PRAC 1

M2.851 - Tipología y ciclo de vida de los datos

Autor: María del Mar Colino García

23/03/2019

Contenido

1 Objetivo.

Este documento aborda la resolución de la segunda prueba de evaluación continua, PRAC1, de la asignatura Tipología y ciclo de vida de los datos, para el segundo semestre del curso 2018-2019.

2 Enunciado

2.1 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

2.2 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 abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de web scraping.

2.3 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.

2.4 Descripción de la Práctica a realizar

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
2. Definir un título para el dataset. Elegir un título que sea descriptivo.
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente
5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.
6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).
7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.
8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
   - 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

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
10. Dataset. Presentar el dataset en formato CSV

2.5 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.

2.6 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.

Otros criterios que se tomarán en cuenta para la evaluación son:

- Idoneidad de las respuestas (deberán ser claras y completas).
- Complejidad del sitio web elegido para la extracción.
- Síntesis y claridad, a través del uso de comentarios, del código resultante.
- Presentación adecuada de los datos.
- Organización y claridad del documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.

2.7 Formato y fecha de entrega

Durante la semana del 8 de abril, 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 (dperez1@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 estén los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, debe aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación por parte del grupo que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	**Firma**

Investigación previa	Integrante 1, Integrante 2, ...
Redacción de las respuestas	Integrante 1, Integrante 2, ...
Desarrollo código	Integrante 1, Integrante 2, ...

3. Una carpeta con el código Python o R generado para obtener los datos.
4. El fichero CSV con los datos.

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 15 de abril**. No se aceptarán entregas fuera de plazo.

3 Solución de la PRAC

3.1 Contexto

El conjunto de datos objeto de la presente práctica se enfoca en:

- La persistencia de los Informes Públicos Periódicos (IPP) en formato XBRL, disponibles de forma pública en la Comisión Nacional del Mercado de Valores (CNMV).
- La persistencia de un conjunto de estados extraídos de los ficheros XBRL, listos para su posterior uso en análisis y procesos derivados.

El proceso de determinación del ámbito y alcance del juego de datos generado será parametrizable.

La CNMV proporciona la información objeto de interés en su papel de organismo encargado de supervisar e inspeccionar los mercados de valores españoles y la actividad de cuantos intervienen en los mismos.

3.2 Título dataset

Spain – XBRL – Public Standardized Financial Reports

3.3 Descripción dataset

El conjunto de datos generados por la aplicación de *web scraping* objeto de esta práctica contempla:

- Un repositorio de documentos XBRL.

  Conjunto de documentos IPP – XBRL con la Información Pública Periódica relativa a los estados financieros de las empresas españolas que cotizan en bolsa (fuente CNMV).

- Un repositorio de *statements*:

  Conjunto de estados y ratios extraídos de los ficheros XBRL del repositorio previo.

Es posible llevar a cabo la configuración del ámbito de datos a ser incluido en dichos repositorios, de acuerdo con las siguientes características:

- Sector financiero
- Periodo financiero

Además, es posible llevar a cabo la configuración del alcance de los datos, en cuanto al conjunto de estados a ser extraídos de los ficheros XBRL:

- Elemento/s
- Contexto/s

3.4 Representación gráfica

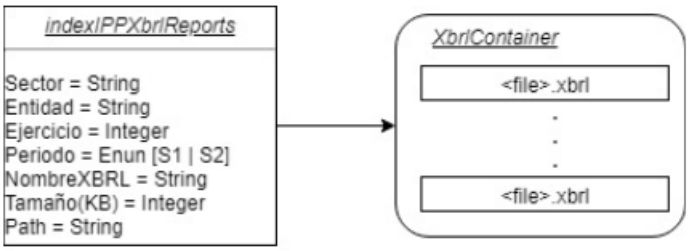
3.5 Contenido

Como se ha explicado previamente, existen dos áreas diferenciadas en el juego de datos generado:

- **Repositorio documentos XBRL.**

  Se ha decidido contemplar, como parte del juego de datos generado, la información en bruto de los ficheros XBRL asociados a cada informe IPP, así como sus datos generales, denominando a estos últimos como "datos de indexación". Se ha optado por esta estrategia, porque esta información en si misma puede resultar de interés para un gran conjunto de usuarios de los datos; disponer de la información en bruto permitirá poder realizar extracciones y procesados posteriores adaptados a las necesidades específicas de cada usuario del dataset.

  El modelo seguido se muestra a continuación:



  Los datos de indexación son persistidos en un fichero CSV. Los campos que define dicho fichero CSV se han incluido en el elemento identificado como "*indexIPPXbrlReports*"; cada entrada (o registro) en el fichero CSV apunta a un fichero XBRL persistido en un sistema de ficheros identificado como "*XbrlContainer*".

  La descripción de los campos es la siguiente:

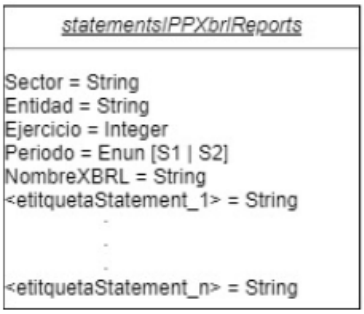
  - *Sector*: Identifica el sector empresarial al que pertenece la empresa que realiza el reporte. Los posibles valores son definidos por la CNMV y son:
    - PETROLEO,
    - ENERGÍA Y AGUA,
    - MINERIA,
    - QUÍMICAS,
    - TEXTIL Y PAPELERAS,
    - COMERCIO,
    - METAL MECÁNICA,
    - ALIMENTACIÓN,

  - CONSTRUCCIÓN,
  - INMOBILIARIAS,
  - TRANSPORTES Y COMUNICACIONES,
  - OTRAS INDUSTRIAS MANUFACTURERAS,
  - OTROS,
  - BANCOS,
  - CAJAS Y COOPERATIVAS DE CREDITO,
  - SEGUROS Y OTRAS INTERMEDIACIONES FINANCIERAS,
  - ESTADO,
  - OTROS ORGANISMOS PUBLICOS.
- *Entidad*: Define el nombre de la empresa o entidad que realiza el reporte.
- *Ejercicio*: Año fiscal al que aplica los datos del reporte.
- *Periodo*: Indica el semestre al que aplica el reporte, puede tomar los valores S1 y S2.
- *NombreXBRL*: Identificador fichero xbrl, se corresponde con el nombre del fichero sin la extensión.
- *Tamaño(KB):* Tamaño en KB del fichero xbrl.
- *Path*: Ruta de acceso al fichero xbrl, donde la ubicación del repositorio de ficheros XBRL será denotada por *<repositoryPath>.*

- **Repositorio *statements*.**

  Adicionalmente, se ha considerado interesante suministrar una extracción muy básica en base a los criterios de nombre de elemento XBRL y contexto aplicable al elemento, de tal forma que, usuarios de los datos con necesidades simples, puedan abstraerse de tener que cargar y procesar los ficheros XBRL.

  El modelo seguido se muestra a continuación:



  Los datos de estados financieros son persistidos en un fichero CSV, cuyos campos se muestran en la imagen previa, los campos identificados como <etiquetaStatement_1 .. n> definen las etiquetas que identifican a un determinado estado o ratio financiero dentro de la taxonomía estándar IPP Taxonomy.

  La descripción de los campos de información general puede encontrarse en el punto previo, y para los datos relativos a estados financieros en IPP Taxonomy.

Tal cual se indica en **3.3 DESCRIPCIÓN DATASET**, el ámbito y alcance de los datos a ser incluidos en el juego de datos resultantes puede ser parametrizado. En el conjunto de datos que se ha puesto a disposición en el repositorio, el ámbito y alcance de la carga y extracción son:

- Se han cargado los datos para todos los sectores financieros definidos por la CNMV, y periodo 2018.
- Se han extraído los estados identificados con las etiqueta y contexto listados en el apartado **4.1 ELEMENTOS XBRL CARGADOS** (más información sobre etiquetas posibles en IPP Taxonomy)

La estrategia seguida para la carga de los datos, se basa en la carga secuencial por sector y dentro de sector por periodo, de tal forma que no cargue en exceso el tráfico en la web de la CNMV.

3.6 Agradecimientos

El origen primario de los datos es cada una de las empresas que genera y envía a la CNMV la información de sus estados financieros semestrales, información que por ley ha de hacerse pública.

La CNMV como receptora y gestora de dicha información suministra en sus web mecanismos para su publicación.

Por lo tanto, ha de agradecerse tanto a las empresas como a la CNMV (CNMV, 2019) el poder disponer de esta información.

Adicionalmente, ha de agradecerse a la asociación española de XBRL su trabajo en difundir e informar sobre los estándares XBRL dentro del territorio nacional (xbrl.es, 2019).

3.7 Inspiración

El presente conjunto de datos tiene un gran ámbito de aplicación, uno de los principales usos que se puede hacer de este tipo de información es el análisis de estados financieros. Este tipo de análisis tiene una gran cantidad de usuarios, entre los que se encuentran:

- Entidades de crédito
- Accionistas
- Proveedores
- Clientes
- Empleados, comités de empresa y sindicatos.
- Auditores de cuentas
- Asesores
- Analistas financieros
- Administraciones Públicas
- Competidores
- Inversores y potenciales compradores de la empresa

Por lo que, este conjunto de datos puede ser usado directamente por usuarios de los tipos comentados, así como por consultoras o empresas de servicios que ofrezcan información o servicios a este tipo de usuarios.

Adicionalmente, y dada la actual corriente en la venta de datos, otro potencial cliente de este tipo de datos son las plataformas de venta de datos, pudiendo nutrirse tanto de los datos en bruto, como de los estados ya extraídos.

La información contenida en este juego de datos, entre las múltiples aplicaciones posibles, es de interés en:

- la aplicación de técnicas estadísticas
- la generación de datos derivados
- en el área de minería de datos, para:
    - elaborar modelos predictivos que permitan, por ejemplo, disponer de una visión de posibles evoluciones del mercado de valores, o de una empresa,
    - aplicar otro tipo de técnicas de minería más descriptivas, que permita justificar el porqué de ciertas reacciones en el mercado de valores, etc.

3.8 Licencia

La licencia seleccionada para llevar a cabo la publicación de los datos es una licencia **CC BY-NC-SA 4.0**, es decir, una licencia ***Creative Commons Non-Commercial*** y ***Share-Alike***.

Se escoge esta licencia ya que se considera las condiciones asociadas a las mismas, con las que mejor encajan al contexto del trabajo realizado.

Esta licencia fomenta la colaboración y la filosofía *open-source*, suministrando la libertad para:

- **Compartir**: Se permite copiar y redistribuir el material en cualquier medio o formato.
- **Adaptar**: Se permite transformar y crear a partir del material

Siempre y cuando se cumplan las siguientes condiciones:

- **Reconocimiento**: Se debe reconocer adecuadamente la autoría de la fuente, por lo que se deberá citar la fuente origina.
- **No comercial**: No se podrá usar el material para fines comerciales, por tanto, su ámbito se reduce al de investigación y académico.
- **Compartir en las mismas condiciones**: El término *Share Alike* de la licencia nos garantiza que cualquier modificación o uso de este material se publique bajo la misma licencia, en aras de promover la colaboración.

3.9 Código

Disponible en url: https://github.com/mmcolino/WebScrapingCNMV/tree/master/src

3.10 Dataset

3.10.1 Repositorio XBRLs

https://github.com/mmcolino/WebScrapingCNMV/tree/master/data/ipp-xbrl

3.10.2 Indexación XBRLs

https://github.com/mmcolino/WebScrapingCNMV/blob/master/data/csv/indexIPPXbrlReports.csv

3.10.3 Propiedades estados financieros

https://github.com/mmcolino/WebScrapingCNMV/blob/master/data/csv/propertiesIPPXbrl.csv

4 Anexos

4.1 Elementos XBRL Cargados

A continuación, se incluirán agrupados por categoría, la lista de elementos considerados en la extracción de estados de los XBRL.

Para cada elemento se indicará su identificador, su contexto y una breve descripción:

4.1.1 Balance consolidado

Id Element	Contexto	Descripción
I1040	Icur_PeriodoActualBalanceMiembro	A) ACTIVO NO CORRIENTE
I1085	Icur_PeriodoActualBalanceMiembro	B) ACTIVO CORRIENTE
I1100	Icur_PeriodoActualBalanceMiembro	TOTAL ACTIVO (A + B)
I1195	Icur_PeriodoActualBalanceMiembro	A) PATRIMONIO NETO (A.1 + A.2 + A.3)
I1120	Icur_PeriodoActualBalanceMiembro	B) PASIVO NO CORRIENTE
I1130	Icur_PeriodoActualBalanceMiembro	C) PASIVO CORRIENTE
I1200	Icur_PeriodoActualBalanceMiembro	TOTAL PASIVO Y PATRIMONIO NETO (A + B + C)

4.1.2 Cuenta de pérdidas y ganancias consolidada

Id Element	Contexto	Descripción
I1245	Dcur_AcumuladoActualMiembro_ImporteMiembro	RESULTADO DE EXPLOTACIÓN
I1256	Dcur_AcumuladoActualMiembro_ImporteMiembro	RESULTADO FINANCIERO
I1265	Dcur_AcumuladoActualMiembro_ImporteMiembro	RESULTADO ANTES DE IMPUESTOS
I1280	Dcur_AcumuladoActualMiembro_ImporteMiembro	RESULTADO DEL EJERCICIO PROCEDENTE DE OPERACIONES CONTINUADAS
I1285	Dcur_AcumuladoActualMiembro_ImporteMiembro	Resultado del ejercicio procedente de operaciones interrumpidas neto de impuestos
I1288	Dcur_AcumuladoActualMiembro_ImporteMiembro	RESULTADO CONSOLIDADO DEL EJERCICIO
I1290	Dcur_AcumuladoActualMiembro_ImporteMiembro	BENEFICIO POR ACCIÓN básico
I1295	Dcur_AcumuladoActualMiembro_ImporteMiembro	BENEFICIO POR ACCIÓN diluido

4.1.3 Estado de Ingresos y Gastos reconocidos consolidado

Id Element	Contexto	Descripción
I1305	Dcur_PeriodoActualMiembro	RESULTADO CONSOLIDADO DEL EJERCICIO (de la cuenta de pérdidas y ganancias)

I1310	Dcur_PeriodoActualMiembro	OTRO RESULTADO GLOBAL – PARTIDAS QUE NO SE RECLASIFICAN AL RESULTADO DEL PERIODO
I1350	Dcur_PeriodoActualMiembro	OTRO RESULTADO GLOBAL – PARTIDAS QUE PUEDEN RECLASIFICARSE POSTERIORMENTE AL RESULTADO DEL PERIODO
I1400	Dcur_PeriodoActualMiembro	RESULTADO GLOBAL TOTAL DEL EJERCICIO (A + B + C)

4.1.4 Plantilla Media

Id Element	Contexto	Descripción
I2295	Icur_IndividualMiembro_PeriodoActualMiembro	Plantilla Media - Total
I2296	Icur_IndividualMiembro_PeriodoActualMiembro	Plantilla Media - Hombres
I2297	Icur_IndividualMiembro_PeriodoActualMiembro	Plantilla Media - Mujeres

4.2 Bibliografía

CNMV. (30 de 03 de 2019). *CNMV*. Obtenido de http://www.cnmv.es

creativecommons.org. (03 de 30 de 2019). *creativecommons.org*. Obtenido de https://creativecommons.org

Lawson, R., & Jarmul, K. (2017). *Python Web Scraping (Second Edition).* Birminghan: Pack Publishing Ltd.

Minguillón, J. (s.f.). *Fundamentos de data science.* Barcelona: UOC.

Selenium. (20 de 03 de 2019). *Selenium*. Obtenido de http://www.seleniumhq.org/

Subirats Maté, L., & Calvo González, M. (s.f.). *Web scraping.* Barcelona: UOC.

xbrl.es. (30 de 03 de 2019). *xbrl.es*. Obtenido de https://xbrl.es/wp/