Universitat Oberta de Catalunya
Tipología y ciclo de vida de los datos

Practica 1 (35% nota final)
Juan Camilo Rivera Palacio
Martin Loizate Sarrionandia

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 2 personas. Tendréis que entregar un solo fichero con el enlace Github (https://github.com) 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 aporta 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.

Contenido

Objetivo

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos de la pagina ballotepedia sobre los requisitos de las votaciones de los Estados Unidos.

El contenido del web scrapping está en:
https://github.com/j-river1/Webscraping-Elecciones-EstadosUnidos

1. Contexto.

La elección presidencial de los Estados Unidos (EEUU) es uno de los eventos más importante del 2020. La importancia de estas elecciones no solo se debe al poder económico y militar de EEUU sino también porque define la situación política, económica y social de muchos países, y en particular los países de Latinoamérica. Además, estas elecciones son atípicas: Por una parte, existe un ambiente altamente polarizado, con una discusión centrada en si es reelecto Donald Trump, un presidente ultra conservador que, ignorando el cambio climático, enarbola la bandera de una América todavía más dominante, frente a Joe Biden que no termina de convencer a sus votantes y además ha sido víctima de la propaganda falsa que los relaciona con el socialismo venezolano y cubano. Por otra parte, se realizarán en medio de una pandemia con más de un millón de muertos, y esta situación añade más incertidumbre en torno a la participación en las elecciones: ¿Participará más gente por el momento delicado que vive la sociedad? ¿La pandemia dificultará la participación? En otros países donde se han celebrado las elecciones bajo la pandemia, se ha visto alterado notablemente la participación.

El objetivo de este trabajo es recopilar toda la información a cerca de los trámites que hay que hacer en cada estado para votar. El sitio web que se usó para la descarga de la información es https://ballotpedia.org/, que es una enciclopedia digital de política y electoral cuyo objetivo es brindar información objetiva y exacta de los procesos políticos en Estados Unidos. Tienen más de cincuenta editores a lo largo del país y es una organización sin ánimo de lucro.

El flujo de trabajo para realizar el *web scraping* está basado en (agregar la citacion) y a continuación se describe la evaluación inicial.

1.1 Archivo robot.

En la [Figura 1] muestra algunos robots que se les permite acceder a todas carpetas, alrededor de 50 robots. Para los demás, hay 30 directorios a los que no podrán acceder. En la figura 2 se muestra algunas de las carpetas con restricción.

```
User-agent: Squider
Disallow: /
User-agent: updown_tester
Disallow: /
User-agent: AhrefsBot
Disallow: /
User-agent: KomodiaBot
Disallow: /
User-agent: Aboundex
Disallow: /
User-agent: Jakarta
Disallow: /
User-agent: 200PleaseBot
Disallow: /
User-agent: 360Spider
Disallow: /
User-agent: SISTRIX
Disallow: /
User-agent: oBot
Disallow: /
```

Figura 1 Archivo de robot de la página ballotpedia

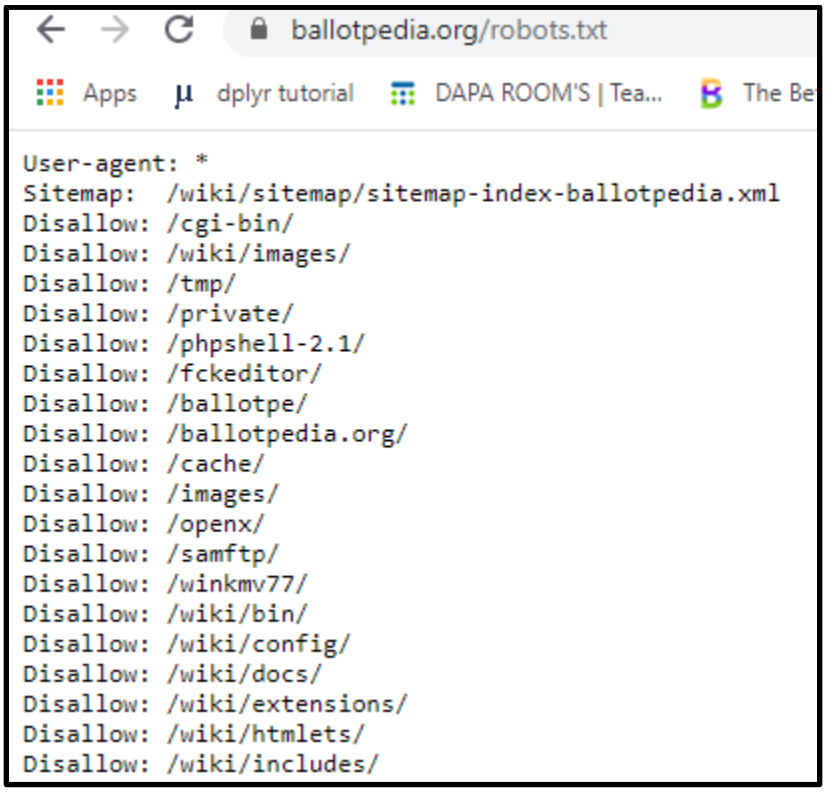

Figura 2 Carpetas con restricción ballotpedia

1.2 Mapa del sitio web.

En la [Figura 3] está un fragmento del mapa del sitio donde está la información de los estados de Alabama, Alaska, Arizona y Arkansas. Además está el link para acceder a la información detallada del estado.

```
<td> <a href="/Voting_in_Alabama" title="Voting in Alabama">Alabama</a></td>
<td>3,609,000</td>
<td>2,490,000</td>
<td>69.0%
</td></tr>
<tr>
<td> <a href="/Voting_in_Alaska" title="Voting in Alaska">Alaska</a></td>
<td>497,000</td>
<td>337,000</td>
<td>67.7%
</td></tr>
<tr>
<td> <a href="/Voting_in_Arizona" title="Voting in Arizona">Arizona</a></td>
<td>4,757,000</td>
<td>3,262,000</td>
<td>68.6%
</td></tr>
<tr>
<td> <a href="/Voting_in_Arkansas" title="Voting in Arkansas">Arkansas</a></td>
<td>2,158,000</td>
<td>1,262,000</td>
<td>58.5%
</td></tr>
```

Figura 3 Ejemplo del mapa del sitio web ballotepedia

1.3 Tamaño.

El tamaño del sitio hace referencia a los enlaces vinculados para el caso de ballotpedia tiene 350.000 enlaces relacionados, como aparece en la [Figura 4].

Figura 4 Tamaño del sitio web

1.4 Tecnología

La tecnología utilizada para esta web se muestra en la [Figura 5]. Cabe la pena mencionar el web frameworks es Twitter Bootstrap y el servidores web es Apache.

Figura 5 Tecnología de ballotepedia

2. Título

Información por estado sobre los requisitos para votar en las elecciones presidenciales de los Estados Unidos.

3. Descripción del dataset.

Las elecciones en EEUU tienen algunas particularidades. Primero, hay que registrarse antes de votar, y existen unos plazos para poder hacerlo que son diferentes en cada estado. Además hay estados donde se puede votar de forma anticipada y en ausencia. En algunos estados se puede votar el mismo día del registro y en otros no, al igual que en algunos se puede registrase el mismo día de las elecciones.

Por todo ello, y el hecho de que EEUU sea un país federal donde cada estado tiene sus propias normas (también en las elecciones) hace que en cada estado sea diferente el proceso de votación. Este dataset recoge las características de cada proceso (registro, verificación, votación, votación a distancia…) en cada uno de los estados.

Las características de cada estado en el proceso de votación a los que hace referencia las columnas del dataset son los siguientes:

4. Representación gráfica.

El flujo del trabajo utilizado el web scrapping [Figura 6]. Empieza por la definición del objetivo, en este caso es recopilar información por estado para las elecciones de los Estados Unidos, luego es escoger el sitio web para la captura de la información, se escogió ballotpedia, que es una página de información donde está toda la información electoral de los Estados Unidos, después se realiza la implementación del código en phyton utilizando varias libreras, como BeatifulSoup para la extracción de las tablas y la creación del dataset y por último la creación del data set que se publica en Zenoodo y Github [Figura 7].

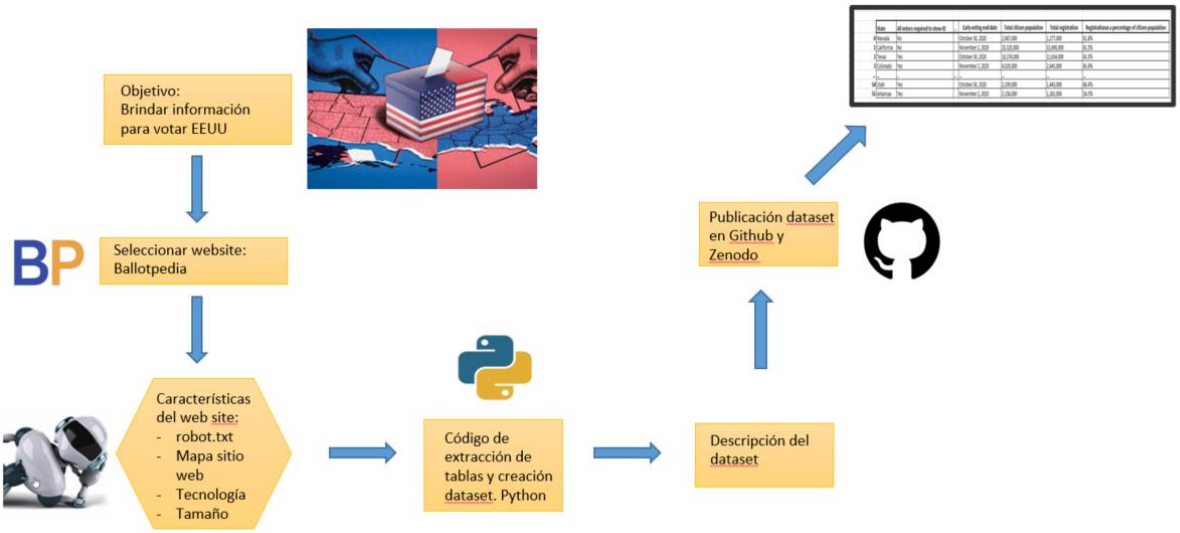

Figura 6 Flujo de trabajo para el webscrapping

	State	All voters required to show ID	...	Early voting end date	Total citizen population	Total registration	Registrationas a percentage of citizen population
0	Nevada	No		October 30, 2020	2,067,000	1,277,000	61.8%
1	California	No		November 2, 2020	25,525,000	15,690,000	61.5%
2	Texas	Yes		October 30, 2020	18,374,000	11,634,000	63.3%
3	Colorado	Yes		November 2, 2020	4,029,000	2,645,000	65.6%
...	...	..	...	...	...	...	...
54	Utah	Yes		October 30, 2020	2,109,000	1,443,000	68.4%
55	Arkansas	Yes		November 2, 2020	2,158,000	1,262,000	58.5%

Figura 7 Imagen del resultado final del dataset

5. Contenido

El dataset está compuesto por 57 filas y 18 columnas. Cada fila hace referencia a cada uno de los estados de EEUU. Cada columna contiene información sobre algún procedimiento para votar en las elecciones presidenciales.

Las características de cada estado en el proceso de votación a los que hace referencia las columnas del dataset son los siguientes:

- **State**. Estado.
- **Total citizen population**. Total de ciudadanos que viven en el estado
- **Total registration**. El registro total de la población que se inscribio en el estado para votar.
- **Registration as a percentage of citizen population**. Porcentaje del total de la poblacion apta para votar.
- **All voters required to show ID**. Todos los votantes deben mostrar una identificación ("Sí" indica que todos los votantes deben presentar una identificación en las urnas).

- **Select voters required to show ID.** Seleccione los votantes que deben mostrar una identificación (una entrada en este campo describe qué votantes deben mostrar una identificación en las urnas si no hay un requisito de identificación universal).
- **Accepted IDs** Identificaciones aceptadas (una lista de formas aceptables de identificación).
- **Registration URL** URL de registro (un enlace al sitio web del estado donde los votantes elegibles pueden registrarse, si corresponde)
- **Registration status URL** URL de estado de registro (un enlace al sitio web del estado donde los votantes pueden verificar su estado de registro, si corresponde).
- **Registration update URL** URL de actualización de registro (un enlace al sitio web del estado donde los votantes pueden actualizar sus registros, si corresponde).
- **In-person registration deadline** Fecha límite de registro en persona (el último día que un votante puede registrarse en persona, si corresponde)..
- **Mail postmark or receipt deadline**. Fecha límite de registro por correo (el último día que un votante puede registrarse por correo, si corresponde).
- **Mail registration deadline**. Matasellos de correo o fecha límite de recepción (ya sea que la fecha límite de registro por correo sea un matasellos o fecha límite de recepción).
- **Online registration deadline** Fechas límite de registro en línea (el último día que un votante puede registrarse en línea, si corresponde).
- **Same-day registration** Registro el mismo día ("Sí" indica que el estado permite que los votantes se registren el día de las elecciones).
- **Early voting same-day registration** Inscripción anticipada el mismo día ("Sí" indica que el estado permite que los votantes se registren durante el período de votación anticipada, si corresponde).
- **Early voting start date** Fecha de inicio de la votación anticipada
- **Early voting end date** Fecha de finalización de la votación anticipada

6. Agradecimientos

Nuestros agradecimientos al equipo de investigadores, escritores y periodistas de Ballotpedia, por recopilar, ordenar y compartir información tan útil en su página web. Ballotpedia es la enciclopedia digital de la política y las elecciones estadounidenses y su objetivo es informar a la gente sobre política proporcionando información precisa y objetiva sobre política en todos los niveles de gobierno, manteniendo la neutralidad en su contenido.

7. Inspiración

Nos parece que este dataset puede ser muy interesante para calcular la facilidad (o dificultad) que existe en cada estado para votar. Partiendo de las características del proceso de votación en cada estado (incluidos en este dataset), se podría construir un algoritmo que calculase la facilidad de voto en cada estado.

Es más fácil votar si te puedes registrar online. Asimismo, es más fácil votar si el periodo de inscripción es más largo. El algoritmo podría ponderar cada una de estas características en cada estado, en base a si facilita el voto o por el contrario lo dificulta, dando una puntuación global a cada estado.

Por ejemplo (estas puntuaciones son ficticias, el objetivo es explicar cómo funcionaría el algoritmo basado en este dataset):

En Alabama es posible registrarse online (+5 puntos), el periodo de inscripción es de 20 días (+5 puntos) y es posible registrarse hasta el mismo día de las elecciones (+1 punto). El estado de Alabama (tomando en cuenta también todas las demás características) sumaría 33 puntos.

En Washington no es posible registrarse online (+0 puntos), el periodo de inscripción es de 25 días (+6 puntos) y no es posible registrarse hasta el mismo día de las elecciones (+0 punto). El estado de Washington (tomando en cuenta también todas las demás características) sumaría 26 puntos.

El algoritmo nos mostraría que es más fácil votar en Alabama que en Washington.

Después, podríamos comparar los resultados obtenidos con la participación en cada estado, para ver si en realidad la facilidad en el proceso de votación influye.

8. Licencia

Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

a. Released Under CC0: Public Domain License
b. Released Under CC BY-NC-SA 4.0 License
c. Released Under CC BY-SA 4.0 License

d. Database released under Open Database License, individual contents under Database Contents License
e. Other (specified above)
f. Unknown License

Para escoger la licencia adecuada para el dataset se debe tener en cuenta los aspectos legales del sitio web de la descarga, en este caso ballotepedia. Este sitio web está construido bajo la licencia GNU Free Documentation License, de acuerdo a wikipe esta licencia le otorga a los lectores los derechos para copiar, redistribuir y modificar (excepto las "secciones invariables") de una obra y requiere que todas las copias y derivados están disponibles bajo la misma licencia. Las copias también se pueden vender comercialmente, pero, si se producen en cantidades mayores (más de 100), el documento original o el código fuente deben estar disponibles para el destinatario de la obra. De acuerdo a lo anterior, se identificaron los siguientes puntos claves:

- Copiar, redistribuir y modificar excepto de las secciones invariantes. Los datos descargados no pertenecen a ninguna sección invariante, más aún en la página de ballotepedia no se identificaron secciones invariantes.
- Las copias se pueden vender comercialmente. Pero nuestro interés no es venderlo comercialmente sino el aprendizaje de web scrapping y generar un dataset informativo sobre las elecciones de los Estados Unidos.

Hasta aquí, se identificaron dos tipos de licencias que se pueden aplicar para este dataset, Open Database License (ODbL) y Creative Commons (CC). Las diferencias entre ambas radican en la operación y alcance de la licencia. Mientras que las licencias ODbL se aplican sólo a los derechos de base de datos sui generis y cualquier derecho de autor en la estructura de la base de datos, las CC aplica también el contenido de la información. En otras palabras, las licencias ODbL no aplican al contenido individual de la base de datos. Teniendo en cuenta que el contenido de este dataset está compuesto por información por estado que está sujeta a derechos de autor entonces esta licencia no aplicaría.

Otra diferencia importante es que las licencias ODbL pueden crear obligaciones contractuales incluso en jurisdicciones donde los derechos de base de datos no existirían de otro modo y serían necesarios sólo para el permiso de la licencia. Estas obligaciones se escapan del objetivo del presente trabajo por lo que esta sería otra razón para no utilizar esta licencia.

Por lo tanto, el tipo de licencias para esta dataset es CC, y hay de tres tipos: 1) Released Under CC0: Public Domain License, 2) Released Under CC BY-NC-SA 4.0 License y 3) Released Under CC BY-SA 4.0 License. La licencia CC BY-SA 4.0 y Public Domain permiten el uso comercial pero el objetivo de este dataset es solo para un fin educativo e informativo luego esta licencia no se ajusta. En

cambio, la licencia Released Under CC BYNC-SA no permite el uso comercial. En conclusión la licencia CC BYNC-SA 4.0 es la que se elige para el dataset.

9. Código

El código se puede encontrar en:
https://github.com/j-river1/Webscraping-Elecciones-EstadosUnidos/tree/main/src

10. Dataset

El dataset lo puede encontrar en:
https://github.com/j-river1/Webscraping-Elecciones-EstadosUnidos/tree/main/dataset

Y en zenodo:
https://zenodo.org/record/4263409#.X6mjr2gzbyQ

Recursos

Los siguientes recursos son de utilidad para 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.

Contribuciones

Contribuciones	Firma
Investigación Previa	Juan Camilo Rivera. Martin Loizate Sarrionandia.
Redacción de las respuestas	Martin Loizate Sarrionandia. Juan Camilo Rivera

Desarrollo condigo	Juan Camilo Rivera. Martin Loizate Sarrionandia