Práctica 3.3

Procesamiento del lenguaje natural

Analizando el contenido de los medios de comunicación

Enunciado

Los medios de comunicación generalistas nos dan una visión sobre qué está ocurriendo en el mundo. Dada la gran cantidad de artículos de noticias en circulación, identificarlos y organizarlos por tema se revela como una tarea que puede ser útil. Ello podría ayudarnos en la tarea de examinar la ingente cantidad de información que, por ejemplo, en el caso de los periódicos online, se publica diariamente, a fin de encontrar noticias que sean relevantes para lo que uno/a necesita.

Con lo anterior nos referimos a que si, por ejemplo, en cierto periódico se repite mucho un término determinado dentro del *corpus* de noticias (publicadas en un mismo día, semana, mes, ...) del que forma parte será porque el tema al que remite ese término resulta importante y es parte de los asuntos de actualidad que conciernen a ese periódico.

Objetivo

En este proyecto, utilizaremos el modelo **"término frecuencia - frecuencia inversa de documentos (tf-idf)"** para analizar el contenido de una serie de artículos de prensa y descubrir qué términos los describen mejor, proporcionando así una visión rápida del tema de que tratan.

Procedimiento

Realizaremos a tal fin los siguientes pasos:

1. Importación y preparación de datos
2. Cálculo de las puntuaciones tf-idf
3. Análisis de resultados y conclusiones

1) Importación y preparación de datos

1. Crea un cuaderno (en Google Colab o en Visual Studio).

2. Lleva al cuaderno la selección de 10 artículos del periódico *The News International* que tienes en el apartado "Recursos" al final de este documento. Cada artículo, almacenado como una cadena, formará parte del *corpus* con el que vamos a trabajar.

   Elimina los valores numéricos que aparecen en algún caso en los artículos como parte de su texto, ya que no nos interesan, y guarda los artículos en una variable *articles_cleaned*.

   Antes de seguir, imprime uno de los artículos y revisa su contenido.

2) Calculando las puntuaciones *tf-idf*

3. Comenzaremos el análisis haciendo un simple recuento de palabras para cada artículo. Inicializa para ello un objeto *CountVectorizer*, asignándolo a una variable cuyo nombre puede ser *vectorizer*.

4. Prepara (es decir, *ajusta* y *transforma*) *vectorizer* para obtener el recuento de palabras de cada artículo. Guarda los recuentos resultantes en una variable que se llame *counts*. Tras guardar el recuento de palabras, visualiza por pantalla un *DataFrame* con el recuento para el conjunto de artículos.

5. Ahora que tienes el recuento de palabras de cada artículo, cabe convertirlos en puntuaciones *tf-idf*. Inicializa un objeto *TfidfTransformer* con el argumento *norm* = *None* y guárdalo en una variable llamada *transformer*.

6. *Ajusta* y *transforma* acto seguido el transformador para convertir la contabilización de palabras en puntuaciones *tf-idf* para cada uno de los artículos, guardando las puntuaciones de *tf-idf* resultantes en una variable a la que puedes llamar *tfidf_scores_transformed*.

   Tras guardar las puntuaciones de *tf-idf* en *tfidf_scores_transformed*, muestra un nuevo *DataFrame* con las puntuaciones tf-*idf* para cada artículo.

7. Ahora tenemos las puntuaciones *tf-idf* para cada artículo. Sin embargo, es conveniente confirmar que *TfidfTransformer* genera los mismos resultados que si usamos directamente *TfidfVectorizer*.

   Inicializa un objeto *TfidfVectorizer* con el argumento *norm* = *None* y guárdalo en una variable *vectorizer*.

8. Prepara el vectorizador para los artículos a fin de calcular las puntuaciones *tf-idf* para cada artículo, guardándolas en una variable a la que puedes llamar *tfidf_scores*.

   ¿Las puntuaciones de *tf-idf* generadas por *TfidfVectorizer* son iguales a las generadas por *TfidfTransformer*?

   Confirmémoslo a través de este código, que debes añadir al script.

```
print('Are the tf-idf scores the same?')
if np.allclose(tfidf_scores_transformed.todense(),
tfidf_scores.todense()):
 print('YES')
else:
 print('No, something is wrong :(')
```

   NOTA: el método ***allclose*** de *Numpy* permite comparar vectores.

3) Análisis de los resultados

9. Una forma sencilla de identificar el "tema" de un documento es etiquetar el documento con el término *tf-idf* de mayor puntuación. Si bien este enfoque no es demasiado sofisticado, es una forma rápida y fácil de obtener información sobre el tema de un documento.

   Si además echamos mano del método de Pandas *idxmax ()*, que devuelve el índice del valor más alto en una columna *DataFrame*, podemos encontrar el término *tf-idf* con la mayor puntuación para cada artículo.

   Escribe un bucle "for" que recorra el conjunto de artículos, mostrando en cada iteración el término con puntuación más alta junto con la puntuación en cuestión. El bucle debería generar una salida de este tipo:

```
Words with the highest tf-idf:
```

```
Word: fares          Score: 10.819
Word: hong           Score: 5.409
Word: sugar          Score: 18.933
Word: petrol         Score: 8.114
Word: engines        Score: 16.228
Word: australia      Score: 8.114
Word: car            Score: 8.114
Word: railways       Score: 9.197
Word: cabinet        Score: 8.114
Word: china          Score: 5.409
```

Echa un vistazo superficial al texto de los artículos y compáralo con el término seleccionado como resultado de aplicar *tf-idf*.

¿La lista de términos anterior dan alguna idea de cuál es el contenido (tema) de los artículos?

Recursos

artículos de prensa

```
article_1 = '''KARACHI: The Sindh government has decided to bring down public transport fares by 7 per cent due to massive reduction in petroleum product prices by the federal government, Geo News reported.Sources said reduction in fares will be applicable on public transport, rickshaw, taxi and other means of traveling. Meanwhile, Karachi Transport Ittehad (KTI) has refused to abide by the government decision.KTI President Irshad Bukhari said the commuters are charged the lowest fares in Karachi as compare to other parts of the country, adding that 80pc vehicles run on Compressed Natural Gas (CNG). Bukhari said Karachi transporters will cut fares when decrease in CNG prices will be made.'''

article_2 = '''HONG KONG:  Hong Kong shares opened 0.66 percent lower Monday following a tepid lead from Wall Street, as the first full week of the new year kicked off. The benchmark Hang Seng Index dipped 158.63 points to 23,699.19.'''

article_3 = '''KARACHI: Wholesale market rates for sugar dropped to less than Rs 50 per kg following the resumption of sugar cane crushing by sugar mills in
```

```
Sindh. Within two days, the rate dropped by Rs 1.70 to Rs 49.80 per kg in Karachi Whole Sale Market. According to dealers, the resumption of sugar cane crushing by the mills stabilised the supply to the market with an immediate effect on price as well. Industry experts said that the quality of sugar cane is excellent in Sindh and approximately 100 kg of sugar cane can produce 11 kg of sugar.'''

article_4 = '''ISLAMABAD: Long queues of vehicles on fuel stations were visible in different parts of the country as the petrol became rare commodity on Thursday. Federal Minister for Petroleum Shahid Khaqan Abbasi says "it may take up to ten days to bring the situation to normality". He claimed that northern areas of Pakistan had been facing the petrol shortage. The minister cited the recent decline in petroleum prices and delay in a shipment as reasons for the shortage.He said situation would improve as soon as shipment reached Pakistan. Sources told Geo News hat due to financial restraints the Pakistan State Oil has been unable import petrol.'''

article_5 = '''KARACHI: The final shipment of Chinese manufactured Rail Engines arrived in Pakistan on Friday. Federal Railways Minister, Khwaja Saad Rafique says, the inclusion of the new engines will help ease the shortfall faced by Pakistan Railways. The shipment includes 2000 and 3000-horse-power engines which will be used to pull freight bogeys. Rafique told journalists, the inclusion of 15 new engines has brought Pakistan Railways total strength to 268 engines however more engines are still required.'''

article_6 = '''SYDNEY: Cricket fever has gripped Australia with the World Cup just days away. Fans from around the world have thronged to the country and hotels are capitalising. Prices of rooms have almost doubled to 300 dollars and hotels are experiencing full bookings. Experts estimate that during the mega event Australia will generate 1.5 million US dollars just from hotel bookings. If the cost of internal air travel, taxis and tickets is taken into consideration, Australia stands to generate two million US dollars during the World Cup.'''

article_7 = '''SAN FRANCISCO: Apple Inc aims to begin producing electric vehicles as early as 2020, Bloomberg reported. The report cited people with knowledge of the matter as saying, a seemingly aggressive target for a mobile devices maker with little experience in car manufacture.The iPhone maker is pushing its "car team" of about 200 people to meet that goal. But Apple may decide to scrap its car-making effort, or delay it, if executives grew unhappy with its progress, the news agency said.'''

article_8 = '''LAHORE: Federal Minister for Railways, Khawaja Saad Rafique Tuesday announced good news of pay-raise for the employees of Pakistan Railways. In a media statement, the Minister disclosed that a summary for increase in salaries for the employees of Pakistan Railways has been forwarded to the Prime Minister. He also said that the government had also chalked out a plan to build houses for the Railways workers. Khawaja Saad Rafique said it was expected that
```

```
the salaries of Railway Police may witness a jump of 20 percent. He also announced the government\x92s plan to launch a new train service between Karachi and Islamabad.'''

article_9 = '''ISLAMABAD: The Federal Cabinet on Tuesday approved the budget strategy paper, sources revealed to Geo News. During the cabinet meeting, Prime Minister Nawaz Sharif said tax rate had to be reduced to increase revenue. He added that people would happily pay taxes if the rate was reduced. The prime minister directed the cabinet to provide maximum relief to people in the budget, emphasising that the economic impact should reach people.'''

article_10 = '''BEIJING: China will keep the yuan basically stable against a basket of currencies and there is no basis for continued yuan depreciation, central bank vice governor Yi Gang said on Sunday. China also will keep foreign exchange reserves at appropriate levels, Yi said.'''

articles = [article_1, article_2, article_3, article_4, article_5, article_6, article_7, article_8, article_9, article_10]
```