RECUPERACIÓN DE LA INFORMACIÓN

Y MINERÍA DE TEXTO

CLUSTERING DE NOTICIAS

PAULA CARBALLO PÉREZ

ESTER CORTÉS GARCÍA

ÍNDICE

1.- INTRODUCCIÓN

Hoy en día, tenemos acceso a una gran cantidad de información que está disponible en la red en cualquier momento. Una parte importante de esta información se encuentra en forma de noticias, que se nos presentan de forma multilingüe. Las noticias permiten al público estar informado y suelen relatar hechos novedosos de carácter político y/o social, y cuya extensión no suele ser demasiado larga.

Las características de las noticias nos permiten procesarlas de diversas formas, siendo una de ellas la clasificación por temas. Es importante no limitar nuestra búsqueda a noticias en un mismo idioma, pues de este modo tendremos acceso a más puntos de vista sobre el mismo tema.

El agrupamiento automático es un proceso de aprendizaje no supervisado cuyo fin es organizar una colección de objetos en diferentes grupos o clústeres, de tal manera que los objetos del mismo clúster sean lo más similares posibles entre sí, manteniendo a su vez el menor grado de similitud posible en relación con los objetos pertenecientes a otros clústeres.

Será el agrupamiento automático lo que aplicaremos en esta práctica, con el fin último de obtener una agrupación de nuestros documentos, lo más aproximada posible (sino igual) a la que se nos proporciona como referencia.

2.- PROCESADO INICIAL DE DOCUMENTOS

Se dispone de una colección de documentos de trabajo en formato html. Dicha colección está compuesta por 22 noticias obtenidas de diferentes fuentes de noticias online, en inglés y español.

En primer lugar, se realiza una inspección de los html. Puesto que lo que nos interesa es quedarnos con el texto de la noticia, en nuestro procesado nos quedaremos con las etiquetas <p>.

La transformación de los documentos a texto la realizamos utilizando la librería BeautifulSoup. Esta biblioteca crea un árbol con todos los elementos del documento y se utiliza para extraer la información.

En la siguiente imagen se muestra el método que, dadas las url, extrae la información de las distintas páginas web.

```
def read_file(file_url):
    try:
        with open(file_url,'r', encoding='utf-8') as f:
            pag = f.read()
        f.close()
    except UnicodeDecodeError:
        with open(file_url,'r', encoding='latin-1') as f:
            pag = f.read()
        f.close()
    bsObj = BeautifulSoup(pag,'html.parser')
    return(bsObj)
```

Una vez que tenemos el contenido, debemos quedarnos con la parte de la página web que nos interesa, en nuestro caso la información almacenada en las etiquetas <p>, que es la que se corresponde con el cuerpo de la noticia. Esa información será almacenada en un array, que tendrá 22 elementos, uno por cada noticia:

```
def get_original_news(url_list):
    print('Se están procesando las 22 noticias')
    text_list = []
    for u in url_list:
        if u.endswith(".html"):
            url = folder + "/" + u
            text = read_file(url)
            original_text = []
            paragraphs = text.find_all('p')
            for paragraph in paragraphs:
                if len(paragraph.getText()) != 0 \
                        and paragraph.getText() != ' ' \
                        and paragraph.getText() != '\xa0':
                    original_text.append(paragraph.getText())
            text_list.append(original_text)
    print('Se han procesado las 22 noticias')
    return text_list
```

3.- MINERÍA DE TEXTO

3.1.- Técnicas empleadas

Tras analizar el código proporcionado, y tal como se comenta en el enunciado de la práctica, se observa que el procesamiento inicial que se ha hecho de los textos es muy sencillo. Por ello, se han aplicado las siguientes técnicas:

- Traducción de textos

De las veintidós noticias con las que tenemos que trabajar, seis de ellas están en español y dieciséis en inglés. Por tanto, se decide traducir al inglés los textos en español. Se utiliza la librería TextBLob y se implementa en el código tal y como se muestra a continuación:

```
def translate_news(list_text):
    print('------------TRADUCCIÓN------------------')
    print('Comienza la traducción de los textos que están en español')
    translated_texts = []
    for text in list_text:
        trans_text = []
        for elem in text:
            t = TextBlob(elem)
            language = t.detect_language()
            if language == 'es':
                new_en = t.translate(to='en')
                trans_text.append(str(new_en))
            else:
                trans_text.append(elem)
        translated_texts.append(trans_text)
    print('Se ha completado la traducción de textos')
    return translated_texts
```

- Eliminación de stopwords

Cuando hablamos de stopwords o "palabras vacías" nos referimos a todas esas palabras que no tienen significado por sí solas. Podemos pensar que estas palabras no van a aportar nada a nuestra clasificación, por lo que probaremos a eliminarlas. La siguiente imagen muestra el método definido para la eliminación de las stopwords.

```
def remove_stopwords(sentence):
    non_stop_sentence = []
    for word in sentence:
        posible = word.lower()
        if posible not in stop:
            non_stop_sentence.append(word)
    return non_stop_sentence
```

- Signos de puntuación

Es importante analizar la relevancia que pueden tener los signos de puntuación en nuestro análisis. Es posible que la eliminación de estos nos lleve a modificar nombres propios, nombres de organizaciones o posibles palabras compuestas.

```
def remove_punctuation(words):
    filter_sentence = []
    for word in words:
        if word not in string.punctuation:
            filter_sentence.append(word)
    return filter_sentence
```

- Entidades nombradas

Tratamos de buscar y clasificar en función de categorías predefinidas (personas, lugares, organizaciones) las entidades que encontramos en los textos. Definimos el siguiente método en nuestro código para la extracción de éstas.

```
def extract_entity_names (sentence):
    entity_names = []
    #Se comprueba que el token tenga etiqueta
    if hasattr(sentence,'label') and sentence.label:
        #print(sentence.label)
        #Si es una entity, entonces lo agregamos con los que ya hemos identificado
        if sentence.label() =='GPE':
            entity_names.append(' '.join([child[0] for child in sentence]))
        # En caso contrario, obtenemos todos los hijos del token
        else:
            for child in sentence:
                entity_names.extend(extract_entity_names(child))
    return entity_names
```

La técnica que emplearemos para la identificación de las entidades nombradas será el "chunking", que asigna una estructura parcial a una sentencia. Vemos así que, en nuestros textos, aparecen tres tipos de Entidades Nombradas, que son "PERSON", "ORGANIZATION" y "GPE" (geopolítico).

3.2 - Otras técnicas

- Stemming y lematización

Stemming realiza una agrupación de las palabras derivadas de una raíz común, mientras que la lematización obtiene directamente la raíz de las palabras.

En nuestro caso, las noticias son todas de carácter político, y con el análisis de las entidades nombradas vemos que hay muchos nombres propios, por lo que consideramos que la aplicación de estos métodos no nos va a aportar información de interés.

4.- RESULTADOS

A la hora de obtener el rand score óptimo para nuestro clúster, dividimos las pruebas realizadas en dos grupos, con los textos originales, tal y como se obtienen de la noticia, y con los textos en español traducidos al inglés. De esta forma, podemos ir comparando los resultados obtenidos y decidimos así cuál es la combinación óptima.

A su vez, para cada caso, se han evaluado los resultados aplicando distancia coseno y distancia euclídea.

Por otro lado, a la hora de asignarle un peso a cada término del vocabulario hemos utilizado únicamente la medida TF, ya que con ella hemos conseguido obtener un ARI = 1.

4.1.- Resultados con textos originales.

1.-Distancia coseno

En primer lugar, tomamos como dato partida el ARI que obtenemos con los textos originales sin traducción alguna y sin eliminar nada.

```
Created a collection of 633 terms.
Unique terms found:  528
Vectors created.
test:  [1 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 3 2 4]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.0294820717131
```

Podemos observa que el ARI obtenido es muy bajo, por lo que procedemos a aplicar las distintas técnicas de minería de texto.

Signos de puntuación

Eliminamos los signos de puntuación de nuestros textos y comprobamos cómo el ARI aumenta ligeramente con respecto a lo que teníamos inicialmente:

```
Created a collection of 19919 terms.
Unique terms found:  4753
Vectors created.
test:  [0 2 2 4 0 0 0 0 4 4 1 1 1 4 0 0 4 2 0 3 0 4]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.0792792792793
```

Stopwords

Comprobamos si las palabras vacías aportan información relevante para realizar el clúster. Lo hacemos manteniendo los signos de puntuación en los textos:

El ARI ha subido de nuevo en este caso, por lo que nos parece lógico realizar ahora el análisis eliminado las stopwords y los signos de puntuación.

Hemos subido de un 0.029 inicial a un 0.151 tras haber eliminado estos dos elementos.

Entidades Nombradas

Como ya se comentó con anterioridad, las entidades nombradas que aparecen en los textos son de tipo “PERSON”, “ORGANIZATION” y “GPE”. Realizamos distintas combinaciones de las entidades con stopwords y signos de puntuación.

Si eliminamos los signos de puntuación y las stopwords y nos quedamos con los tres tipos de entidades nombradas citados:

Se observa un claro aumento en el ARI, hasta el 0.466, lo que ya nos induce a pensar que las entidades nombradas van a tener gran importancia a la hora de obtener un rand score alto.

Realizamos también la comprobación sin eliminar las stopwords ni los signos de puntuación, pero quedándonos con las tres entidades nombradas. Es curioso ver cómo en este caso sube el ARI, lo que nos indica que las stopwords y los signos son elementos importantes a la hora de analizar el texto y no debemos eliminarlos, pues podrían formar parte de los nombres de las organizaciones.

```
Created a collection of 1389 terms.
Unique terms found:  573
Vectors created.
test:  [2 4 2 2 2 3 3 3 1 1 4 4 4 4 1 0 0 0 0 2 0 3 1]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.58533145275
```

Es por ello que las siguientes pruebas se realizarán ya sin eliminar ambos. Se muestra a continuación un cuadro resumen con las distintas combinaciones realizadas.

FILTROS	STOPWORDS	PUNTUACIÓN	CHUNKING	ENTIDAD NOMBRADA SELECCIONADA			TERMS IN COLLECTION	UNIQUE TERMS	RAND SCORE
				PERSON	ORGANIZATION	GPE			
COSENO	NO	NO	NO	NO	NO	NO	633	528	0,02948207
	NO	SI	NO	NO	NO	NO	19919	4753	0,07927928
	SÍ	NO	NO	NO	NO	NO	17079	4588	0,08125
	SÍ	SÍ	NO	NO	NO	NO	13578	4566	0,15109344
	SÍ	SÍ	SÍ	SÍ	SÍ	SÍ	1222	558	0,46632783
	NO	NO	SÍ	SÍ	SÍ	SÍ	1389	573	0,58533145
	NO	NO	SÍ	SÍ	SÍ	NO	804	388	0,57931781
	NO	NO	SÍ	SÍ	NO	SÍ	1108	446	0,6918239
	NO	NO	SÍ	NO	SÍ	SÍ	866	346	0,58533314
	NO	NO	SÍ	SÍ	NO	NO	523	257	0,48339483
	NO	NO	SÍ	NO	NO	SÍ	585	213	0,58533145

Se extraen las entidades nombradas por parejas, y puesto que la combinación "GPE" y "PERSON" es la que nos da un mayor un resultado, se analizan éstas individualmente. De esta forma, el ARI desciende.

Con los textos originales, aplicando distancia coseno, no se obtiene un resultado que supere el 0.70, por lo que pasamos a realizar el análisis aplicando la distancia euclídea.

2.-Distancia euclídea

Al igual que con la distancia coseno, partimos del resultado que nos dan los textos sin aplicar ningún método, y obtenemos un ARI negativo.

```
Created a collection of 633 terms.
Unique terms found:  528
Vectors created.
test:  [0 0 0 3 0 0 0 0 0 4 0 0 0 1 0 0 2 0 0 0 0 0]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  -0.0453141091658
```

El cuadro siguiente muestra las combinaciones realizadas para la distancia euclídea. El número de combinaciones es inferior al de la distancia coseno. Esto se debe a que según íbamos ejecutando el código con las

distintas combinaciones, comprobábamos que el ARI era igual o inferior al otro. Por ello no ejecutamos más, pues el método es claramente peor.

FILTROS	STOPWORDS	PUNTUACIÓN	CHUNKING	ENTIDAD NOMBRADA SELECCIONADA			TERMS IN COLLECTION	UNIQUE TERMS	RAND SCORE
				PERSON	ORGANIZATION	GPE			
EUCLIDEA	NO	NO	NO	NO	NO	NO	633	528	-0,04531411
	SÍ	SÍ	NO	NO	NO	NO	13578	4566	0,0792792
	NO	NO	SÍ	SÍ	NO	SÍ	1108	446	0,17226277
	NO	NO	SÍ	NO	NO	SÍ	585	213	0,07609988

4.2.- Resultados con textos traducidos.

De las 22 noticias, sólo 6 de ellas están en español por lo que decidimos traducir éstas al inglés. Al igual que con los textos originales, aplicamos *distancia coseno* y *distancia euclídea*.

1.-Distancia coseno

El orden en el que se han realizado las pruebas es el mismo que para los textos originales. Por ello, partimos de nuevo del resultado base obtenido con la traducción, pero sin aplicar ningún otro método.

```
Created a collection of 633 terms.
Unique terms found:  528
Vectors created.
test:  [1 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 3 2 4]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.0294820717131
```

El ARI inicial coincide con el de los textos sin traducir, lo que en un principio nos podría indicar que la traducción no es necesaria. Es preciso realizar más pruebas para comprobar que esto no es así.

Signos de puntuación

Al contrario de lo que pasaba con los textos originales, si eliminamos únicamente los signos de puntuación, el ARI en este caso baja, por lo que es evidente que la traducción de los textos sí va a influir en nuestros resultados

```
Created a collection of 19830 terms.
Unique terms found:  4037
Vectors created.
test:  [4 3 3 0 0 1 1 0 0 0 0 0 0 0 0 0 0 3 0 2 0 0]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.0222127296027
```

Stopwords

Para el caso en el que mantenemos los signos de puntuación, pero eliminamos las stopwords, tenemos que el ARI sube con respecto a los dos casos anteriores.

```
Created a collection of 15691 terms.
Unique terms found:  3867
Vectors created.
test:  [0 4 4 0 0 0 0 0 0 1 1 1 1 1 3 0 0 4 0 2 0 1]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.27122263748
```

Se realiza también la prueba eliminando stopwords y signos de puntuación, y el ARI aumenta considerablemente.

```
Created a collection of 12178 terms.
Unique terms found:  3845
Vectors created.
test:  [0 3 3 0 0 0 0 0 0 4 2 2 2 2 2 1 1 1 3 0 1 0 2]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.44875
```

Entidades Nombradas

La primera prueba que se hace con las entidades nombradas es aquella en la que se eliminan los signos de puntuación, las stopwords y se seleccionan los tres tipos de entidades "PERSON", "ORGANIZATION" y "GPE". El ARI obtenido (0,8372093) es el mejor hasta el momento.

```
Created a collection of 1158 terms.
Unique terms found:  501
Vectors created.
test:  [0 2 0 0 0 1 1 1 0 2 2 2 2 2 4 4 4 4 0 3 1 2]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.837209302326
```

A continuación, mantenemos las stopwords y la puntuación y seleccionamos las tres entidades, obteniendo el mismo resultado que para el caso anterior, aunque el número de términos de la colección y los términos únicos es superior:

```
Created a collection of 1347 terms.
Unique terms found:  525
Vectors created.
test:  [0 2 0 0 0 1 1 1 0 2 2 2 2 2 3 3 3 3 0 4 1 2]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  0.837209302326
```

Al igual que con los textos originales, se realizan distintas combinaciones de las entidades, primero por parejas y después de forma individual. El siguiente cuadro es un resumen de los resultados obtenidos:

FILTROS	STOPWORDS	PUNTUACIÓN	CHUNKING	ENTIDAD NOMBRADA SELECCIONADA			TERMS IN COLLECTION	UNIQUE TERMS	RAND SCORE
				PERSON	ORGANIZATION	GPE			
COSENO	NO	NO	NO	NO	NO	NO	633	528	0,02948207
	NO	SI	NO	NO	NO	NO	19830	4037	0,02221273
	SÍ	NO	NO	NO	NO	NO	15691	3867	0,27122263
	SÍ	SÍ	NO	NO	NO	NO	12178	3845	0,44875
	SÍ	SÍ	SÍ	SÍ	SÍ	SÍ	1158	501	0,8372093
	NO	NO	SÍ	SÍ	SÍ	SÍ	1347	525	0,8372093
	NO	NO	SÍ	SÍ	SÍ	NO	771	368	0,58607095
	NO	NO	SÍ	SÍ	NO	SÍ	1051	391	0,8372093
	NO	NO	SÍ	NO	SÍ	SÍ	872	321	0,8159136
	NO	NO	SÍ	SÍ	NO	NO	475	228	0,55847568
	NO	NO	SÍ	NO	NO	SÍ	576	181	**1**

Puede verse en el cuadro que la mejor combinación de todas es aquella en la que no se eliminan los stopwords ni los signos de puntuación, y nos quedamos únicamente con las entidades de tipo "GPE". Con esto conseguimos el **ARI = 1**, por lo que nuestro clúster es exactamente igual al de la referencia del código proporcionado.

```
Created a collection of 576 terms.
Unique terms found:  181
Vectors created.
test:  [0 4 0 0 0 2 2 2 1 4 4 4 4 4 4 3 3 3 3 1 0 2 4]
reference:  [0, 5, 0, 0, 0, 2, 2, 2, 3, 5, 5, 5, 5, 5, 4, 4, 4, 4, 3, 0, 2, 5]
rand_score:  1.0
```

No se realizan, por tanto, más pruebas con la distancia coseno. Sí se realizan unas pruebas con la distancia euclídea para comprobar si sucede como en el caso de los textos originales, esto es, que el ARI obtenido es claramente inferior al de la distancia coseno.

2.-Distancia euclídea

En el cuadro de resultados siguiente, se observa que sí ocurre como con los textos sin traducir, es decir, el ARI es inferior al de la distancia coseno.

FILTROS	STOPWORDS	PUNTUACIÓN	CHUNKING	ENTIDAD NOMBRADA SELECCIONADA			TERMS IN COLLECTION	UNIQUE TERMS	RAND SCORE
				PERSON	ORGANIZATION	GPE			
EUCLIDEA	NO	NO	NO	NO	NO	NO	633	528	-0,04531411
	SÍ	SÍ	NO	NO	NO	NO	12178	3845	-0,0237812
	NO	NO	SÍ	SÍ	NO	SÍ	1051	391	0,6440677
	NO	NO	SÍ	NO	NO	SÍ	576	181	0,6024096

5.- CONCLUSIONES

Se comprueba en el apartado anterior que la mejor combinación es aquella en la que se mantienen los signos de puntuación y los stopwords, y la entidad seleccionada es "GPE", con la que obtenemos un ARI = 1.

Estamos trabajando con 22 de noticias de carácter claramente político, por lo que es lógico que la entidad "GPE" nos proporcione la mejor aproximación. Asimismo, el hecho de mantener las stopwords y la puntuación es indicativo de que los nombres propios de las organizaciones presentes en los textos pueden tener '- , .' o palabras como 'the' como parte de su nombre.

Es importante señalar de nuevo que la aplicación de la distancia euclídea es mucho menos efectiva que la distancia coseno.

Se utiliza la medida TF para asignar pesos a los términos del vocabulario y, a la vista de los resultados, decidimos no aplicar TF-IDF pues ya habíamos conseguido nuestro objetivo.