Anexo 14

Proyecto 14: Resumen de textos con Vectorización

Mg. Luis Felipe Bustamante Narváez

En este proyecto, desarrollaremos un generador de resúmenes a través de vectorización, tomando una base de datos importante, de diferentes artículos de prensa para analizar cómo pueden sumarizarse y permitir los beneficios de la inteligencia artificial.

Librerías

In [...
```
import pandas as pd
import numpy as np
import textwrap
import nltk
from nltk.corpus import stopwords
from nltk import word_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [...
```
nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\luis_\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\luis_\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[... True

In [...
```
path = 'data/df_total.csv'
df = pd.read_csv(path, encoding='utf-8')
```

In [...
```
df
```

Out[...

	url	news	Type
0	https://www.larepublica.co/redirect/post/3201905	Durante el foro La banca articulador empresari...	Otra
1	https://www.larepublica.co/redirect/post/3210288	El regulador de valores de China dijo el domin...	Regulaciones
2	https://www.larepublica.co/redirect/post/3240676	En una industria históricamente masculina como...	Alianzas
3	https://www.larepublica.co/redirect/post/3342889	Con el dato de marzo el IPC interanual encaden...	Macroeconomia
4	https://www.larepublica.co/redirect/post/3427208	Ayer en Cartagena se dio inicio a la versión n...	Otra
...	...	...	...
1212	https://www.bbva.com/es/como-lograr-que-los-in...	En la vida de toda empresa emergente llega un ...	Innovacion
1213	https://www.bbva.com/es/podcast-como-nos-afect...	La espiral alcista de los precios continúa y g...	Macroeconomia
1214	https://www.larepublica.co/redirect/post/3253735	Las grandes derrotas nacionales son experienci...	Alianzas
1215	https://www.bbva.com/es/bbva-y-barcelona-healt...	BBVA ha alcanzado un acuerdo de colaboración c...	Innovacion
1216	https://www.larepublica.co/redirect/post/3263980	Casi entrando a la parte final de noviembre la...	Alianzas

1217 rows × 3 columns

In [...
```
print(df['news'][2][:500])
```

```
En una industria históricamente masculina como lo es la aviación Viva presentó su avión rosad
o A320NEO que apuesta por la equidad de género la lucha contra el cáncer de mama la inclusión
y la diversidad.Desde Francia llegó Go Pink que tuvo un precio promedio de US50 millones conv
irtiéndose en la aeronave número 20 de las 21 con las que finalizará el año esta aerolínea.En
Viva estamos trabajando muy fuerte para que haya más mujeres. Actualmente el grupo ejecutivo
está compuesto por 42 mujeres per
```

In [...
```
# Buscamos una noticia larga para tomar como ejemplo a la hora de hacer el resumen
doc = df['news'].sample()
```

In [...
```
print(doc.iloc[0][:500])
```

```
El actual brote de inflación es un momento de déjà vu para las personas que vivieron las subi
das de precios de principios de la década de 1980.La inflación de EE.UU. se aceleró a una tas
a anual de 75 en enero alcanzando un máximo de cuatro décadas. El índice de precios al consum
idor que mide lo que la gente paga por bienes y servicios estuvo el mes pasado en su nivel má
s alto desde febrero de 1982 en comparación con enero de hace un año según el Departamento de
Trabajo.Blaise Jones recuerda a su
```

In [...
```
# Obtener el índice de la noticia para manipulación de datos
indice = df.index[df['news'] == doc.iloc[0]].tolist()
print(indice)
```

```
[297]
```

In [...
```
# hacemos la prueba
print(df['news'][297][:100])
```

```
El actual brote de inflación es un momento de déjà vu para las personas que vivieron las subi
das de
```

Procesamiento de Datos

TextWrap

In [...
```
# Eliminamos las palabras cortadas de las líneas
doc2 = textwrap.fill(doc.iloc[0], replace_whitespace=False, fix_sentence_endings=True)
```

In [...
```
print(doc2[:500])
```

```
El actual brote de inflación es un momento de déjà vu para las
personas que vivieron las subidas de precios de principios de la
década de 1980.La inflación de EE.UU. se aceleró a una tasa anual de
75 en enero alcanzando un máximo de cuatro décadas.  El índice de
precios al consumidor que mide lo que la gente paga por bienes y
servicios estuvo el mes pasado en su nivel más alto desde febrero de
1982 en comparación con enero de hace un año según el Departamento de
Trabajo.Blaise Jones recuerda a s
```

Separación de líneas

In [...
```
# Podemos separar por líneas, por puntos o comas, la idea es conservar oraciones
# con ideas claras.
lineas = doc2.split('. ')  #Usamos punto espacio, por las siglas o números que pueden haber
lineas[:3]
```

Out[…
```
['El actual brote de inflación es un momento de déjà vu para las\npersonas que vivieron las subidas de precios de principios de la\ndécada de 1980.La inflación de EE.UU',
 'se aceleró a una tasa anual de\n75 en enero alcanzando un máximo de cuatro décadas',
 ' El índice de\nprecios al consumidor que mide lo que la gente paga por bienes y\nservicios estuvo el mes pasado en su nivel más alto desde febrero de\n1982 en comparación con enero de hace un año según el Departamento de\nTrabajo.Blaise Jones recuerda a su madre hablando sobre el aumento del\nprecio de la leche y la determinación de su padre de mantener baja la\nfactura de calefacción de su hogar tácticas que incluían bajar el\ntermostato a 62 grados a la hora de acostarse.Juro que podía ver mi\nrespiración cuando me levantaba dijo el doctor Jones ahora de 59 años\ny neurorradiólogo pediátrico en Cincinnati']
```

Tokenización

In […
```
# Creamos la tokenización usando las stopwords descargadas
tokenizar = TfidfVectorizer(stop_words=stopwords.words('spanish'), norm='l1')
```

In […
```
# Creamos la matriz
X = tokenizar.fit_transform(lineas)
X
```

Out[…
```
<26x433 sparse matrix of type '<class 'numpy.float64'>'
        with 560 stored elements in Compressed Sparse Row format>
```

In […
```
# Mostramos la matriz
filas, columnas = X.shape

for i in range(10):  #aquí ponemos las filas, pero al ser muchas el resultado es extenso.
    for j in range(10):
        print(X[i, j], end=' ') # imprime el elemento y un espacio en blanco
    print()  #deja el renglón

# Cada fila va a representar una palabra
# Cada columna va a representar cada una de las oraciones
```

```
0.0 0.0 0.0 0.0 0.0 0.04756870737569473 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 0.018470715419332654 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 0.08059297416517343 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 0.0 0.0 0.0
```

In […
```
# Promediamos los puntajes de cada una de las oraciones
def obtener_score(tfidf_row):
    x = tfidf_row[tfidf_row != 0]  # Elimina las oraciones que no tienen puntuación
    return x.mean()
```

In […
```
# Creamos el vector de puntuación y lo llenamos
scores = np.zeros(len(lineas))
for i in range(len(lineas)):
    score = obtener_score(X[i,:])
    scores[i] = score
```

In […
```
scores
```

Out[…
```
array([0.06666667, 0.11111111, 0.01754386, 0.05      , 0.07692308,
       0.0625    , 1.        , 0.25      , 0.04761905, 0.03703704,
       0.04545455, 0.02439024, 0.03846154, 0.02083333, 0.06666667,
       0.03225806, 0.04545455, 0.0625    , 0.05263158, 0.03225806,
       0.125     , 0.09090909, 0.03703704, 0.07692308, 0.03030303,
       0.07142857])
```

Resumen

In […
```
# Ordenamos los scores y mostramos las posiciones de mayor a menor
sort_index = np.argsort(-scores)
```

In […
```
sort_index
```

Out[…
```
array([ 6,  7, 20,  1, 21, 23,  4, 25, 14,  0, 17,  5, 18,  3,  8, 10, 16,
       12,  9, 22, 15, 19, 24, 11, 13,  2], dtype=int64)
```

In […
```
# Resumen desordenado
oraciones = []
cantidad_oraciones = 10
for i in range(cantidad_oraciones):
    oraciones.append([sort_index[i], scores[sort_index[i]], lineas[sort_index[i]]])
    print(f'{scores[sort_index[i]]}:\n{lineas[sort_index[i]]}\n')
```

```
1.0:
 Jones

0.25:
 Él y su esposa han sido frugales durante mucho
tiempo

0.125:
 Está posponiendo la
instalación de un nuevo revestimiento de aluminio en su casa porque
los precios han subido

0.1111111111111111:
se aceleró a una tasa anual de
75 en enero alcanzando un máximo de cuatro décadas

0.09090909090909093:
 Pagó el préstamo de su automóvil hace unos
dos años pero continúa conduciendo un BMW 2014 golpeado

0.07692307692307694:
El fundador de la empresa de ferias comerciales Shamrock
Productions condujo su Oldsmobile durante 450.000 millas hasta que el
motor explotó

0.07692307692307694:
 dos veces por semana
para llenar la camioneta de la familia y tratar de evitar las colas en
la bomba durante la crisis energética de 1979

0.07142857142857144:
 Ella también se está saltando vacaciones costosas
por ahora.A cada paso la gente se ve afectada por el aumento de los
precios dijo la señora Navratil.

0.0666666666666667:
 Cuando cerró la compra de la casa
la tasa hipotecaria que el corredor había ofrecido saltó a cerca de
135 desde alrededor de 1275 que tenía cuando había iniciado el proceso
dijo

0.0666666666666667:
El actual brote de inflación es un momento de déjà vu para las
personas que vivieron las subidas de precios de principios de la
década de 1980.La inflación de EE.UU
```

In […

```python
# Ordenamiento de la lista por el primer elemento de cada sublista
oraciones_sort = sorted(oraciones, key=lambda x:x[0])

#Imprimimos la lista ordenada
for item in oraciones_sort:
    print(item[2]) # el 2 es la columna de las líneas
```

El actual brote de inflación es un momento de déjà vu para las personas que vivieron las subidas de precios de principios de la década de 1980.La inflación de EE.UU se aceleró a una tasa anual de 75 en enero alcanzando un máximo de cuatro décadas dos veces por semana para llenar la camioneta de la familia y tratar de evitar las colas en la bomba durante la crisis energética de 1979 Jones Él y su esposa han sido frugales durante mucho tiempo Cuando cerró la compra de la casa la tasa hipotecaria que el corredor había ofrecido saltó a cerca de 135 desde alrededor de 1275 que tenía cuando había iniciado el proceso dijo Está posponiendo la instalación de un nuevo revestimiento de aluminio en su casa porque los precios han subido Pagó el préstamo de su automóvil hace unos dos años pero continúa conduciendo un BMW 2014 golpeado El fundador de la empresa de ferias comerciales Shamrock Productions condujo su Oldsmobile durante 450.000 millas hasta que el motor explotó Ella también se está saltando vacaciones costosas por ahora.A cada paso la gente se ve afectada por el aumento de los precios dijo la señora Navratil.

Conclusiones

Hace algunos años, la vectorización era la manera más adecuada para generar resúmenes, como podemos notar en los resultados, hace falta un poco de coherencia, pero se puede entender la idea del texto que se pretende resumir.

Mg. Luis Felipe Bustamante Narváez