Quintana Martínez Erick
Tratamiento de lenguaje natural

Similitud coseno entre textos usando TF-IDF

Código

```
import numpy as np
from math import *

def minusculas(texto):
    return texto.lower()

def remover_puntuaciones(texto):
    puntuaciones = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

    for c in texto:
        if c in puntuaciones:
            texto = texto.replace(c, '')

    return texto

def tokenizar(texto):
    return texto.split(' ')

def cuenta_palabra_documentos(textos):
    ocurrencias = {}

    for palabra in vocabulario:
        ocurrencias[palabra] = 0
        for texto in textos:
            if palabra in texto:
                ocurrencias[palabra] += 1

    return ocurrencias

def frecuencia_termino(texto, palabra):
    n = len(texto)
    ocurrencia = len([token for token in texto if token == palabra])

    return ocurrencia / n

def frecuencia_inversa_documentos(palabra):
    try:
        ocurrencia_palabra = ocurrencias[palabra] + 1
    except:
        ocurrencia_palabra = 1

    return np.log(total_documentos / ocurrencia_palabra)

def tf_idf(texto):
    tf_idf_vec = np.zeros((len(vocabulario),))

    for palabra in texto:
        tf = frecuencia_termino(texto, palabra)
        idf = frecuencia_inversa_documentos(palabra)

        valor = tf * idf
        tf_idf_vec[diccionario_palabras[palabra]] = valor
```

```
    return tf_idf_vec

textos = ["La inteligencia artificial (IA) actual funciona bien cuando la
aplicas a un ámbito muy concreto: desde jugar al ajedrez a hacer un
diagnóstico médico. Pero el gran reto en el siglo XXI, consiste en
alcanzar una inteligencia artificial general (AGI por sus siglas en
inglés), que es aquella IA capaz de aprender tareas intelectuales como lo
hacen los humanos. En este contexto, destaca GPT-3, un modelo de IA que
permite generar lenguaje escrito. Es lo que se conoce como un modelo de
lenguaje auto-regresivo, es decir, 'un algoritmo que permite crear la
siguiente mejor palabra que seguiría a un texto dado', explica César de
Pablo, científico de datos en BBVA Data &Analytics.",
          "Desde Yo, robot la humanidad ha vivido aterrada con la idea de
que los robots cobren vida y consciencia propia. Puede que la inteligencia
artificial nos ponga un paso más cerca de ese día, y según algunos eso ya
habría ocurrido en los cuarteles de una de los gigantes tecnológicos más
importantes del siglo XXI: Google. El programa LaMDA, una IA
especializada en conversaciones, pudo haber cobrado consciencia según uno
de los ingenieros participantes en el proyecto. LaMDA (Language Model for
Dialogue Applications, modelo de lenguaje para aplicaciones de diálogo en
español) es un programa diseñado para tener conversaciones realistas con
un ser humano prestando atención a 'un aparente sinnúmero de temas', como
sucede normalmente en nuestras pláticas.",
          "El francés Aspect, el estadounidense Clauser y el austriaco
Zeilinger son tres pioneros de los revolucionarios mecanismos de la física
cuántica. Han sido premiados por sus descubrimientos sobre el
'entrelazamiento cuántico', un fenómeno por el que dos partículas
cuánticas están perfectamente correlacionadas, independientemente de la
distancia que las separe, segñun ha anunciado el jurado del Nobel."]

total_documentos = len(textos)

i = 0
vocabulario = []

for texto in textos:
    texto_aux = minusculas(texto)
    texto_aux = remover_puntuaciones(texto_aux)
    texto_tokenizado = tokenizar(texto_aux)

    textos[i] = texto_tokenizado

    for palabra in textos[i]:
        if palabra not in vocabulario:
            vocabulario.append(palabra)

    i += 1

diccionario_palabras = {}
i = 0

for palabra in vocabulario:
    diccionario_palabras[palabra] = i

    i += 1

ocurrencias = cuenta_palabra_documentos(textos)

vectores = []
```

```python
for texto in textos:
    vec = tf_idf(texto)
    vectores.append(vec)

print(vectores)

def raiz_cuadrada(x):
    return round(sqrt(sum([a * a for a in x])),4)

def similitudCoseno(x, y):
    numerador = sum(a * b for a, b in zip(x, y))
    denominador = raiz_cuadrada(x) * raiz_cuadrada(y)

    sim_cos = round(numerador/float(denominador),4)

    return sim_cos

print(f"Similitud entre texto 1 y texto 2 = {similitudCoseno(vectores[0], vectores[1])}")
print(f"Similitud entre texto 1 y texto 3 = {similitudCoseno(vectores[0], vectores[2])}")
print(f"Similitud entre texto 2 y texto 3 = {similitudCoseno(vectores[1], vectores[2])}")
```

Salida

Texto 1

```
[-0.00770577  0.          0.          0.          0.00362022  0.00362022
  0.00362022  0.00362022  0.00362022  0.         -0.01541154  0.00362022
  0.00362022  0.00362022  0.          0.00362022  0.00362022  0.00362022
  0.00362022  0.00362022  0.00362022  0.00362022 -0.00513718  0.00362022
  0.00362022  0.          0.          0.          0.00362022  0.00362022
  0.          0.00362022  0.00362022  0.          0.          0.00362022
  0.00362022 -0.01284295  0.          0.00362022  0.00362022 -0.01284295
  0.00362022  0.00362022  0.00362022  0.          0.00724045  0.00362022
 -0.00256859  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022
  0.          0.00724045  0.00362022  0.          0.00362022  0.00362022
  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022
  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022
  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022  0.00362022
  0.00362022  0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.        ]
```

Texto 2

```
[-0.00725249  0.          0.          0.          0.          0.
  0.          0.          0.          0.         -0.00966999  0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.         -0.00483499  0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.         -0.00483499  0.          0.          0.         -0.01933997
  0.          0.          0.          0.          0.          0.
 -0.00966999  0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.00340727  0.00340727  0.00340727  0.          0.00340727
  0.00340727  0.00681454  0.00340727  0.00340727  0.00340727  0.00340727
  0.          0.00681454  0.00340727  0.00340727  0.00340727  0.00340727
  0.00340727  0.00681454  0.00340727  0.00340727  0.00340727  0.00681454
  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727
  0.00340727  0.00340727  0.00340727  0.          0.00340727  0.00681454
  0.00681454  0.00340727  0.00340727  0.00681454  0.00340727  0.00340727
  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727
  0.00340727  0.00340727  0.00340727  0.00340727  0.00681454  0.00340727
  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727
  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727  0.00340727
  0.00340727  0.00340727  0.00340727  0.00340727  0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.        ]
```

Texto 3

```
[-0.01027436  0.          0.          0.          0.          0.
  0.          0.          0.          0.         -0.00513718  0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.         -0.03082308  0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.         -0.01027436  0.          0.          0.         -0.01541154
  0.          0.          0.          0.          0.          0.
 -0.00513718  0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.          0.
  0.          0.          0.          0.          0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045  0.00724045
  0.00724045  0.00724045  0.00724045]
```

Similitud entre textos

```
Similitud entre texto 1 y texto 2 = 0.3535
Similitud entre texto 1 y texto 3 = 0.2975
Similitud entre texto 2 y texto 3 = 0.2834
```