Práctica 1. Herramientas de análisis del corpus. Extracción de terminología

Objetivos

Esta práctica tiene como objetivo fundamental la utilización de una herramienta de análisis del corpus para hacer una extracción de terminología.

Materiales

Para hacer esta práctica se propone usar un corpus formado por una serie de documentos de la Wikipedia que tratan acerca del cambio climático. Dichos artículos son los siguientes:

- Cambio climático: https://es.wikipedia.org/wiki/Cambio_clim%C3%A1tico.
- Calentamiento global: https://es.wikipedia.org/wiki/Calentamiento_global.
- Gas de efecto invernadero: https://es.wikipedia.org/wiki/Gas_de_efecto_invernadero.
- Huella de carbono: https://es.wikipedia.org/wiki/Huella_de_carbono.
- Protocolo de Kioto: https://es.wikipedia.org/wiki/Protocolo_de_Kioto_sobre_el_cambio_clim%C3%A1tico.
- Impuesto sobre el carbono: https://es.wikipedia.org/wiki/Impuesto_sobre_el_carbono.
- Combustible fósil: https://es.wikipedia.org/wiki/Combustible_f%C3%B3sil.
- Contaminación: https://es.wikipedia.org/wiki/Contaminaci%C3%B3n.

Se utilizará la siguiente herramienta de análisis del corpus: KWIC Concordance for Windows v5.3 (http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html). Esta herramienta nos permite obtener la lista de frecuencias de las palabras, las concordancias y las colocaciones.

Se hará uso puntualmente en esta práctica de una lista de palabras a ignorar (StopList): https://meta.wikimedia.org/wiki/Stop_word_list/google_stop_word_list.

También se utilizará una herramienta de extracción de términos automática llamada TermoStat (http://termostat.ling.umontreal.ca/).

Procedimiento

El primer paso consiste en añadir a la herramienta la información que conforma el corpus. Los artículos de la Wikipedia deben guardarse en forma de ficheros de texto y combinarse en un único fichero con extensión TXT. Una vez lo hayamos hecho, seleccionamos la pestaña “Corpus Files” y hacemos clic en “Add”. A continuación seleccionamos nuestro fichero.

Actividad 1. Lista de frecuencia

Vamos a obtener la lista de frecuencia de palabras del texto propuesto. Para ello ejecutaremos la opción de menú “WordList”. Esto nos genera una tabla con la información que queremos. Tengamos en cuenta que si utilizamos varios ficheros, KWIC nos genera una tabla por cada documento del corpus. Por ello es conveniente unir todos los documentos en uno, de lo contrario habría que hacer manualmente la suma de los resultados de frecuencia de palabras.

La tabla resultado tiene dos columnas: “Word” y “Count”, que representan el término y el número de veces que aparece en el texto, respectivamente. Podemos ordenar la tabla de distintas formas, de acuerdo a la manera en que queramos visualizar la información.

Se deben entregar dos tablas. Cada tabla incluirá las diez palabras más frecuentes sin utilizar una StopList y lo mismo utilizando una StopList. Recordemos que la StopList es la lista de términos que no se incluirán en el análisis de terminología por ser artículos, conjunciones o preposiciones.

Nuestra herramienta permite el uso de StopList. Para utilizar una, iremos a la pestaña “Corpus Settings” y haremos clic en “Stop Word”. Seleccionaremos el fichero de palabras a ignorar en el análisis y marcaremos la casilla “Activate”. La StopList que nos corresponda (inglés o francés) la obtendremos del enlace proporcionado al principio de este documento de prácticas, y la guardaremos en formato texto plano (TXT) para su uso con KWIC.

Palabra	Frecuencia

Tabla 1. Modelo de tabla para la actividad 1

Actividad 2. Estudio de concordancia

Una vez realizado el estudio de la lista de frecuencia de palabras, pasaremos a realizar un estudio de concordancia monolingüe para identificar el contexto en el que se usan las palabras y detectar posibles términos formados por más de una palabra. Para hacerlo, utilizaremos la opción de menú “Concordance”, que nos devolverá una tabla.

Buscaremos las concordancias ordenadas por la derecha y por la izquierda de las **seis** palabras más frecuentes de la lista obtenida en la actividad anterior, con el uso de la StopList. Una vez encontrados estos términos los insertaremos en una tabla, tal y como se muestra en el modelo de tabla a utilizar. Como son seis palabras a analizar, se entregarán seis tablas.

Línea	Izquierda	Palabra	Derecha

Tabla 2. Modelo de tabla para la actividad 2, concordancias

Además se incluirá otra tabla por cada palabra que mostrará la lista de términos compuestos que se identifiquen.

Palabra compuesta	Frecuencia

Tabla 3. Modelo de tabla para la actividad 2, lista de términos compuestos

Actividad 3. Colocaciones lingüísticas

En el punto tres haremos un estudio de colocaciones léxicas para las palabras “climate” y “pollution” (inglés) o “climatique” y “pollution” (francés). Para ello utilizaremos la opción de menú “Collocate”. Indicaremos las cinco filas con mayor frecuencia total, para cada término.

Palabra	Frecuencia	L5	L4	L3	L2	L1	Nodo	R1	R2	R3	R4	R5

Tabla 4. Modelo de tabla para la actividad 3

Actividad 4. Extracción de términos con TermoStat

A continuación se va a utilizar una herramienta automática de extracción de términos para analizar el corpus con el que estamos trabajando en esta práctica. Lo primero que tendremos que hacer es registrarnos en la página web de los desarrolladores de esta aplicación para obtener las credenciales de acceso.

El funcionamiento de la herramienta es muy sencillo. Primero seleccionaremos la opción “Seleccionar archivo” en el apartado “File” y le pasaremos el fichero correspondiente al corpus. En “Language” pondremos la opción adecuada (inglés o francés) a fin de obtener buenos resultados. Por último seleccionamos “Analyze” para lanzar el análisis.

Rellenaremos una tabla con los 25 primeros resultados de la tabla que nos aparece en la pestaña “List of terms”, los cuales están ordenados por la columna “Score”. Haremos lo mismo con los 25 primeros resultados que nos aparecen en la pestaña “Structuration”. Los modelos de tabla se muestran a continuación. Una vez tengamos las listas anteriores, compararemos los términos obtenidos con los ya extraídos anteriormente de manera manual.

Candidato	Frecuencia	Puntuación (especificidad)	Variantes	Patrón

Tabla 5. Modelo de tabla para la actividad 4, lista de términos

Candidato	Frecuencia	Inclusión del término

Tabla 6. Modelo de tabla para la actividad 4, estructuración

Entregable

Se debe entregar un breve informe generado mediante un procesador de textos sobre cada una de las actividades descritas anteriormente. Asimismo se debe entregar un fichero en formato Excel con 40 términos extraídos del corpus de entre todos los encontrados por las herramientas, habiendo aplicado la StopList.