PRÀCTICA 1
Tipologia i cicle de vida de les dades

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realització, s'han de complir els següents punts:

1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

La informació per crear un dataset amb totes les direccions postals, noms i telèfons de les farmàcies d'Espanya s'ha extret del següent lloc web:

http://esfarmacia.es/

Prèviament s'ha analitzat el fitxer Robots.txt per assegurar que no incorríem en cap mala praxis:

http://esfarmacia.es/robots.txt

Quan ja estàvem desenvolupament el codi en llenguatge Python ens hem adonat que la informació estava disponible en format XML en el següent enllaç. Usant el sitemap hauria estat més senzill el codi font, a la vegada que el sitemap ens ofereix una data de modificació per a cada farmàcia, que és una informació molt útil per optimitzar el temps del scraper. Malgrat tot hem seguit el camí de fer-ho amb un scraper que anés recorrent les diferents pàgines de la web, ja que aquesta complexitat afegida ja ens anava bé per practicar en l'ús de les llibreries de Phyton:

http://esfarmacia.es/sitemap.xml

El principal motiu d'escollir aquest lloc web és el fet que treballo en una empresa de venta de productes cosmètics, i recentment s'ha constituït una àrea comercial per vendre productes a les farmàcies de tot l'estat. Des de fa poc soc el tècnic responsable de dades de la companyia, i veient que el nou director de l'àrea comercial de farmàcies estava interessat en comprar un directori amb totes les direccions de farmàcies d'espanya, vaig pensar que podria usar la pràctica de la UOC per facilitar-li aquest fitxer sense cost, i posteriorment jo podria usar-lo en alguna de les eines de visualització de cartografia de les que disposem a la companya per tal de que ell pogués panificar les rutes de prospecció dels seus comercials per les diferents ciutats on tenen previst obrir mercat.

2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

El dataset l'he anomenat farmaDataset. Descriu que el que hi trobarem és informació relativa a farmàcies. Hi ha la informació imprescindible per plasmar-ho en eines de cartografia (codi postal, ciutat, provincia i direcció) i el telèfon per si el comercial hi vol telefonar abans de fer-hi la seva visita. Ens hauria anar bé que el dataset també inclogués els correus electrònics de la farmàcia, però aquest lloc web no disposa d'aquesta informació.

El títol és el mateix que hem usat per definir el repositori a GitHub on hi hem adjuntat tota la informació:

https://github.com/jordipuiggros/PharmaScraper

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

Per a cada farmàcia trobem en aquest web aquests atributs:

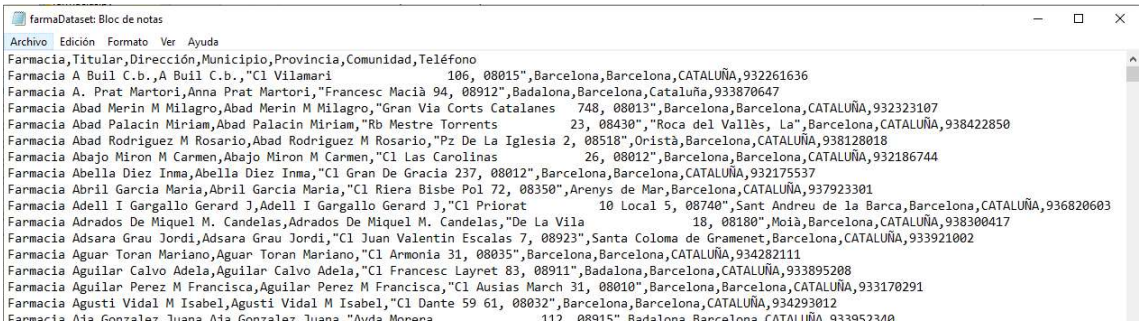
La informació que extraurem amb el Scraper i guardarem al dataset és exactament la que correspon als camps de la fitxa: Farmacia (nom de la farmàcia), Titular (nom de la empresa o persona particular propietària), Direcció (adreça postal), Municipi (ciutat on es troba), Provincia (provincia on es troba la farmàcia) i Teléfon.

En el cas de la provincia ens adonem que no sempre està escrita en el mateix format. Algunes posa Cataluña, altres Catalunya, en el cas de les que duen accent a vegades la trobem amb accent i a vegades sense. Passa quelcom similar amb les ciutat i a vegades trobem camps sense informar. Això fa que el dataset amb més de 15.000 farmacies per tal de ser útil al departament comercial i es pugui usar en eines de cartografia, calgui netejar-lo prèviament. Aquest procés ja no forma part d'aquesta pràctica. Normalment ho faria usant procediments T-SQL, que és com treballo normalment, però si és possible miraré de fer-ho usant els coneixements que adquirim en el següent mòdul de la UOC que veig que es titula neteja de dades i em pot ser útil per donar continuïtat al projecte aquest.

4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment

A continuació adjunto una captura del fitxer csv amb les columnes descrites en l'apartat anterior, de manera que gràficament ens vegi quin tipus d'informació estem extraient. La

primera fila correspon a la capçalera i les següents ja corresponen a la informació extreta del lloc web:



5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

En l'apartat 3, on descric el dataset faig una descripció de tots els camps. Per tal de no repetir-me ni allargar-me simplement els menciono a continuació:

Farmacia,Titular,Dirección,Municipio,Provincia,Comunidad,Teléfon

El període de temps del dataset i la data d'actualització de cada una de les farmàcies el trobem recollit en el sitemap. Una millora per quan automatitzem aquest procés perquè alimenti el DW de la companyia, serà revisar aquests dates i optimitzar el procés perquè només extregui informació quan hi hagi informació actualitzada. Sembla que la majoria de farmàcies es van generar al octubre del 2014 i tenen una freqüència d'actualització de canvis de període setmanal:

http://esfarmacia.es/sitemap.xml

El procés com s'ha recollit la informació és similar com ho faria una persona. No s'ha fet usant el sitemap (que hauria estat el més òptim i pràctic) sinó que s'ha fet entrant a les url en l'ordre que ho faria una persona que volgués obtenir la informació de manera manual per incorporar-la en un fitxer csv. Fer-ho aquesta manera ha afegit una complexitat addicional al codi que m'ha estat molt útil per posar en pràctica aquests coneixements, ja que prèviament mai havia programat en llenguatge phyton i molt menys havia usant la llibreria Beautifulsoup4.

Primer em recorregut les url provincials i allà hem guardat totes les urls que fan referencia a les fitxes de les farmàcies. Posterioment s'ha entrat a cada una d'aquestes i s'han guardat els atributs que en el web estan ressaltats en negreta.

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Segons el lloc web les dades són propietat de: Miguel Ángel Sesma con NIF 10867797K y Manuel Carbajal con NIF 53548099M.

http://esfarmacia.es/aviso-legal.html

En el mateix lloc trobem el següent advertiment, concretament en l'apartat 4 del avís legal que hi tenen publicat:

4.- Propiedad Intelectual. La propiedad intelectual vinculada a la página web (logotipos, diseño, contenidos, código de programación, animaciones,…) son propiedad de esFarmacia.es, quedando prohibida su reproducción total o parcial sin el expreso consentimiento de la empresa. A tal efecto, en el caso de producirse un uso indebido de la información vinculada a la propiedad intelectual de este portal por terceros, la empresa se reserva el derecho de desarrollar la actuaciones civiles o penales que considere con el fin de defender sus derechos.

Com que el cas que ens ocupa no és el d'usar aquesta informació per reproduir-la en un altre lloc web i permetre que tercers ho puguin usar, sinó que l'ús que en farem serà el d'emmagatzemar-ho en un datawarehouse per una explotació interna de consulta per part del director comercial, no estaríem vulnerant el fi pel qual s'ha fet pública aquesta informació.

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Aquest dataset per la companyia on treballo és interessant per varis motius:

-Permetrà mitjançant una eina de cartografia representar les dades sobre un mapa, i per tant el director comercial podrà dissenyar les rutes dels seus agents per tal de fer prospecció.

-Permetrà als comercials disposat dels telèfons i adreces cada dia de les farmàcies que han de visitar. El telèfon permetrà planificar la visita amb anterioritat si ho desitgen.

-Incorporant aquesta informació al nostra datawarehouse, podrem creuar-la amb la informació de punts de venta que ja tenim (on hi ha algunes farmàcies ja), i podrem indicar quines ja tenim assolides com a clients i per tant descartar-les del mapa de prospecció.

8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció: ○ 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

Basant-me amb la descripció que fa el portal en el seu avís legal sobre l'ús que permet fer d'aquesta informació, el dataset haurà de regir-se per les restriccions que hi puguin haver en el lloc origen d'on hem extret la informació:

2.-Objeto de la actividad y condiciones de uso. El portal www.esfarmacia.es tiene por objeto suministrar información de carácter público vinculada al mundo de las farmacias y ponerla a disposición de los usuarios que visiten el portal o de terceros portales interesados en disponer de dicha información. La información vinculada a los establecimientos se ha obtenido de fuentes públicas, ha sido suministrada por los propios establecimientos o sugerida por los usuarios con el fin de prestar un servicio a todos los posibles demandantes a dichos servicios. Los usuarios que visiten la web www.esfarmacia.es y usen los servicios ofrecidos en la misma se comprometen a respetar las condiciones indicadas en este Aviso Legal, debiendo leer y aceptar dichas condiciones como paso previo a la utilización de algunas de las funcionalidades del portal.

Així doncs, el llicenciament sembla que hauria de ser el menys restrictiu de tots, donat que explica que es tracta de dades obtingudes de fonts públiques, però com ja hem comentat a la pregunta 6, la mateixa web ens posa una restricció de no usar les dades per reproduir en altres llocs, i atenent al criteri de la advertència més restrictiva, donaria al dataset que em construït el següent llicenciament:

CC BY-NC-ND, que pel que sembla és el més restrictiu de tots:

> *The most restrictive Creative Commons (CC) license is the "Attribution-NonCommercial-NoDerivatives" (CC BY-NC-ND) license. This license is the most restrictive of the six (6) main CC licenses and only allows users to download the materials and share them with others as long as they credit the author.*

9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi del Scraper amb el qual s'ha obtingut el dataset de Farmacies s'ha elaborat en llenguatge phyton, usant la IDE de PyCharm, que disposa d'una versió amb llicenciament gratuït.

El codi el podeu trobar en dos fitxers amb extensió .py a Github:

Main.py: https://github.com/jordipuiggros/PharmaScraper/blob/master/main.py

Com veureu en els comentaris adjunts al codi font, aquest és el fil principal del Scraper. Aquí s'importen les llibreries necessàries, es crea l'objecte de la classe Farm, per tal d'usar-ne els seus mètodes i s'inicia la url principal de la web on volem extraure la informació: esfarmacia.es

Farmacias.py: https://github.com/jordipuiggros/PharmaScraper/blob/master/farmacias.py

Els mètodes que hi definim són els següents:

-obtenirlinksProvincies: a la pàgina principal cerquem els enllaços que fan referencia a les pàgines de províncies i en guardem els enllaços en una llista per recórrer més tard.

-obtenirlinksFarmacies: recorrem en l'enllaç d'una provincia de la llista d'enllaços de províncies que ens em construït abans, (totes les pàgines del 1 a la 120 per assegurar-nos que

entrem a totes a través d'un bucle al main) i obtenim els enllaços a les pàgines de cada farmàcia on hi ha la informació que ens interessa. Guardem tots aquests enllaços en una llista.

-obtenirFarmacia: A través del enllaç d'una farmàcia, amb aquest mètode obtenim els atributs que volem. Com que a la pàgina estan en negreta (strong) buscarem aquesta etiqueta html per trobar-los. Retornem els atributs de la farmàcia.

-guardarFarmaciesCSV: donada una ruta on guardar el dataset i una llista amb tots els atributs de farmàcies, procedirem a guardar-ho en un fitxer. Podem dir que aquest mètode és el que genera el dataset pròpiament dit.

10.Dataset. Presentar el dataset en format CSV

El dataset generat a partir del Scraper s'ha deixat penjat a GitHub en format CSV. El podeu consultar en la següent URL:

https://github.com/jordipuiggros/PharmaScraper/blob/master/farmaDataset.csv

La capçalera del fitxer es la següent:

Farmacia,Titular,Dirección,Municipio,Provincia,Comunidad,Teléfono

ANNEX. Llibreria Selenium

Després de la preentrega, i seguint les recomanacions de la tutora, he optat per experimentar usant la llibreria Selenium, que desconeixia fins al moment.

És una llibreria que permet usar els buscadors de les pàgines web.

He fet un script que permet buscar, usant el buscador de la mateixa pàgina web, les farmàcies d'una ciutat concreta i exportar-les en un fitxer csv usant la classe FARMA implementada anteriorment.

Per fer la proba ho he fet usant la ciutat on visc: "Igualada".

He deixat el codi d'aquest annex a github:

https://github.com/jordipuiggros/PharmaScraper/blob/master/Buscador.py

El dataset del buscador el podeu trobar aquí:

https://github.com/jordipuiggros/PharmaScraper/blob/master/farmaDatasetIgualada.csv