# Tipologia i cicle de vida de les dades

## PRÀCTICA 1: Web Scraping

**ESTUDIANT: David Gil del Rosal (dgilros@uoc.edu)**

## 1.- Context

El present treball usa web scraping per a recopilar informació sobre l'evolució dels preus i de les vendes de tabac a Espanya durant els anys 2005 a 2018.

A Espanya el tabac és un producte sotmès a una forta regulació degut al seu gran impacte sanitari i fiscal. El monopoli de la seva venda és de l'Estat, que l'exerceix a través de concessionaris anomenats estancs i màquines expedidores amb autorització de venda amb recàrrec [1]. Els preus del tabac no són lliures: han de ser aprovats mitjançant Resolucions de l'organisme regulador Comissionat per al Mercat de Tabacs (CMT) que es publiquen en la seu electrònica del Butlletí Oficial de l'Estat (BOE). D'altra banda, les estadístiques sobre les vendes de tabac són elaborades pel CMT mensualment amb les dades proporcionades pels estancs i es publiquen al lloc web del Ministeri d'Hisenda.

En conclusió, les dades s'obtindran de les següents fonts:

- Evolució dels preus de les labors de tabac: Resolucions sobre modificació de preus publicades a la seu electrònica del BOE al lloc web www.boe.es
- Estadístiques sobre vendes de tabac en unitats i euros: lloc web del Ministeri d'Hisenda www.hacienda.gob.es.

## 2.- Títol

El títol dels jocs de dades generats serà: **Evolució dels Preus i les Vendes de Tabac a Espanya**.

## 3.- Descripció del joc de dades

Generarem dos jocs de dades relacionats: un contindrà l'evolució dels preus del tabac i l'altre estadístiques sobre les vendes agrupades anualment i per Comunitat Autònoma.

Com s'explicarà a l'apartat 7 això ens permetrà analitzar l'evolució dels preus del tabac i el seu impacte sobre el consum.

# 4.- Representació gràfica

La següent imatge mostra diverses marques de tabac comercialitzades a Espanya:


# 5.- Contingut

En aquest apartat es descriu el contingut i estructura dels dos jocs de dades capturats.

El primer registrarà l'evolució dels preus de labors de tabac des de 2005 a 2018. Les dades es referiran a la Península Ibèrica i les Illes Balears, ja que Ceuta i Melilla tenen preus diferenciats per motius fiscals i poc impacte sobre el volum total de vendes, i Canàries no pertany a l'àrea del monopoli del tabac. S'obtindrà del BOE i tindrà els següents camps:

- **Marca**: marca de la labor de tabac, per exemple «Fortuna Rojo Blando».
- **Fecha**: data en la que el nou preu va entrar en vigor en format YYYYMMDD. Normalment, el dia següent a la publicació en el BOE de la Resolució de modificació de preus.
- **Precio**: nou preu de la marca en euros a la Península Ibèrica i les Illes Balears.

El segon joc de dades registrarà les estadístiques anuals de vendes de labors de tabac en quantitat i valor monetari en euros. S'obtindrà del lloc web del Ministeri d'Hisenda que publica aquestes estadístiques en format Excel. Els camps seran:

- **Comunidad**: Comunitat Autònoma a la que es refereix la dada.
- **Anyo**: any en que s'han produït les vendes.
- **Labor**. Tipus de labor de tabac: cigarrets, cigarros o picadures.
- **Unidad**. Unitat de mesura del total, «euros» o «cantidad».
- **Total**. Quantitat total venuda o recaptació en euros.

S'ha seleccionat l'any 2005 com a inici del període ja que a partir de 2006 va haver canvis normatius que van introduir més restriccions al consum de tabac provocant perturbacions en el mercat d'aquest producte [2]. Comparar amb les dades de 2005 ens permetrà analitzar el seu impacte. El fi del període és 2018 perquè és el darrer any amb estatístiques anuals.

# 6.- Agraïments

Cal agrair al BOE i al Ministeri d'Hisenda les facilitats per a la captura i reutilització de les dades. Especialment al BOE: el seus termes de servei permeten l'obtenció i utilització de la informació que publica gairebé sense restriccions [3].

Respecte a recerques anteriors, hi ha nombrosos precedents de l'ús de web scraping per a l'anàlisi de preus i transaccions comercials de productes[1] com per exemple la recerca de Cavallo i Rigobon per a la construcció d'un índex de preus per a diferents països [4].

# 7.- Inspiració

Com es va indicar al primer apartat, el tabac és un producte especial amb importants repercussions sanitàries i impacte fiscal, raons per la qual és d'interès ser capaç de respondre a preguntes com les següents:

- Com han evolucionat els preus del tabac els darrers anys? Han pujat tots els anys o hi ha hagut davallades?
- Com han evolucionat les vendes de tabac, en unitats i valor monetari?
- Hi ha una clara correlació negativa entre el nivell de preus i el consum de tabac? Són realment efectives les alçades de preu per a combatre el tabaquisme?.

Els jocs de dades recopilats permetran respondre a aquestes qüestions.

# 8.- Llicència

Tenint en compte que les dades han estat recopilades emparant-se en la Llei de reutilització d'informació al sector públic [5] la llicència més adequada és «CC BY-SA 4.0 License». Aquesta llicència permet copiar i redistribuïr les dades i adaptar-les, transformar-les i usar-les per a qualsevol propòsit, comercial o no, però té els següents requisits [6]:

- **Atribució**: si s'usen les dades s'ha d'indicar que provenen del BOE i el Ministeri d'Hisenda.
- **Compartició**: si les dades originals o transformades es redistribueixen cal usar la mateixa llicència i no es poden aplicar mesures legals o tecnològiques que limiten exercir aquests drets.

1 A Google Scholar, la cerca "web scraping prices" retorna més de 15.000 resultats.

# 9.- Codi

L'script de Python 3 **scraper_tabaco.py** que executa els processos de web scraping està disponible en la carpeta **src** del repositori GitHub https://github.com/dgilros/WebScraping. L'Annex A descriu el disseny i implementació del script.

# 10.- Datasets

Els dos *datasets* generats en format CSV separat per «;» **TabacoPrecios.csv** i **TabacoVentas.csv** estan disponibles a la carpeta **csv** del repositori GitHub https://github.com/dgilros/WebScraping.

# ANNEX A. Detalls tècnics.

Aquest annex descriu l'anàlisi, disseny i implementació de l'script de web scraping emprat a aquesta pràctica, així com un exemple de la utilització de les dades capturades per a respondre a certes preguntes d'interès analític.

## A.1.- Anàlisi

En primer lloc es va procedir a analitzar la viabilitat del procés de web scraping i les possibles alternatives, considerant els següents aspectes indicats per Subirats i Calvo [7]:

- **Legalitat:** la captura i ús de les dades està emparada per la normativa de reutilització de continguts del sector públic [5].
- **Existència d'APIs**: no hi ha APIs per a accedir aquestes dades. El BOE facilita la descàrrega de continguts en formats HTML, XML i PDF però ha de fer-se mitjançant peticions web especificant la data del butlletí. El Ministeri d'Hisenda no disposa d'un API per a la descàrrega de les estadístiques de vendes de tabac.
- **Obstacles al web scraping**. Els llocs web accedits no requereixen inici de sessió o ús de cookies ni tenen cap mecanisme antirobot tipus CAPTCHA. El lloc web del BOE sí disposa d'un fitxer robots.txt que limita la indexació de certes pàgines, però com es veurà després s'ha tingut en compte i no ha afectat als documents a obtenir.
- **Grandària:** reduïda. Del 2005 al 2018 s'han publicat 569 Resolucions de preus amb unes 20 marques de mitjana. Les estadístiques per any, Comunitat Autónoma i labor de tabac són 1.536. En total menys de 12.000 registres.
- **Tecnologies**. Les pàgines no fan un ús excessiu de JavaScript ni de tecnologies que compliquin el web scraping. Les pàgines a les que s'accedirà no tenen continguts audiovisuals. Una dificultat és que les estadístiques de vendes de tabac estan en fulles de càlcul d'Excel que presenten canvis de format en alguns anys. Això s'haurà de tenir en compte per a l'extracció de dades d'aquests fitxers.

# A.2.- Disseny i implementació

Les principals alternatives per a realitzar web scraping amb el llenguatge Python són la llibreria BeautifulSoup i el framework Scrapy [8, 9].

BeautifulSoup facilita la construcció de scrapers senzills i ofereix funcionalitats per a navegar pel DOM dels documents HTML o XML accedits.

Scrapy és més potent: permet la construcció de *crawlers* amb possibilitat de descàrrega en paral·lel i la definició de *pipelines* per a crear *workflows* de processament de les dades. Com a contrapartida la seva configuració i programació és més complicada.

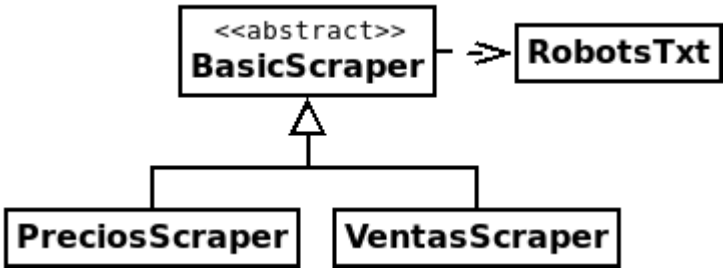
En aquest treball, degut al nombre reduït de llocs web i documents a accedir, hem optat per emprar BeautifulSoup però amb un interfície propi orientat a objectes que facilita la reutilització de funcionalitat comú per als dos lloc web als que accedim: el del BOE  i el del Ministeri d'Hisenda.

En concret, definirem una classe abstracta **BasicScraper** amb dos subclasses per a processar cada lloc web: **PreciosScraper** i **VentasScraper**. En síntesi, la funcionalitat és:

- Es recopilen els enllaços dels documents que contenen els preus i les estadístiques de vendes de tabac accedint al cercador del BOE i la pàgina corresponent del Ministeri d'Hisenda, respectivament. El URL de consulta generat pel cercador del BOE és complex i s'ha introduït literalment a l'script. Els paràmetres de consulta introduïts al formulari de cerca avançada de legislació del BOE per a generar el URL van ser:
  - **Título: «**comisionado para el mercado de tabacos».
  - **Rango:** Resolución.
  - **Materias:** «Tabaco Precios».
  - **Documentos por página**: 2000.
- Per a cada enllaç es processa el document corresponent analitzant el XML de les Resolucions de preus i la fulla de càlcul Excel de les estadístiques, i inserint els registres extrets en una llista. Els fitxers Excel s'analitzen amb la funció *read_excel()* de Pandas que permet llegir els fitxers des d'un URL i especificar quines files i columnes s'han de processar. Això és molt convenient perquè els fitxers de 2005 a 2014 tenen un format diferent que els de 2015 i posteriors. Els detalls poden observar-se al codi.
- Finalment, per a generar els CSV es creen *dataframes* de Pandas amb la llista de registres extrets al pas anterior i s'escriuen en fitxers CSV.

També es defineix una classe **RobotsTxt** com un *proxy* per a parsejar el fitxer robots.txt de la web del BOE, encara que cap dels documents accedits està restringit per aquest fitxer. D'altra banda, no s'ha considerat necessari introduir un retard entre peticions degut a que el volum de peticions es moderat: uns centenars en el cas de les Resolucions de preus i unes desenes en les estadístiques.

El següent diagrama UML mostra les classes que s'han descrit:



## A.3.- Exemple d'ús de les dades capturades

A la carpeta **csv** del repositori GitHub https://github.com/dgilros/WebScraping s'ha inclós un notebook de Jupyter **stats_tabaco.ipynb** amb exemples d'utilització dels fitxers CSV generats per a respondre a algunes preguntes presentades a l'apartat 7, en concret: evolució de preus d'una marca i evolució anual de vendes de cigarrets en quantitat i facturació.

# ANNEX B. Exemple avançat amb Selenium

Aquest Annex presenta un exemple de *web scraping* avançat amb Selenium, un API que permet invocar un navegador web (Chrome, Firefox, etc.) per a realitzar la navegació [8]. Amb Selenium es delega la interpretació de les pàgines web i el contingut dinàmic com JavaScript o AJAX al navegador, i l'script té accés als recursos accedits mitjançant el DOM.

El lloc web accedit és http://example.webscraping.com, un lloc de prova emprat al llibre de Richard Lawson [9]. Aquest lloc web presenta els següents reptes:

- Conté un formulari que permet cercar països al que enviem una consulta.
- Els resultats de la cerca es generen amb AJAX, per la qual cosa hem hagut d'introduïr retards per a obtenir-los.
- La paginació es realitza mitjançant una funció de JavaScript a la que invocarem.
- El propietari del lloc web controla el número de peticions consecutives, bloquejant la IP temporalment si són excessives per la qual cosa hem hagut d'introduïr retards.

L'script generat és a la carpeta **selenium** del repositori https://github.com/dgilros/WebScraping. L'script conté una classe que instància el webdriver amb Firefox, executa una cerca de països que contenen la lletra «z» al nom i navega pels resultats de cerca i les pàgines dels països per a extreure certa informació que s'emmagatzema al fitxer Countries.csv.

## Referències

[1] Ley 13/1998, de 4 de mayo, de Ordenación del Mercado de Tabacos y Normativa Tributaria. https://boe.es/buscar/act.php?id=BOE-A-1998-10407

[2] Ley 28/2005, de 26 de diciembre, de medidas sanitarias frente al tabaquismo y reguladora de la venta, el suministro, el consumo y la publicidad de los productos del tabaco. https://www.boe.es/buscar/doc.php?id=BOE-A-2005-21261

[3] BOE. Aviso Legal: Condiciones generales de reutilización. https://www.boe.es/informacion/aviso_legal/index.php#reutilizacion

[4] Cavallo, A.; Rigobon, R. (2016). «The Billion Prices Project: Using Online Prices for Measurement and Research». *Journal of Economic Perspectives* (vol. 30, núm. 2, pàg. 151-178). http://www.thebillionpricesproject.com/wp-content/papers/BPP_JEP.pdf

[5] Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público. https://www.boe.es/eli/es/l/2007/11/16/37/con

[6] https://creativecommons.org/licenses/by-sa/4.0/

[7] Subirats, L.; Calvo, M. (2018). *Web Scraping*. Editorial UOC.

[8] Mitchell, R. (2018). *Web Scraping with Python: collecting more data from the modern web*. 2nd. ed. O'Reilly.

[9] Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing.