Práctica final. II Parte de Minería de Datos y Texto.

MADM. Curso 2021-22

Bernat Amengual Mesquida

El objetivo de esta práctica es el de identificar y crear una ontología de aves, encontrades en entradas de un blog. Vamos a dividir el problema en cuatro partes. En primer lugar, realizaremos un preprocesado del texto, buscaremos y procesaremos las aves de la dbpedia, buscaremos si esas aves se encuentran en el texto procesado y, finalmente, crearemos una ontología con los resultados obtenidos.

Los datos se encuentran dentro de un fichero llamado files, allí dentro se encuentran las 10 entradas del blog dentro de un .txt con el nombre file_x, donde x es el número del archivo. Además hay un fichero de prueba fuera de la carpeta llamado aves.txt, con el cual se han ido realizando las pruebas previas.

Preprocesado del texto

Para el procesado del texto hemos creado una función que se puede encontrar en la chunk numero 2. Para la identificación de los pájaros vamos a quedarnos solamente con nombres, adjetivos y verbos. Los verbos se han dejado por si después se quería hacer una ontología más completa aunque finalmente no se ha hecho. Asumimos que los nombres de pájaros van a contener algunos adjetivos ya que a veces contiene el color o algún rasgo distintivo del pájaro en el nombre. Los determinantes se han eliminado pese a que algún nombre de pájaro los pueda contener, pero cuanto hagamos el siguiente paso también los eliminaremos.

Finalmente, realizamos una tokenización del texto filtrado. El motivo es para eliminar posibles plurales o conjugaciones de dichas palabras. Como en el caso de los determinantes en el siguiente paso también vamos a tokenizar los nombres de los pájaros.

Búsqueda y procesado de aves en la dbpedia

Con el objetivo de crear un diccionario para poder buscar si un ave se encuentra en el texto, vamos a realizar una query a la dbpedia. En este caso nos interesa tanto la label del pájaro como su enlace en la dbpedia.

Para facilitar la identificación y posterior creación de la ontología, vamos a crear un diccionario con la siguiente estructura:

{ bird : (tokenized_bird, dbpedia_url)}

Encontrar aves en el texto

Al haber aplicado la tokenización y el filtrado de palabras a los pájaros y al texto, asumimos que van a tener la misma forma, por tanto, podremos identificarlos fácilmente con el comando *in*.

Definimos una función que se encargue de ello, que posteriormente será usada para crear la ontología.

Creación de la ontología

Las ontologías constan principalmente de dos partes, las entidades y las relaciones entre ellas. Para este caso se han definido 3 tipos de entidad: el pájaro, la url del pájaro en la dbpedia y el archivo de donde se ha extraído el pájaro.

Finalmente hay dos relaciones posibles: url y find_in. Url es el enlace del pájaro a la dbpedia y find_in nos indica en que fichero se ha encontrado el pájaro.

El paso final es iterar sobre todos los ficheros para construir la ontología.

Valoración critica

En este apartado se discuten posibles mejoras o problemas del algoritmo utilizado.

El hecho de tokenizar tanto el texto como los pájaros asume que las palabras van a quedar con la misma forma, aunque no hemos comprobado que realmente sea así. Cuando tenemos una palabra que puede ser derivada de algo relacionado con un nombre de un pájaro y dicha palabra es algo más larga puede que el tokenizado. Una posible solución seria la de crear una función que en lugar de tokenizar, convirtiese cada pájaro en una expresión regular, lo que evitaría estos problemas.

La query saca todos los pájaros de la dbpedia lo que hace que no sea realmente eficiente. Una opción podría ser tratar de sacar la ubicación de donde se publica la entrada (la mayoría de ellas lo especifica claramente) y hacer una búsqueda de pájaros regionales. Aunque esto puede tener una limitación, si el pájaro del que esta hablando es una especie invasora seguramente no lo identificáramos.