Prova d'avaluació continuada 1 – Tipologia i Cicle de Vida de les Dades

Enunciat

L'objectiu d'aquesta activitat serà la creació d'un data set a partir de les dades contingudes al web. Heu d'indicar les següents característiques del data set general:

1. *Títol del data set. Cal que poseu un títol que sigui descriptiu.*

El data set generat es pot titular "Catàleg diari de sabates Pompeii".

2. *Subtítol del data set. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.*

Podríem definir el conjunt de dades amb la definició següent: "Conjunt de models de sabatilles en oferta per la marca Pompeiibrand, tant per home com per a dona, amb les seves principals característiques i preus en el moment de l'extracció de la pàgina web oficial."

3. *Imatge. Agregueu una imatge que identifiqui el vostre data set visualment.*

Com a comentari addicional, recalcar que totes les imatges utilitzades **són propietat de la marca Pompeiibrand™** i que com a tal estan subjectes a drets d'autor. Aquesta imatge és per tant propietat privada i no es vol utilitzar amb fins comercials, sinó que es basa en la realització d'un estudi concret amb un objectiu purament de recerca i investigació.

4. *Context. Quina és la matèria del conjunt de dades?*

El conjunt de dades extret correspon al catàleg de models de sabatilles que apareixen en venda a la pàgina web de la marca Pompeiibrand, segmentant per model i sexe, de cara a obtenir les característiques pròpies del model tals com a color, preu i si el model. Aquesta extracció està prevista de manera diària per a obtenir el conjunt de productes disponibles en funció del a temporalitat, ja que no és possible obtenir dades històriques de la pàgina web directament.

5. *Contingut. Quins camps inclou? Quin és el període de temps de les dades i com s'ha recollit?*

L'objectiu del codi es generar un data set amb l'estructura següent:

Atribut	*Definició*
Data	Data de l'extracció de les dades en format 'YYYY/MM/DD'.
Marca	Marca del producte, concretament 'Pompeii' en aquest cas, tot i que es podria ampliar amb altres marques.
Gènere	Gènere al que està dedicat el producte, a escollir entre 'Home' o 'Dona'.
Model	Model del producte.
Color	Combinació de colors del producte, p.e. 'EVERGLADE CARAMEL'
Oferta	Tarifa del producte, que pot variar entre 'Precio habitual' i altres categories en cas de descomptes o promocions.
Preu	Preu en format *número real* amb dos decimals. El valor obtingut segueix el format 'XX,XX€'.
Talla	Talla del model. Cada talla d'un mateix model i color generarà un registre nou en el data set.

Com s'ha esmentat prèviament, l'objectiu de l'extracció es obtenir el conjunt diari de productes ofertats, ja que no es disposa de registres històrics per dates, i el generarem doncs internament. La periodicitat de les extraccions serà diària i incremental, recorrent tots els productes mostrats a les dues url's (home/dona) i emmagatzemant les dades en una taula conjunta.

6. *Agraïments. Qui és propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.*

En darrera instància, les dades són propietat única i exclusiva de la marca Pompeiibrand™ com ja s'ha esmentat prèviament, i s'utilitzen en aquest data set amb una finalitat de recerca i investigació, així com amb fins de docència pròpia.

Agrair doncs als propietaris de la marca Pompeiibrand™ la cessió de les dades, disponibles en obert. En relació amb la pregunta, presentar el fitxer robots.txt de la pagina web obtingut mitjançant l'enllaç següent:

https://www.pompeiibrand.com/robots.txt

```
# we use Shopify as our ecommerce platform

User-agent: *
Disallow: /admin
Disallow: /cart
Disallow: /orders
Disallow: /checkout
Disallow: /362709002/checkouts
Disallow: /362709002/orders
Disallow: /carts
Disallow: /account
Disallow: /collections/*+*
Disallow: /collections/*%2B*
Disallow: /collections/*%2b*
Disallow: /blogs/*+*
Disallow: /blogs/*%2B*
Disallow: /blogs/*%2b*
Disallow: /*design_theme_id*
Disallow: /*preview_theme_id*
Disallow: /*preview_script_id*
Disallow: /gift_cards/
Disallow: /policies/
```

```
Disallow: /apple-app-site-association
Sitemap: https://www.pompeiibrand.com/sitemap.xml

# Google adsbot ignores robots.txt unless specifically named!
User-agent: adsbot-google
Disallow: /checkout
Disallow: /carts
Disallow: /orders
Disallow: /362709002/checkouts
Disallow: /362709002/orders
Disallow: /gift_cards/
Disallow: /*design_theme_id*
Disallow: /*preview_theme_id*
Disallow: /*preview_script_id*

User-agent: Nutch
Disallow: /

User-agent: MJ12bot
Crawl-Delay: 10

User-agent: Pinterest
Crawl-delay: 1
```

Com a informació de la pàgina web, obtenim els següents resultats:

In[467]: builtwith.parse('https://www.pompeiibrand.com')
Out[467]:
{'web-servers': ['Nginx'],
'font-scripts': ['Google Font API'],
'tag-managers': ['Google Tag Manager'],
'ecommerce': ['Shopify'],
'analytics': ['TrackJs'],
'javascript-frameworks': ['jQuery']}

In[468]: print(whois.whois('pompeiibrand.com'))
{
"domain_name": "POMPEIIBRAND.COM",
"registrar": "DonDominio (SCIP)",
"whois_server": "whois.scip.es",
"referral_url": null,
"updated_date": [
"2017-06-07 15:00:14",
"2017-06-07 17:01:03"
],
"creation_date": "2013-11-20 17:10:04",
"expiration_date": "2019-11-20 17:10:04",
"name_servers": [
"NS-1234.AWSDNS-26.ORG",
"NS-164.AWSDNS-20.COM",
"NS-1933.AWSDNS-49.CO.UK",
"NS-612.AWSDNS-12.NET"
],
"status": [
"clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
"ok http://www.icann.org/epp#ok",

```
  "clientTransferProhibited http://www.icann.org/epp#clientTransferProhibited"
 ],
 "emails": "abuse@scip.es",
 "dnssec": [
  "unsigned",
  "Unsigned"
 ],
 "name": "Redacted for privacy",
 "org": "Paper Plane Partners S.L.",
 "address": "Redacted for privacy",
 "city": "Redacted for privacy",
 "state": "Madrid",
 "zipcode": "Redacted for privacy",
 "country": "ES"
}
```


7. *Inspiració. Per què és interessant aquest conjunt de dades? Quines preguntes li agradaria respondre la comunitat?*

El principal motiu per el qual s'ha escollit aquest set de dades ha estat el fet de ser un consumidor habitual de la marca. Després d'un temps seguint els productes que s'ofereixen, he pogut apreciar certs detalls que m'han semblat interessants per a un anàlisi, i que giren entorn de com els propietaris de la marca promouen els seus productes.

De fet, la oferta de productes és molt variable, amb canvis de tarifa, promocions, productes antics que tornen a entrar en el catàleg,... Un dels punts que em cridaven l'atenció era poder analitzar l'estratègia de màrketing de la marca, podent veure la disponibilitat dels productes, quan de temps estan en oferta, temporalitat de les promocions, creixement o reducció de la oferta, però també com a consumidor, ser capaç de preveure promocions i descomptes en funció de dades antigues.

La idea es doncs exportar diàriament el conjunt de productes disponibles (models, colors i talles per gènere) i veure l'evolució de l'stock.

Com a apunt interessant, m'agradaria remarcar un intent frustrat inicial d'obtenir dades estadístiques de la NBA, de cara a poder establir un model predictiu de resultats a partir dels resultats de la temporada present i de les actuacions dels jugadors en els darrers partits. Malauradament, un alt percentatge de les dades estava protegit contra descàrregues (un cop accedit al contingut desitjat es retornava un missatge de *Protected content – Copyright* tot i que la inspecció dels elements al *browser* mostrava la informació que es volia extreure.

8. *Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir perquè l'heu seleccionada:*
   - *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*

Entre les llicències disponibles, aquest codi es vol distribuir subjecte a la no-comercialitat de les dades, i de manera que no es pugui distribuir sense assegurar que aquesta no-comercialitat es manté, escollint per tant la llicència *Released Under CC BY-NC-SA 4.0.*

De fet, la llicència *Released Under CC0: Public Domain License* no permet assegurar que aquesta no-comercialitat es mantingui, i, d'altra banda, la llicencia *Released Under CC BY-SA 4.0 License* pot implicar que no es reconeguin els drets sobre les dades dels propietaris de la marca, i per tant es descarta igualment. La llicència *Database released under Open Database License* fa referència a un set de dades concret i no aplica per tant al codi compartit.

Donat que el codi s'ha dissenyat amb fins educatius i de recerca, qualsevol distribució o modificació ha de quedar registrada en el codi respecte a la versió original, i respectant que la llicència es manté per a eventuals modificacions.

Addicionalment, és important destacar que les dades extretes són propietat registrada de la marca Pompeiibrand™ i que l'ús del codi per a extreure aquestes dades ha de respectar les condicions indicades per els propietaris en el fitxer robots.txt, i utilitzar-se amb bona voluntat, sense cap objectiu maliciós de cara a pertorbar el servei de la pàgina web.

9. *Codi: Cal adjuntar el codi amb el que heu generat el data set, preferiblement amb R o Python, que us ha ajudat a generar el data set*

El codi ve adjunt al repositori de git compartir amb motiu de la pràctica sota el nom:

'TipoiCicledeVidadelesDades-PEC1.py'

TipoiCicledeVidadel
esDades-PEC1.py

10. *Data set: Data set en format CSV.*

El data set també està adjuntat sota el format .csv en el repositori de github sota el nom 'shoe-catalog.csv'.

shoe-catalog.csv