

OBTENCIÓN DE DATOS

Práctica integral de la asignatura
Curso 2016-2017



OBJETIVO DE LA PRÁCTICA

La práctica de la asignatura tiene como objetivo proporcionar datos en formato RDF, para su uso posterior en la obtención de rutas accesibles, de los transportes METRO, METRO Ligero y Cercanías RENFE.

RESUMEN DE LA PRÁCTICA

Partiendo de la fuente de datos abiertos del Consorcio Regional de Transportes de Madrid (CRTM) [1], en formato GTFS [2], habrá que obtener los datos de las estaciones (contenidas en el fichero *stops.txt*) de los transportes METRO, METRO Ligero y cercanías RENFE.

Dado que en estos ficheros GTFS, las estaciones no están ordenadas según el itinerario de la línea, ni en ellos aparece la línea a la que pertenece la estación, será necesario obtener dicho orden y dicha pertenencia a partir de los datos de las líneas de cada medio de transporte, accesibles en el Consorcio Regional de Transportes de Madrid [3].

REALIZACIÓN DE LA PRÁCTICA

El CRTM proporciona, en formato abierto, datos de los medios de transporte de la Comunidad de Madrid. Uno de los formatos que ofrece es el de la especificación de los *feeds* de transporte de Google (GTFS). Para obtener estos ficheros GTFS, será necesario desarrollar un *scraper* sobre la página de datos abiertos del consorcio, que descargue los ficheros de Metro, Metro Ligero y Cercanías en su correspondiente [enlace GTFS](#).

Dado que estos ficheros no contienen el orden del itinerario de las líneas, ni las líneas a las que pertenece cada estación, es necesario buscar esta información realizando otro *scraper* en [3]. Será necesario acceder al [enlace Líneas](#) de cada medio de transporte especificado con anterioridad. Una vez allí, habrá que recorrer todas las líneas, accediendo a su contenido, con el fin de obtener la ordenación de las estaciones.

La integración de estas dos fuentes de datos se realizará en un fichero de texto en formato CSV.

Posteriormente, será necesario elaborar un diagrama conceptual con las entidades y las relaciones entre ellas, que soporte la información del dominio de este problema. Finalmente, se llevará a cabo la creación de un grafo RDF utilizando *rdflib*. El resultado de este último paso será un fichero RDF/XML.

Se pide:

- Código de los *scrapers* desarrollados para obtener la información.
- Fichero de texto, en formato CSV, con información del medio de transporte, la línea y el conjunto de estaciones en el orden marcado por el itinerario de cada línea. La cabecera de dicho fichero debe incluir los siguientes campos (en cursiva los propios de GTFS):

transportmean_name,line_number,order_number,stop_id,stop_code,stop_name,stop_desc,stop_lat,stop_lon,zone_id,stop_url,location_type,parent_station,stop_timezone,wheelchair_boarding

- *transportmean_name*: indicará el nombre del medio de transporte. Los valores posibles son: METRO, ML, CR.
- *line_number*: indicará el número de la línea a la que pertenece la estación.
- *order_number*: indicará el número de la línea concatenado con el número de secuencia de la estación en la línea, según el itinerario.

OBTENCIÓN DE DATOS

Práctica integral de la asignatura
Curso 2016-2017

- Diagrama conceptual del dominio del problema.
- Fichero RDF, en formato RDF/XML, con la información procedente del fichero anterior.
- Código de creación del fichero RDF.

FECHA DE ENTREGA DE LA PRÁCTICA

La fecha tope de entrega de la práctica será el día 9 de enero de 2017 a través del Aula Virtual, en un enlace disponible para dicha entrega.

FORMATO DE ENTREGA DE LA PRÁCTICA

La entrega será un fichero ZIP con el nombre “inicial/es del nombre de pila del alumno” + “primer apellido” (si sois dos alumnos, concatenad de la misma forma el nombre del siguiente alumno). El ZIP deberá tener el siguiente contenido:

- Código fuente en Python de los scrapers desarrollados.
- Fichero CSV.
- Fichero RDF/XML.
- Memoria descriptiva del trabajo realizado:

El formato de la memoria es libre pero el contenido mínimo será (a) el diagrama de clases del modelo de dominio, (b) ontologías analizadas, (c) ontologías consideradas (de las analizadas o de nueva creación).

ENLACES DE INTERÉS PARA LA REALIZACIÓN DE LA PRÁCTICA

- [1]. Datos abiertos del Consorcio Regional de Transporte de Madrid (CRTM), <http://datos.crtm.es/>
- [2]. Google Transit Feed Specification (GTFS), <https://developers.google.com/transit/gtfs/reference?hl=es-419>
- [3]. Información de los medios de transporte en el CRTM, <http://www.crtm.es/tu-transporte-publico.aspx>