaccard=0.1956
  suspicious-document0061.txt: Jaccard=0.1282
  suspicious-document2289.txt: Jaccard=0.1253
  suspicious-document0067.txt: Jaccard=0.1220
  suspicious-document0063.txt: Jaccard=0.1197

Documentos más similares a source-document0008.txt:
  suspicious-document0080.txt: Jaccard=0.1722
  suspicious-document0079.txt: Jaccard=0.1507
  suspicious-document0792.txt: Jaccard=0.1411
  suspicious-document1484.txt: Jaccard=0.1404
  suspicious-document0183.txt: Jaccard=0.1380

Documentos más similares a source-document0009.txt:
  suspicious-document0090.txt: Jaccard=0.1735
  suspicious-document0390.txt: Jaccard=0.1670
```

```
  suspicious-document0384.txt: Jaccard=0.1551
  suspicious-document1239.txt: Jaccard=0.1457
  suspicious-document1240.txt: Jaccard=0.1402

Documentos más similares a source-document0010.txt:
  suspicious-document0099.txt: Jaccard=0.1645
  suspicious-document0100.txt: Jaccard=0.1422
  suspicious-document2040.txt: Jaccard=0.1362
  suspicious-document0096.txt: Jaccard=0.1311
  suspicious-document1048.txt: Jaccard=0.1257

Documentos más similares a source-document0011.txt:
  suspicious-document0110.txt: Jaccard=0.2188
  suspicious-document0108.txt: Jaccard=0.1671
  suspicious-document2309.txt: Jaccard=0.1670
  suspicious-document0249.txt: Jaccard=0.1609
  suspicious-document0038.txt: Jaccard=0.1587

Documentos más similares a source-document0012.txt:
  suspicious-document0119.txt: Jaccard=0.2045
  suspicious-document0112.txt: Jaccard=0.1543
  suspicious-document0120.txt: Jaccard=0.1385
  suspicious-document0116.txt: Jaccard=0.1335
  suspicious-document0113.txt: Jaccard=0.1322

Documentos más similares a source-document0013.txt:
  suspicious-document0130.txt: Jaccard=0.1947
  suspicious-document0129.txt: Jaccard=0.1744
  suspicious-document0123.txt: Jaccard=0.1575
  suspicious-document0127.txt: Jaccard=0.1561
  suspicious-document0125.txt: Jaccard=0.1432

Documentos más similares a source-document0014.txt:
  suspicious-document0139.txt: Jaccard=0.1947
  suspicious-document0140.txt: Jaccard=0.1826
  suspicious-document2018.txt: Jaccard=0.1689
  suspicious-document1896.txt: Jaccard=0.1680
  suspicious-document1885.txt: Jaccard=0.1667

Documentos más similares a source-document0015.txt:
  suspicious-document0150.txt: Jaccard=0.1902
  suspicious-document0149.txt: Jaccard=0.1786
  suspicious-document0141.txt: Jaccard=0.1482
  suspicious-document0143.txt: Jaccard=0.1347
  suspicious-document0499.txt: Jaccard=0.1326

Documentos más similares a source-document0016.txt:
  suspicious-document1820.txt: Jaccard=0.1965
  suspicious-document0177.txt: Jaccard=0.1940
  suspicious-document0183.txt: Jaccard=0.1939
  suspicious-document0245.txt: Jaccard=0.1931
  suspicious-document0194.txt: Jaccard=0.1926

Documentos más similares a source-document0017.txt:
  suspicious-document0169.txt: Jaccard=0.2521
  suspicious-document0170.txt: Jaccard=0.1651
  suspicious-document0161.txt: Jaccard=0.1244
  suspicious-document0162.txt: Jaccard=0.1081
  suspicious-document0163.txt: Jaccard=0.1051
```

```
Documentos más similares a source-document0018.txt:
  suspicious-document0179.txt: Jaccard=0.2144
  suspicious-document0172.txt: Jaccard=0.1569
  suspicious-document0180.txt: Jaccard=0.1436
  suspicious-document1931.txt: Jaccard=0.1382
  suspicious-document1882.txt: Jaccard=0.1375

Documentos más similares a source-document0019.txt:
  suspicious-document0189.txt: Jaccard=0.1937
  suspicious-document0182.txt: Jaccard=0.1402
  suspicious-document1066.txt: Jaccard=0.1364
  suspicious-document1678.txt: Jaccard=0.1339
  suspicious-document1086.txt: Jaccard=0.1327

Documentos más similares a source-document0020.txt:
  suspicious-document0199.txt: Jaccard=0.2244
  suspicious-document0200.txt: Jaccard=0.1529
  suspicious-document1734.txt: Jaccard=0.1527
  suspicious-document0219.txt: Jaccard=0.1522
  suspicious-document0477.txt: Jaccard=0.1491

--------------------------------------------
Resultados por Dice:
Documentos más similares a source-document0001.txt:
  suspicious-document0010.txt: Dice=0.3079
  suspicious-document2205.txt: Dice=0.3014
  suspicious-document2048.txt: Dice=0.2997
  suspicious-document2121.txt: Dice=0.2987
  suspicious-document1267.txt: Dice=0.2958

Documentos más similares a source-document0002.txt:
  suspicious-document0020.txt: Dice=0.2531
  suspicious-document0019.txt: Dice=0.2267
  suspicious-document1463.txt: Dice=0.2085
  suspicious-document0013.txt: Dice=0.2052
  suspicious-document1603.txt: Dice=0.2033

Documentos más similares a source-document0003.txt:
  suspicious-document0030.txt: Dice=0.3339
  suspicious-document0029.txt: Dice=0.2955
  suspicious-document1000.txt: Dice=0.2019
  suspicious-document0023.txt: Dice=0.1979
  suspicious-document0022.txt: Dice=0.1949

Documentos más similares a source-document0004.txt:
  suspicious-document0040.txt: Dice=0.3025
  suspicious-document0039.txt: Dice=0.2961
  suspicious-document0031.txt: Dice=0.2844
  suspicious-document0033.txt: Dice=0.2830
  suspicious-document0151.txt: Dice=0.2563

Documentos más similares a source-document0005.txt:
  suspicious-document0049.txt: Dice=0.3314
  suspicious-document0044.txt: Dice=0.2355
  suspicious-document2070.txt: Dice=0.2335
  suspicious-document0050.txt: Dice=0.2312
  suspicious-document1779.txt: Dice=0.2308

Documentos más similares a source-document0006.txt:
  suspicious-document0059.txt: Dice=0.2853
```

```
  suspicious-document0060.txt: Dice=0.2822
  suspicious-document0055.txt: Dice=0.2568
  suspicious-document0839.txt: Dice=0.2442
  suspicious-document0720.txt: Dice=0.2373

Documentos más similares a source-document0007.txt:
  suspicious-document0069.txt: Dice=0.3272
  suspicious-document0061.txt: Dice=0.2273
  suspicious-document2289.txt: Dice=0.2227
  suspicious-document0067.txt: Dice=0.2174
  suspicious-document0063.txt: Dice=0.2138

Documentos más similares a source-document0008.txt:
  suspicious-document0080.txt: Dice=0.2938
  suspicious-document0079.txt: Dice=0.2619
  suspicious-document0792.txt: Dice=0.2473
  suspicious-document1484.txt: Dice=0.2462
  suspicious-document0183.txt: Dice=0.2426

Documentos más similares a source-document0009.txt:
  suspicious-document0090.txt: Dice=0.2957
  suspicious-document0390.txt: Dice=0.2862
  suspicious-document0384.txt: Dice=0.2685
  suspicious-document1239.txt: Dice=0.2544
  suspicious-document1240.txt: Dice=0.2459

Documentos más similares a source-document0010.txt:
  suspicious-document0099.txt: Dice=0.2825
  suspicious-document0100.txt: Dice=0.2489
  suspicious-document2040.txt: Dice=0.2397
  suspicious-document0096.txt: Dice=0.2319
  suspicious-document1048.txt: Dice=0.2233

Documentos más similares a source-document0011.txt:
  suspicious-document0110.txt: Dice=0.3590
  suspicious-document0108.txt: Dice=0.2863
  suspicious-document2309.txt: Dice=0.2862
  suspicious-document0249.txt: Dice=0.2772
  suspicious-document0038.txt: Dice=0.2740

Documentos más similares a source-document0012.txt:
  suspicious-document0119.txt: Dice=0.3396
  suspicious-document0112.txt: Dice=0.2674
  suspicious-document0120.txt: Dice=0.2432
  suspicious-document0116.txt: Dice=0.2356
  suspicious-document0113.txt: Dice=0.2335

Documentos más similares a source-document0013.txt:
  suspicious-document0130.txt: Dice=0.3259
  suspicious-document0129.txt: Dice=0.2969
  suspicious-document0123.txt: Dice=0.2721
  suspicious-document0127.txt: Dice=0.2701
  suspicious-document0125.txt: Dice=0.2505

Documentos más similares a source-document0014.txt:
  suspicious-document0139.txt: Dice=0.3259
  suspicious-document0140.txt: Dice=0.3088
  suspicious-document2018.txt: Dice=0.2890
  suspicious-document1896.txt: Dice=0.2877
  suspicious-document1885.txt: Dice=0.2857
```

Documentos más similares a source-document0015.txt:
  suspicious-document0150.txt: Dice=0.3196
  suspicious-document0149.txt: Dice=0.3031
  suspicious-document0141.txt: Dice=0.2582
  suspicious-document0143.txt: Dice=0.2374
  suspicious-document0499.txt: Dice=0.2341

Documentos más similares a source-document0016.txt:
  suspicious-document1820.txt: Dice=0.3284
  suspicious-document0177.txt: Dice=0.3249
  suspicious-document0183.txt: Dice=0.3248
  suspicious-document0245.txt: Dice=0.3237
  suspicious-document0194.txt: Dice=0.3230

Documentos más similares a source-document0017.txt:
  suspicious-document0169.txt: Dice=0.4027
  suspicious-document0170.txt: Dice=0.2834
  suspicious-document0161.txt: Dice=0.2213
  suspicious-document0162.txt: Dice=0.1951
  suspicious-document0163.txt: Dice=0.1902

Documentos más similares a source-document0018.txt:
  suspicious-document0179.txt: Dice=0.3532
  suspicious-document0172.txt: Dice=0.2712
  suspicious-document0180.txt: Dice=0.2511
  suspicious-document1931.txt: Dice=0.2428
  suspicious-document1882.txt: Dice=0.2417

Documentos más similares a source-document0019.txt:
  suspicious-document0189.txt: Dice=0.3246
  suspicious-document0182.txt: Dice=0.2459
  suspicious-document1066.txt: Dice=0.2400
  suspicious-document1678.txt: Dice=0.2363
  suspicious-document1086.txt: Dice=0.2342

Documentos más similares a source-document0020.txt:
  suspicious-document0199.txt: Dice=0.3666
  suspicious-document0200.txt: Dice=0.2653
  suspicious-document1734.txt: Dice=0.2649
  suspicious-document0219.txt: Dice=0.2642
  suspicious-document0477.txt: Dice=0.2595