Nombre: Ivan Antony Luque Suca

LABORATORIO: Manipulación de Word embbeding

En este laboratorio, aplicará operaciones de álgebra lineal usando NumPy para encontrar relaciones entre palabras para lo cual se tiene como objetivo:

- Analizar el código
- Escribir conclusiones del Laboratorio de Manipulación de word embbeding

1. Manipulación de word embeddings

Los **word embeddings** son representaciones de palabras con vectores.

```
import pandas as pd
import numpy as np
import pickle

word_embeddings = pickle.load( open( "word_embeddings_subset.p", "rb" ) )
len(word_embeddings) #  243 words
```

```
243
```

```
word_embeddings
```

```
 -0.03027297,  0.09082031, -0.20751812,  0.08886719,  0.05322266,
 -0.22265625, -0.09667969, -0.19726562, -0.09570312,  0.05786133,
  0.109375  , -0.16113281,  0.06738281,  0.07421875,  0.21289062,
  0.0168457 ,  0.06689453,  0.18066406,  0.13378906, -0.27148438,
  0.00805664, -0.13867188, -0.03613281,  0.11181641, -0.18066406,
  0.03076172, -0.13085938,  0.10058594,  0.25976562,  0.20605469,
 -0.03271484, -0.05932617,  0.10009766, -0.171875  ,  0.17578125,
  0.18554688, -0.19042969, -0.01721191,  0.23144531,  0.18847656,
 -0.33398438, -0.01599121, -0.37890625,  0.03637695,  0.05786133,
 -0.0612793 ,  0.07226562,  0.03344727, -0.06079102, -0.078125  ,
 -0.08447266,  0.19824219, -0.15917969, -0.16992188,  0.00305176,
 -0.06298828,  0.25      , -0.0456543 ,  0.16503906, -0.12207031,
  0.02050781,  0.01916504, -0.06494141, -0.24023438,  0.23730469,
 -0.00921631,  0.04516602, -0.19042969,  0.17285156,  0.01757812,
 -0.15234375, -0.12158203, -0.14746094,  0.15917969,  0.12402344,
  0.0859375 , -0.03930664,  0.02246094,  0.04492188, -0.14257812,
  0.41601562, -0.16113281,  0.26953125,  0.16015625, -0.44726562,
  0.13671875, -0.00148773, -0.13964844, -0.14453125, -0.07128906,
  0.11669922,  0.06030273, -0.1640625 , -0.05444336,  0.07666016,
 -0.12890625, -0.01696777, -0.08642578,  0.0300293 , -0.00320435,
  0.21484375,  0.0111084 ,  0.29101562, -0.02294922,  0.32421875,
  0.16308594,  0.01660156,  0.13183594, -0.08056641, -0.2421875 ,
 -0.04174805, -0.01062012,  0.12792969,  0.03515625, -0.03063965,
  0.12109375, -0.12304688,  0.22363281, -0.0625    ,  0.40429688,
  0.36328125,  0.10839844,  0.05053711,  0.12890625, -0.01867676,
  0.25      ,  0.15234375,  0.0035553 , -0.09814453,  0.11523438,
 -0.12060547, -0.05078125,  0.19238281, -0.08935547, -0.15527344,
 -0.27148438, -0.27734375,  0.06884766,  0.21582031, -0.06591797,
 -0.23828125,  0.15039062, -0.43359375, -0.05761719,  0.13867188,
  0.11376953,  0.01275635,  0.02026367,  0.0703125 , -0.24511719,
  0.01806641, -0.09765625, -0.13867188,  0.2109375 , -0.28515625,
 -0.23632812,  0.10546875,  0.11572266, -0.01525879,  0.00775146,
 -0.04321289, -0.11425781, -0.25976562,  0.0168457 , -0.10400391,
 -0.01080322, -0.21289062,  0.20898438, -0.34960938,  0.27734375,
 -0.11621094, -0.21484375,  0.2421875 , -0.3125    ,  0.20605469,
 -0.22363281,  0.0612793 ,  0.09814453, -0.15820312, -0.05688477,
  0.23535156, -0.33789062,  0.05786133, -0.0300293 ,  0.24121094],
dtype=float32)}
```

Ahora que el modelo está cargado, podemos echar un vistazo a las representaciones de palabras. Primero, tenga en cuenta que **word_embeddings** es un diccionario. Cada palabra es la clave de la entrada, y el valor es su correspondiente vector de presentación.

Recuerde que los corchetes permiten el acceso a cualquier entrada si existe la clave.

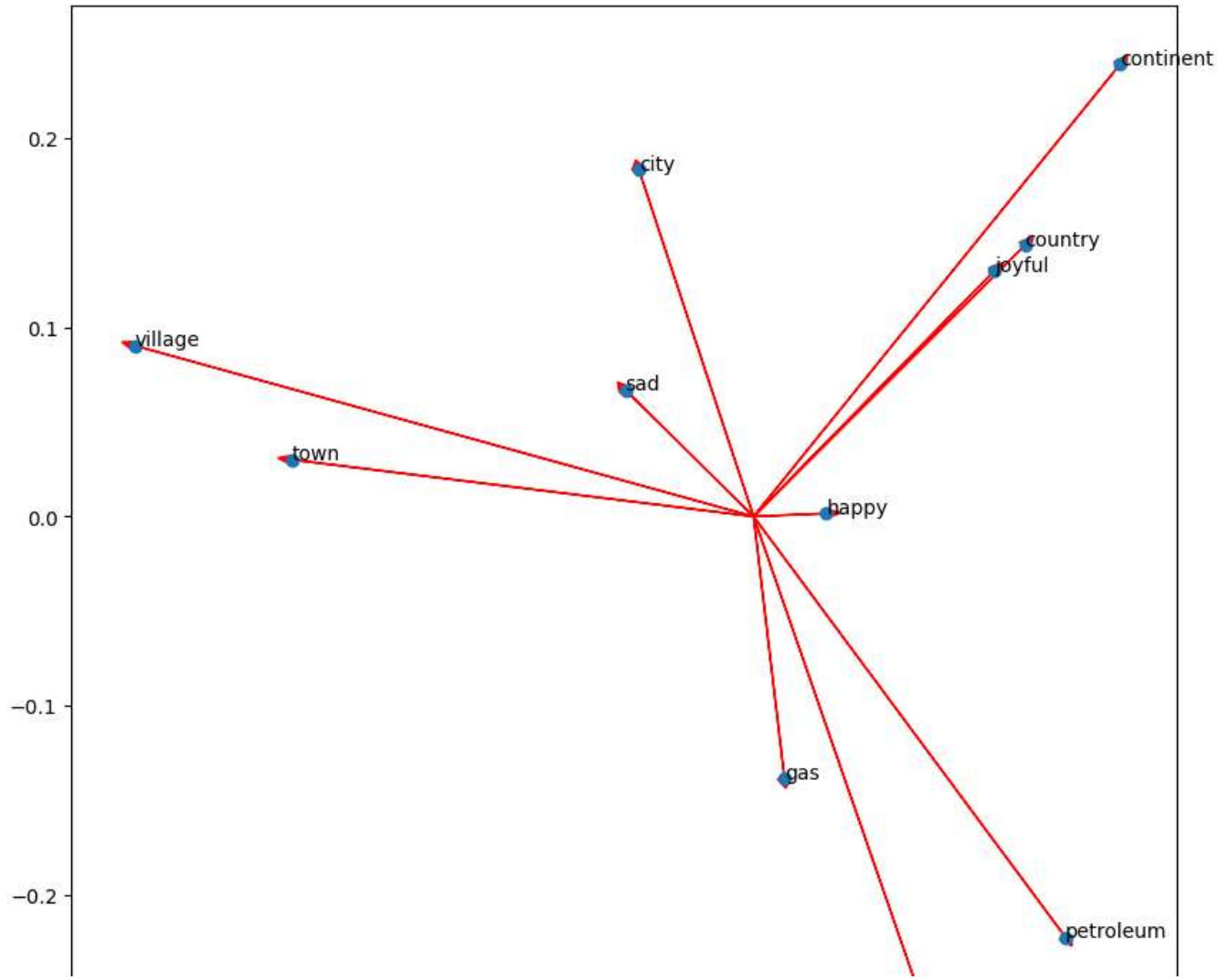
```
# Obtiene la representación vectorial de la palabra 'country'
countryVector = word_embeddings['Paris']
# Imprime el tipo del vector. Tenga en cuenta que es una matriz numpy
print(type(countryVector))
# Imprime los valores del vector.
print(len(countryVector))
```

```
<class 'numpy.ndarray'>
300
```

Es importante tener en cuenta que almacenamos cada vector como una arrayNumPy. Nos permite usar las operaciones de álgebra lineal en él.

Los vectores tienen un tamaño de 300, mientras que el tamaño del vocabulario de Google News es de alrededor de 3 millones de palabras.

```
# Función que obtiene el vector para una palabra dada:
def vec(w):
    return word_embeddings[w]


import matplotlib.pyplot as plt

words = ['oil', 'gas', 'happy', 'sad', 'city', 'town', 'village', 'country', 'continent', 'pe

# Convierte cada palabra a su representación vectorial
bag2d = np.array([vec(word) for word in words])

# Crea imagen de tamaño personalizado
fig, ax = plt.subplots(figsize = (10, 10))

col1 = 3 # Seleccione la columna para el eje x
col2 = 2 # Seleccione la columna para el eje y

# Imprime una flecha para cada palabra
for word in bag2d:
    ax.arrow(0, 0, word[col1], word[col2], head_width=0.005, head_length=0.005, fc='r', ec='r

# Plot un punto para cada palabra
ax.scatter(bag2d[:, col1], bag2d[:, col2]);

# Agrega la  etiqueta sobre cada punto en el diagrama de dispersión
for i in range(0, len(words)):
    ax.annotate(words[i], (bag2d[i, col1], bag2d[i, col2]))


plt.show()
```



Operando en word embeddings

Recuerde que comprender los datos es uno de los pasos más críticos en Data Science. Las word embeddingsson el resultado de procesos de aprendizaje automático y serán parte de la entrada para procesos posteriores. Este word embeddings debe validarse o al menos entenderse porque el rendimiento del modelo derivado dependerá en gran medida de su calidad.

Los word embeddings son arrays multidimensionales, generalmente con cientos de atributos que plantean un desafío para su interpretación.

En este Laboratorio, inspeccionaremos visualmente word embeddings de algunas palabras usando un par de atributos. Los atributos sin procesar no son la mejor opción para la creación de dichos gráficos, pero nos permitirán ilustrar la parte mecánica en Python.

En la siguiente celda, hacemos un gráfico para las word embeddings de algunas palabras. Incluso si trazar los puntos da una idea de las palabras, las representaciones de flechas también ayudan a visualizar la alineación del vector.

Tenga en cuenta que palabras similares como "village" y "town" o "petroleum", "oil" y "gas" tienden a apuntar en la misma dirección. Además, tenga en cuenta que 'sad' y 'happy' se parecen mucho; sin embargo, los vectores apuntan en direcciones opuestas.

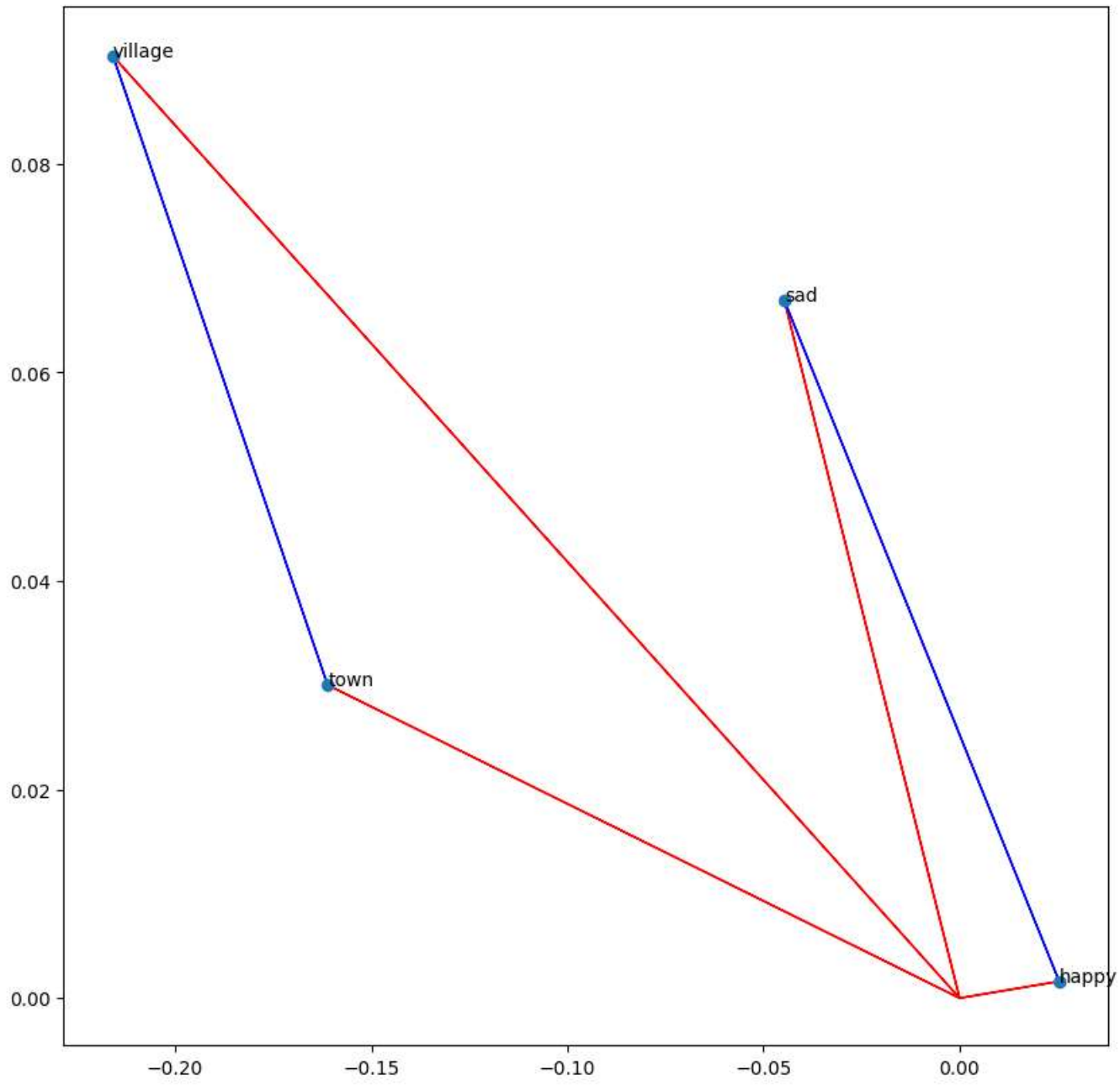
En este cuadro, uno puede calcular los ángulos y las distancias entre las palabras. Algunas palabras están cerca en ambos tipos de métricas de distancia.

Distancia entre palabras

Ahora escribe las palabras 'sad', 'happy', 'town' y 'village'. En este mismo gráfico, muestra el vector de 'village' a 'town' y el vector de 'sad' a 'happy'. Usemos NumPy para estas operaciones de álgebra lineal.

```
words = ['sad', 'happy', 'town', 'village']

# Convierte cada palabra a su representación vectorial
bag2d = np.array([vec(word) for word in words])

fig, ax = plt.subplots(figsize = (10, 10)) # Create custom size image

col1 = 3 # Select the column for the x axe
col2 = 2 # Select the column for the y axe

# Imprime un flecha para cada palabra
for word in bag2d:
    ax.arrow(0, 0, word[col1], word[col2], head_width=0.0005, head_length=0.0005, fc='r', ec=

# Imprime el vector diferencia entre village y town
village = vec('village')
town = vec('town')
diff = town - village
ax.arrow(village[col1], village[col2], diff[col1], diff[col2], fc='b', ec='b', width = 1e-5)

# Imprime el vector diferencia entre sad y happy
sad = vec('sad')
happy = vec('happy')
diff = happy - sad
ax.arrow(sad[col1], sad[col2], diff[col1], diff[col2], fc='b', ec='b', width = 1e-5)


ax.scatter(bag2d[:, col1], bag2d[:, col2]); # Plot a dot for each word

# Agregue la  etiqueta sobre cada punto en el diagrama
for i in range(0, len(words)):
    ax.annotate(words[i], (bag2d[i, col1], bag2d[i, col2]))
```

```
plt.show()
```



Algebra Lineal en word embeddings

En clases se vio las relaciones entre las palabras usando álgebra lineal en word embeddings. Veamos cómo hacerlo en Python con Numpy. Para empezar, obtenga la **norma** de una palabra en word embedding.

```
# Imprime la norma de la palabra town
print(np.linalg.norm(vec('town')))
# Imprime la norma de la palabra sad
print(np.linalg.norm(vec('sad')))
```

```
2.3858097
2.9004838
```

1. Predicción de capitales

Ahora, aplicando la diferencia y la suma de vectores, se puede crear una representación vectorial para una nueva palabra. Por ejemplo, podemos decir que el vector diferencia entre **France** y **Paris** representa el concepto de **Capital**.

Uno puede moverse desde la ciudad de Madrid en la dirección del concepto de Capital, y obtener algo cercano al país correspondiente al cual Madrid es la Capital.

```
capital = vec('France') - vec('Paris')
country = vec('Madrid') + capital

print(country[0:5]) # Imprime los primeros 5 valores del vector
```

```
[-0.02905273 -0.2475586   0.53952026  0.20581055 -0.14862823]
```

Podemos observar que el vector **country** que esperábamos que fuera el mismo que el vector de **Spain** no lo es exactamente.

```
diff = country - vec('Spain')
print(diff[0:5])
```

```
[-0.06054688 -0.06494141  0.37643433  0.08129883 -0.13007355]
```

Entonces, tenemos que buscar las palabras más cercanas en el word embbeding que coincidan con el país candidato. Si el word embbeding funciona como se esperaba, la palabra más similar debe ser 'Spain'. Definamos una función que nos ayude a hacerlo. Almacenaremos nuestro word embedding como un DataFrame, que facilita las operaciones de búsqueda basadas en los vectores numéricos.

```
# Crea u dataframe a partir del diccionario embedding.
# Esto facilita las operaciones algebraicas.
keys = word_embeddings.keys()
data = []
for key in keys:
```

```
    data.append(word_embeddings[key])

embedding = pd.DataFrame(data=data, index=keys)

# Defina una función para encontrar la palabra más cercana a un vector:
def find_closest_word(v, k = 1):
    # Calcula el vector diferencia de cada palabra al vector de entrada
    diff = embedding.values - v
    # Obtenga la norma de cada vector de diferencia.
    # Significa la distancia euclidiana al cuadrado de cada palabra al vector de entrada
    delta = np.sum(diff * diff, axis=1)
    # Encuentre el índice de la distancia mínima en la matriz
    i = np.argmin(delta)
    # Devuelve el nombre de la fila para este elemento
    return embedding.iloc[i].name



# Imprime algunas filas de embedding como un Dataframe
embedding.head(10)
```

	0	1	2	3	4	5	6	
country	-0.080078	0.133789	0.143555	0.094727	-0.047363	-0.023560	-0.008545	-0.18652
city	-0.010071	0.057373	0.183594	-0.040039	-0.029785	-0.079102	0.071777	0.01330
China	-0.073242	0.135742	0.108887	0.083008	-0.127930	-0.227539	0.151367	-0.04565
Iraq	0.191406	0.125000	-0.065430	0.060059	-0.285156	-0.102539	0.117188	-0.35156
oil	-0.139648	0.062256	-0.279297	0.063965	0.044434	-0.154297	-0.184570	-0.49804
town	0.123535	0.159180	0.030029	-0.161133	0.015625	0.111816	0.039795	-0.19628
Canada	-0.136719	-0.154297	0.269531	0.273438	0.086914	-0.076172	-0.018677	0.00625
London	-0.267578	0.092773	-0.238281	0.115234	-0.006836	0.221680	-0.251953	-0.05542
England	-0.198242	0.115234	0.062500	-0.058350	0.226562	0.045898	-0.062256	-0.20214
Australia	0.048828	-0.194336	-0.041504	0.084473	-0.114258	-0.208008	-0.164062	-0.26953

10 rows × 300 columns

Ahora busquemos el nombre que corresponde a nuestro país numérico:

```
find_closest_word(country)
```

```
'Spain'
```

2. Predicción de otros países

```
find_closest_word(vec('Italy') - vec('Rome') + vec('Madrid'))
```

```
'Spain'
```

```
print(find_closest_word(vec('Berlin') + capital))
print(find_closest_word(vec('Beijing') + capital))
```

```
Germany
China
```

Sin embargo, no siempre funciona.

```
print(find_closest_word(vec('Lisbon') + capital))
```

```
Lisbon
```

Representar una oración como un vector

Una oración completa se puede representar como un vector sumando todos los vectores de palabras que conforman la oración.

```
doc = "Spain petroleum city king"
vdoc = [vec(x) for x in doc.split(" ")]
doc2vec = np.sum(vdoc, axis = 0)
doc2vec
```

```
array([ 2.87475586e-02,  1.03759766e-01,  1.32629395e-01,  3.33007812e-01,
       -2.61230469e-02, -5.95703125e-01, -1.25976562e-01, -1.01306152e+00,
       -2.18544006e-01,  6.60705566e-01, -2.58300781e-01, -2.09960938e-02,
       -7.71484375e-02, -3.07128906e-01, -5.94726562e-01,  2.00561523e-01,
       -1.04980469e-02, -1.10748291e-01,  4.82177734e-02,  6.38977051e-01,
        2.36083984e-01, -2.69775391e-01,  3.90625000e-02,  4.16503906e-01,
        2.83416748e-01, -7.25097656e-02, -3.12988281e-01,  1.05712891e-01,
        3.22265625e-02,  2.38403320e-01,  3.88183594e-01, -7.51953125e-02,
       -1.26281738e-01,  6.60644531e-01, -7.89794922e-01, -7.04345703e-02,
       -1.14379883e-01, -4.78515625e-02,  4.76318359e-01,  5.31127930e-01,
        8.10546875e-02, -1.17553711e-01,  1.02050781e+00,  5.59814453e-01,
       -1.17187500e-01,  1.21826172e-01, -5.51574707e-01,  1.44531250e-01,
       -7.66113281e-01,  5.36102295e-01, -2.80029297e-01,  3.85986328e-01,
       -2.39135742e-01, -2.86865234e-02, -5.10498047e-01,  2.59658813e-01,
       -7.52929688e-01,  4.32128906e-02, -7.17773438e-02, -1.26708984e-01,
        4.40673828e-02,  5.12939453e-01, -5.15808105e-01,  1.20117188e-01,
```

```
-5.52978516e-02, -3.92089844e-01, -3.15917969e-01,  1.57226562e-01,
-3.19702148e-01,  1.75170898e-01, -3.81835938e-01, -2.07031250e-01,
-4.72717285e-02, -2.79296875e-01, -3.29040527e-01, -1.69067383e-01,
 1.61132812e-02,  1.71569824e-01,  5.73730469e-02, -2.44140625e-03,
 8.34960938e-02, -1.58203125e-01, -3.10119629e-01,  5.28564453e-02,
 8.60595703e-02,  5.12695312e-02, -7.22900391e-01,  4.97924805e-01,
-5.85937500e-03,  4.49951172e-01,  3.82446289e-01, -2.80029297e-01,
-3.28125000e-01, -6.27441406e-02, -4.81933594e-01,  1.93176270e-02,
-1.69326782e-01, -4.28649902e-01,  5.39062500e-01, -1.28417969e-01,
-8.83789062e-02,  5.13916016e-01,  9.13085938e-02, -1.60156250e-01,
 6.86035156e-02, -9.74121094e-02, -3.70712280e-01, -3.27270508e-01,
 1.77978516e-01, -4.65332031e-01,  1.70410156e-01,  9.08203125e-02,
 2.76857376e-01, -1.69677734e-01,  3.27728271e-01, -3.12500000e-02,
-2.20809937e-01, -3.46679688e-01,  4.67407227e-01,  5.31860352e-01,
-1.30615234e-01, -2.36816406e-02, -6.56250000e-01, -5.79589844e-01,
-2.05810547e-01, -3.03222656e-01,  1.94259644e-01, -7.28515625e-01,
-4.92522240e-01, -5.37109375e-01, -3.47656250e-01,  1.08642578e-01,
-1.41601562e-01, -2.07031250e-01,  2.52441406e-01, -7.78808594e-02,
-5.02441406e-01,  1.53808594e-02,  8.64257812e-02,  2.59765625e-01,
 6.64062500e-02, -7.12890625e-01, -1.45751953e-01,  7.56835938e-03,
 4.87792969e-01,  1.39160156e-01,  1.15722656e-01,  1.28662109e-01,
-4.75585938e-01,  2.21191406e-01,  3.25317383e-01,  1.06323242e-01,
-6.11083984e-01, -3.59619141e-01,  6.54296875e-02, -2.41699219e-01,
-6.29882812e-02, -1.62109375e-01,  4.26269531e-01, -4.38354492e-01,
 1.93725586e-01,  4.89562988e-01,  5.31494141e-01, -7.29370117e-02,
 1.77246094e-01,  9.39941406e-02,  2.92236328e-01, -2.74047852e-01,
 2.63366699e-02,  4.36035156e-01, -3.76953125e-01,  3.10546875e-01,
 4.87304688e-01, -2.43041992e-01,  1.21612549e-02, -3.80371094e-01,
 3.80493164e-01, -6.22436523e-01, -3.98071289e-01,  1.24206543e-01,
-8.20312500e-01, -2.72583008e-01, -6.21582031e-01, -4.87060547e-01,
 3.06671143e-01, -2.61230469e-01,  5.12451172e-01,  5.55694580e-01,
 5.66894531e-01,  7.33886719e-01, -1.75781250e-01,  4.13574219e-01,
-2.54272461e-01,  1.32507324e-01, -4.78515625e-01,  4.63256836e-01,
-6.21948242e-02, -1.80664062e-01, -5.46386719e-01, -6.31103516e-01,
-1.47949219e-01, -3.15185547e-01, -7.12890625e-02, -7.67578125e-01,
 3.92272949e-01, -1.97753906e-01,  2.23144531e-01, -5.07324219e-01,
 8.39843750e-02, -4.98657227e-02,  1.01074219e-01,  2.07885742e-01,
-2.77343750e-01,  1.03027344e-01, -1.38671875e-01,  2.87353516e-01,
-4.81895447e-01, -1.66748047e-01, -1.47277832e-01,  3.61633301e-01,
 6.38504028e-02, -6.69189453e-01,  1.95312500e-03, -7.34375000e-01,
-1.28158569e-01,  9.76562500e-04, -7.08007812e-02,  3.72558594e-01,
 8.31176758e-01,  5.94482422e-01,  5.37109375e-02,  3.00140381e-01
```

```
find_closest_word(doc2vec)
```

```
'petroleum'
```

ACTIVIDAD

A partir del desarrollo del laboratorio escriba 4 conclusiones del Laboratorio

- Para comenzar word embeddings son una representación vectorial que captura el significado semántico de las palabras, permitiendo agrupar palabras con similitudes a áreas cercanas del espacio vectorial.
- Los word embeddings son útiles en actividades o tareas de PLN para la clasificación, generación de texto y traducción automática, pero presentan limitaciones al no poder capturar completo el contexto y la similitud de las palabras.
- Para un mejora y garantizar una representación mas precisa de los word embeddings es a su uso con mas datos durante un entrenamiento, considerando la cantidad, calidad y relevancia de los datos a utilizar.
- Los word embeddings por mas limitaciones que tenga, siguen siendo una herramienta principal para el PLN, logra facilitarnos la identificacion entre palabras con similitud.

✓ 0 s se ejecutó 14:20