PRÁCTICA DE LA ASIGNATURA DE OBTENCIÓN DE DATOS

CURSO 2021-2022

REALIZACIÓN DE LA PRÁCTICA

La práctica se llevará a cabo en grupos de 2 ó 3 personas. Cada grupo se identificará con un nombre que estará formado por el primer apellido de cada miembro del grupo, ordenados alfabéticamente y concatenados entre sí sin espacios ni guiones entre ellos. Un ejemplo para el conjunto de profesores de esta asignatura llamados Almudena Sierra, Jesús Sanchez-Oro y Paloma Cáceres sería: caceressanchez-orosierra

CONTEXTO DE LA PRÁCTICA

El dominio de interés seleccionado para la práctica son datos medioambientales sobre la calidad del aire en la Comunidad de Madrid. Estos datos están definidos a través de mediciones de contaminantes presentes en el aire y mediciones de datos meteorológicos, así como datos sobre las estaciones que realizan las mediciones obtenidas que forman parte de la Red de Calidad del Aire de la Comunidad de Madrid (http://gestiona.madrid.org/azul_internet/html/web/3.htm?ESTADO_MENU=3). Más información: https://www.comunidad.madrid/servicios/salud/calidad-aire-salud.

En este dominio, se pide desarrollar tanto un scraper como un conjunto de datos semánticos. Ambos desarrollos se especifican de forma detallada en las siguientes secciones PARTE 1 y PARTE 2 respectivamente.

PARTE 1: SCRAPER

Acceder a http://gestiona.madrid.org/azul_internet/html/web/2_1.htm?ESTADO_MENU=2_1_1; indicar "Selección por estación (datos de las últimas 24 horas)"; seleccione una estación concreta, a su elección, y desarrolle un scraper de la pantalla resultante; almacenar los datos un fichero csv de nombre, la estación seleccionada.csv (por ejemplo, si se ha seleccionado Móstoles, el fichero se denominará mostoles.csv).

PARTE 2: DATOS SEMÁNTICOS:

En este ámbito de la calidad del aire, la Comunidad de Madrid a través del Catálogo de Datos Abiertos (https://datos.comunidad.madrid/catalogo/) proporciona datos publicados de forma abierta producidos o recopilados por las administraciones públicas a disposición de la ciudadanía para que puedan ser utilizados libremente. Este catálogo incluye datos en varios formatos de diferentes temáticas, proporcionando una descripción sobre los datos, datos de ejemplo y un diccionario de datos con el nombre de cada campo y el tipo de datos asociado. Para esta práctica los ficheros de datos fuente seleccionados sobre la calidad del aire han sido los siguientes:

- **calidad_aire_datos_meteo_mes.csv**: Datos meteorológicos horarios del mes en curso recogidos por las estaciones de medición de la Red de Calidad del Aire de la Comunidad de Madrid: valores medios horarios de velocidad del viento, dirección del viento, temperatura, humedad relativa, presión atmosférica, radiación solar y precipitación. Más información: https://datos.comunidad.madrid/catalogo/dataset/calidad_aire_datos_meteo_mes/resource/3da128fa-fbb7-491d-a316-675d5ecf99c2. El código completo de una estación se

construye con el de provincia (pp) más el de municipio (mmm) más el de estación (eee), como ppmmmeee y es el que servirá como conector con el fichero siguiente.

- **calidad_aire_estaciones.csv**: Conjunto de las estaciones que integran la Red de Calidad del Aire de la Comunidad de Madrid. Más información: https://datos.comunidad.madrid/catalogo/dataset/calidad_aire_estaciones/resource/132cd18b-c81e-45c1-a358-baa930252353. De este fichero nos interesan los datos de nombre de la estación, latitud y longitud (UTM). El código completo de la estación, en este fichero, es ppmmmeee donde pp es el código de la provincia, mmm es el código del municipio y eee el de la estación.

Se pide un diagrama de clases UML que represente el modelo conceptual de los datos.

Y además se pide el conjunto de datos semánticamente anotados (es decir, la anotación semántica de los datos integrados de ambos ficheros), organizándolos en tripletas donde el sujeto sea el "código completo de una estación".

Una ontología que también podéis usar (además de las vistas en clase y que tenéis en las transparencias) es la de *geonames*, cuya documentación está disponible en http://www.geonames.org/ontology/documentation.html. Si necesitáis crear vuestros propios términos podéis crear un listado de términos y entregarlo en un readme.pdf. Luego podréis utilizar dichos términos utilizando el alias od (Obtención de Datos).

Puesto que las estaciones están geográficamente situadas en provincia y municipios, se solicita el enlace de los datos de este ejercicio con los recursos de dbpedia, en la medida en la que esto sea posible. Un ejemplo de enlazar datos sería: buscamos en Google "Getafe dbpedia", encontramos el recurso Getafe en dbpedia con la URI https://dbpedia.org/page/Getafe. Por lo tanto, podremos enlazar las estaciones que estén en el municipio de Getafe con dicha URI y nos quedaría un código similar a este:

```
….
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:od="http:// www.masterdatascience.es/contenidos/obteniendatos#">
…..
<rdf:Description rdf:about= ”
http://gestiona.madrid.org/azul_internet/html/web/DescripcionEstacionAccion.icm?ESTADO_MENU=3_2&idEstacion=1”>
…..
   <od:municipio rdf:resource=”https://dbpedia.org/page/Getafe” />
…..
</rdf:Description>
……
</rdf:RDF>
```

ENTREGA

Se realizará una única entrega por grupo a través del enlace disponible en Aula Virtual. La entrega será un fichero comprimido con el nombre del grupo (ver primera sección).

El fichero contendrá los siguientes seis ficheros:

1. Readme.pdf
2. Código del scraper.
3. Código de generación del conjunto de datos semánticos.
4. Fichero csv con los datos obtenidos (con nombre datosscraper.csv).
5. Diagrama de clases UML en formato pdf (con nombre diagramaUML.pdf).
6. Fichero con el conjunto de datos semánticos (con nombre datossemanticos.rdf).

FECHA DE ENTREGA

La fecha límite de entrega será el día 13 de enero a las 23:55.