Práctica 1: Web scraping

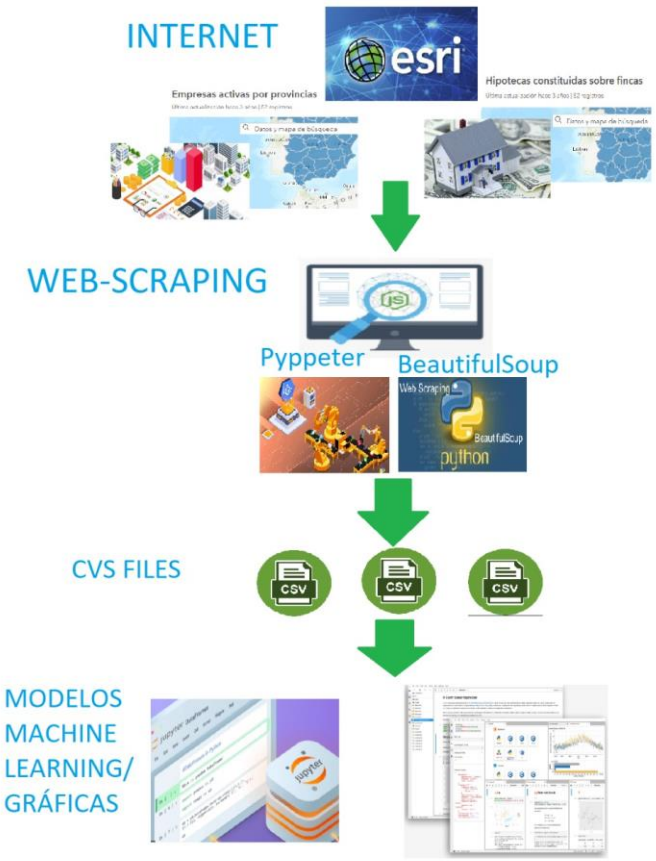
Miembros del equipo

La actividad ha sido realizada de manera grupal por Juan Pablo Uphoff Salas y Miguel Alejandro Ponce.

Descripción

La información se ha seleccionado en un contexto de análisis de recursos información económica e hipotecas de empresas por provincia en España. El sitio web es un recurso de open data para el análisis de información de diversos contenidos de datos abiertos con geo-referenciación para que se puedan consultar y analizar. Se elige el sitio web debido a que presentan diversos conjuntos de datos y posibles relacionados por tanto es posible realizar cruces de información y generar nuevos sub-conjuntos con mayores características.

Esquema



Detalle solución y complejidad

Se presenta una solución que permite la extracción de contenido web dinámico, en donde la presentación del contenido corresponde a cargas parciales de secciones cargadas de manera asíncrona. Se hace uso de programación multihilo y procesamiento de tareas para optimizar los

tiempos de respuesta diferenciando entre los procesos de extracción de contenido web y la generación de los documentos csv. En este caso se generan inicialmente varios ficheros luego de extraer el contenido de las diferentes pestañas de una página. Luego, se realizará un procesamiento de tratamiento de los datos e integración de la fuentes de información.

Dataset

En este caso debido a la versatilidad de la implementación del código fuente es posible utilizar el mismo código para extraer información de diferentes repositorios del portal. En este caso se realizará la integración de dos fuentes tenemos:

- Empresas activas: Este servicio muestra las empresas activas por provincia y condición jurídica (2017-1999).
- Hipotecas: Este servicio muestra las Hipotecas constituidas sobre el total de fincas por provincias y por meses (2003-2018).

Datos sin tratamiento:

Las listas de los archivos cvs con los datos sin tratamiento se encuentran en la capeta *src/data*, en donde el índice de en el nombre se remplaza x por el índice de la página tabulada:

- empresas-activas_x.csv
- hipotecas-constituidas_x.csv

Datos procesados:

Las listas de los archivos cvs con los datos sin tratamiento se encuentran en la capeta *csv*. En este caso se presentan dos conjuntos de datos luego del tratamiento el primero corresponde a la unión de los datos de las empresas e hipotecas, en donde se busca conservar los datos de información geográficas y los índices comunes que los relaciones. El segundo es un sub-conjunto de datos del cual se extrae únicamente aquellos tipos que tienen hipotecas reportadas (están son registradas a partir del año 2012).

- empresas_hipotecas_españa.csv
- empresas_hipotecas_españa_2012_2017.csv

Dataset:

Los campos que incluye el dataset son los siguientes:

- *total*: número total de (importe de hipotecas o número de hipotecas)
- *texto*: provincias del estado español
- *cod*: código de identificación para cada una de las provincias (“texto”)
- *anio*: año.
- *shape_Area*: indentificador geográfico I
- *shape_Length*: identificador geográfico II
- *TipoEmpresa*: tipo de empresa según su forma jurídica.

- *TipoHipoteca*: toma los valores “importe_hipotecas”; importe total de hipotecas por provincias y por meses, en miles de euros, o “numero_hipotecas”; número total de hipotecas constituidas sobre el total de fincas por provincias y por meses.
- *M01*,…,*M12*: meses, donde M01 refiere a enero y M12 a diciembre.

Los datos recogidos incluyen información desde enero de 2003 hasta 2018, aunque como se ha citado con anterioridad, no siempre se reportan datos de las hipotecas . Han sido recogidos mediante web scraping, a través de la librería BeautifulSoup de Python.

Propietario de los datos, análisis previos y posibles estudios.

El portal web www.opendata.esri.es, operado por Esri España, pone a disposición de cualquier usuario multitud de contenidos de acceso libre, los cuales se pueden consultar, analizar o descargar. Además de ofrecer sus propios datos, recopila y ofrece al público diversos datasets de otros portales de acceso libre. En nuestro caso, los datos son del Instituto Nacional de Estadística Español (INE).

Tanto el dataset sobre empresas activas como el de hipotecas constituidas sobre el total de fincas, han sido utilizados en distintos estudios, tanto en el ámbito académico como fuera de él. Antón y Matarazzo (2015) estudian la relación existente entre la crisis económica en España, utilizando las hipotecas como proxy con el que medir el desarrollo económico, y la vuelta a Ecuador de aquellos que emigraron a España en la década de los dos mil. Por otro lado, los datos de empresas activas son a menudo empleados en informes y estudios sobre la coyuntura económica española. Un ejemplo son los boletines del Banco de España, en cuyo último número (enero de 2021) se analiza el impacto de la crisis causada por la COVID-19 en el tejido empresarial español, con el resultado obvio de una tremenda disminución de las empresas activas.

En nuestro caso, hemos considerado que el análisis conjunto de dos indicadores concluyentes de la marcha de la economía española, puede servir para potenciar y optimizar los análisis previos. Por lo general, los estos estudios no tratan simultáneamente la parte de la demanda y la de la oferta, por lo que nuestro nuevo conjunto de datos, al incluir variables elementales de la demanda privada (hipotecas) y de la oferta, utilizando el número de empresas como proxy para la oferta de bienes y servicios, puede ser de gran utilidad para el análisis de la situación de la economía. Especialmente interesante será cuando los datos de 2020 estén disponibles, pues incluye los datos de la pandemia. Un posible caso de estudio podría ser un análisis de la deslocalización de las hipotecas y las empresas, pues son muchos los ciudadanos que han trasladado su residencia –y en algunos casos, también sus empresas- de grandes centros urbanos a pequeños municipios. Algunas de las tesis a responder pueden ser las siguientes: ¿En qué provincias ha aumentado más la compra de viviendas? ¿Hay relación entre compra de vivienda y aumento de empresas activas en tal región? ¿Qué tipo de empresa se constituye más?

Licencia

Nuestros datos estarán bajo la licencia CC0. Bajo esta licencia, se liberan los derechos de propiedad intelectual y cualquier usuario puede hacer un uso libre de ella. Somos firmes defensores del software libre y del código abierto, pues consideramos que en un mundo tecnológico cada vez más

monopolizado por grandes empresas, ofrecer un trabajo que puede beneficiar a cualquier persona del mundo, es un paso hacia la libertad individual y colectiva.

Ficheros del código fuente

- src/web-scraper-esri.py punto de entrada al programa. Inicia el proceso de scraping. Contiene la implementación multihilo para extraer un conjunto de datos inicala a partir de la base de las bases de datos online [Hipotecas(2003-2018)](https://opendata.esri.es/datasets/hipotecas-constituidas-sobre-fincas/data?geometry=-36.558%2C29.677%2C22.725%2C42.089) y [Empresas(2017-1999)](https://opendata.esri.es/datasets/empresas-activas-por-provincias/data?geometry=-36.558%2C29.677%2C22.725%2C42.089).
- src/mdl-procesar-csv Notebook con los métodos para realizar el procesamiento y tratamiento de datos.
- src/mdl-graficos-csv.ipynb Notebook con las representaciones gráficas de los datos procesados.
- src/mdl-regresion-csv.ipynb Notebook con un ejemplo de regresión aplicada al conjunto de datos.
- src/data: contiene los datos extraídos de las fuentes sin tratamiento.
- csv/empresas_hipotecas_españa.csv: Conjunto de datos que representa la unión de los datos de las empresas e hipotecas.
- csv/empresas_hipotecas_españa_2012_2017.csv Conjunto de datos del cual se extrae únicamente aquellos tipos que tienen hipotecas reportadas (están son registradas a partir del año 2012).
- csv/empresas_hipotecas_españa_2012_2017_importe_hipotecas Conjunto de datos del cual se extrae únicamente aquellos tipos que tienen hipotecas reportadas (están son registradas a partir del año 2012) y son del tipo importe hipotecas.
- empresas_hipotecas_españa_2012_2017_numero_hipotecas.csv Conjunto de datos del cual se extrae únicamente aquellos tipos que tienen hipotecas reportadas (están son registradas a partir del año 2012) y son del tipo número hipotecas.

Recursos

1. Lawson, R. (2015). _Web Scraping with Python_. Packt Publishing Ltd. Chapter 2. Scraping the Data.

2. Mitchel, R. (2015). _Web Scraping with Python: Collecting Data from the Modern Web_. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.

3. Antón Hurtado, Fina, & Matarazzo, Claudio (2015). Invirtiendo la ruta: proce-sos de retorno de los ecuatorianos en España. Universitas,XIII (23), pp. 35-64.

4. Banco de España (2021). Boletín Trimestral de la Economía Española.