Práctica 1 (35 % nota final) Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Para hacer esta práctica tendréis que trabajar en grupos de 3 o 2 personas, o si preferís, también podéis hacerlo de manera individual. Tendréis que entregar un solo fichero con el enlace Github (https://github.co m) donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Podéis mirar estos ejemplos como guía:

- Ejemplo: https://github.com/rafoelhonrado/foodPriceScraper
- Ejemplo complejo: https://github.com/tteguayco/Web-scraping

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes que su tratamiento aportan valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como queries, API y scraping).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos al web. Tenéis que indicar las siguientes características del dataset general:

1. Título del dataset. Poned un título que sea descriptivo.
   Subastas BOE (Boletín Oficial del Estado)

1.

2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.
   Listado de subastas extraido de la página web del Boletín Oficial del Estado.

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente

4. Contexto. ¿Cuál es la materia del conjunto de datos?
   Subastas de bienes e inmuebles realizadas por la agencia tributuaria.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?
   Se han recogido estos datos el día 12/11/2018 de la página https://subastas.boe.es/

Los datos se han extraido realizando web scrapping con una aplicación java que recorre todos los enlaces listados en la búsqueda y extrae la información de las miniaturas encontradas en cada una de las páginas listadas en la búsqueda.

Incluye estos campos según estemos viendo el dataSet completo o el dataSet reducido:

Data set completo:

- InfoGen_Fecha_de_inicio
- InfoGen_Fecha_de_conclusion
- InfoGen_Cantidad_reclamada
- InfoGen_Lotes_Sin_lotes
- InfoGen_Anuncio_BOE
- InfoGen_Valor_subasta
- InfoGen_Tasacion
- InfoGen_Puja_mnima
- InfoGen_Tramos_entre_pujas
- InfoGen_Importe_del_depósito
-
- InfoAutGestor_codigo
- InfoAutGestor_Codigo
- InfoAutGestor_Descripcion
- InfoAutGestor_Dirección
- InfoAutGestor_Telefono
- InfoAutGestor_Fax
- InfoAutGestor_Correo_electrónico
-
- Bienes_Titulo
- Bienes_Descripción
- Bienes_Dirección
- Bienes_Código Postal
- Bienes_Localidad

- Bienes_Provincia
- Bienes_Situación posesoria
- Bienes_Visitable
-
- Acreedor_nombre
- Acreedor_nif
- Acreedor_direccion
- Acreedor_localidad
- Acreedor_provincia
-
- Es_Subaste_Por_Lotes
-
- Lotes_Lote1_Cantidad_reclamada
- Lotes_Lote1_Valor_de_tasación
- Lotes_Lote1_Importe_del_depósito
- Lotes_Lote1_Puja_mínima
- Lotes_Lote1_Tramos_entre_pujas
- Lotes_Lote1_Descripcion
- Lotes_Lote1_Referencia_catastral
- Lotes_Lote1_Dirección
- Lotes_Lote1_Código Postal
- Lotes_Lote1_Localidad
- Lotes_Lote1_Provincia
- Lotes_Lote1_Vivienda_habitual
- Lotes_Lote1_Situación_posesoria
- Lotes_Lote1_Visitable
- Lotes_Lote1_Inscripción_registral
- Lotes_Lote1_Informacion_adicional
-
- Lotes_Lote2_Cantidad_reclamada
- Lotes_Lote2_Valor_de_tasación
- Lotes_Lote2_Importe_del_depósito
- Lotes_Lote2_Puja_mínima
- Lotes_Lote2_Tramos_entre_pujas

- Lotes_Lote2_Descripcion
- Lotes_Lote2_Referencia_catastral
- Lotes_Lote2_Dirección
- Lotes_Lote2_Código Postal
- Lotes_Lote2_Localidad
- Lotes_Lote2_Provincia
- Lotes_Lote2_Vivienda_habitual
- Lotes_Lote2_Situación_posesoria
- Lotes_Lote2_Visitable
- Lotes_Lote2_Inscripción_registral
- Lotes_Lote2_Informacion_adicional
-
- Lotes_Lote3_Cantidad_reclamada
- Lotes_Lote3_Valor_de_tasación
- Lotes_Lote3_Importe_del_depósito
- Lotes_Lote3_Puja_mínima
- Lotes_Lote3_Tramos_entre_pujas
- Lotes_Lote3_Descripcion
- Lotes_Lote3_Referencia_catastral
- Lotes_Lote3_Dirección
- Lotes_Lote3_Código Postal
- Lotes_Lote3_Localidad
- Lotes_Lote3_Provincia
- Lotes_Lote3_Vivienda_habitual
- Lotes_Lote3_Situación_posesoria
- Lotes_Lote3_Visitable
- Lotes_Lote3_Inscripción_registral
- Lotes_Lote3_Informacion_adicional
-
- Lotes_masLotes
-
- Pujas
- Enlace_documentoPujas

Data set reducido incluye estos campos:

- MIN_Enlace
- MIN_titulo
- MIN_Juzgado
- MIN_expediente
- MIN_estado

Funcionamiento:

1. Se navega a la URL base del sitio : https://subastas.boe.es/subastas_ava.php?..
2. Descargar enlaces busqueda/SubastaMiniatura: crea ontologias de tipo subastaMiniatura y las rellena.
3. Exportar a CSV. Crea un data set reducido con estos datos.
4. Abre el CSV y creado y ahora utiliza el campo enlace de este CSV y hace scrapping para constuir el data set completo.
5. Almacena el data set completo en formato CSV.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

   El propietario de los datos es la Agencia Estatal de Administración Tributaria. No he visto ningún estudio o investigación acerca de las subastas realizadas por la agencia tributaria.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

   Podemos ver en que localidades hay más subastas, ordenarlas por precio, por tipo de subasta, por localidad, podemos ver las imágenes asociadas a cada subasta, podemos realizar cálculos estadísticos acerca de las subastas que la agencia lleva acabo.

   Me gustaría responder:

   - ¿En que localidades se producen más subastas?
   - ¿Cuál es el precio máximo, mínimo, medio de las subastas?
   - ¿Cuántas subastas quedan desiertas?

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:
    - Released Under CC0: Public Domain License
    - Released Under CC BY-NC-SA 4.0 License
    - Released Under CC BY-SA 4.0 License
    - Database released under Open Database License, individual contents under Database Contents License
    - Other (specified above)
    - Unknown License

Licencia: CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/legalcode)

He escogido esta licencia debido a que:
- Se puede compartir: Está permitido compartir y redsitribuir el material en cualquier medio o formato.
- Se puede adaptar: Se puede adaptar, mezclar, transformar, construir y utilizar el material para cualquier propóstio, aunque tenga fines comerciales.

Bajo estos términos se protege lo siguiente:
- Atribución: Se debe otorgar el crédito correspondienten, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Se puede otorgar el crédito de cualquier manera razonable, pero no de ninguna forma que sugiera que el licenciante (persona a la que se atribuyen los derechos) está respaldando la obra derivada o su uso.
- “ShareAlike” : Si se remezcla, transforma o construye sobre el material, se deben distribuir las contribuciones bajo la misma licencia que el original.

- Sin restricciones adicionales: No puede aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

El código se puede descargar aquí:
https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS

Se puede consultar la documentación genereada sobre el código en:
https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/javadoc

10. Dataset: Dataset en formato CSV

Los datasets generados se pueden descargar aqui:

https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/dataSet

El HTML generado se puede descargar aquí:

https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/html_descargado

Recursos

Los siguientes recursos son de utilidad por la realización de la PEC: ●
Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

- Los apartados 1, 2, 3 y 4 valen 0,25 puntos cada uno.
- Los apartados 5, 6, 7, 8 valen 1 punto cada uno.
- Los apartados 9 y 10 valen 2,5 puntos cada uno.

Formato y fecha de entrega

Durante la semana del 29 de octubre el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica. En la entrega parcial los estudiantes deberán entregar por correo electrónico (lsubirats@uoc.edu) el enlace al repositorio Github con lo que hayan avanzado.

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace a Github donde haya:

1. Una Wiki donde haya los nombres de los componentes del grupo y una descripción de los ficheros.
   - Ver en https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/Wiki.pdf
2. Un documento Word, Open Office o PDF con las respuestas a las preguntas y los nombres de los componentes del grupo.
   - Ver en https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/respuestas.pdf
3. Una carpeta con el código Python o R generado para obtener los datos.
   - Descargar de https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/
4. El fichero CSV con los datos:
   - Descargar de https://github.com/ovarelag/PRA1_TIPOLOGIA_DATOS/dataSets

Este documento de la entrega final se tiene que entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59 del día 12 de noviembre**. No se aceptarán entregas fuera de plazo.