Documentació de la pràctica

1. Context

En el context del món assegurador, pel bon disseny d'un producte és molt important conèixer el tipus de cobertures que ofereix la competència, ja que aquestes poden ser força heterogènies. Per tal, no només cal competir en preus per captar clients sinó també és necessari fer una comparativa ràpida entre les cobertures ofertes en el mercat, que poden canviar ràpidament en funció de les circumstàncies (e.g. COVID-19)

L'objectiu d'aquest projecte és desenvolupar un codi de web-scrapping que recopili les cobertures ofertes per diferents companyies asseguradores per un producte assegurador d'automòbil, consultant a la pàgina web de cada entitat. A continuació es faciliten algunes de les pàgines utilitzades com a font per a alimentar el dataset:

- Zurich
- Mapfre
- Allianz

2. Títol del dataset

En funció del context aportat a l'apartat anterior, el títol serà el següent:

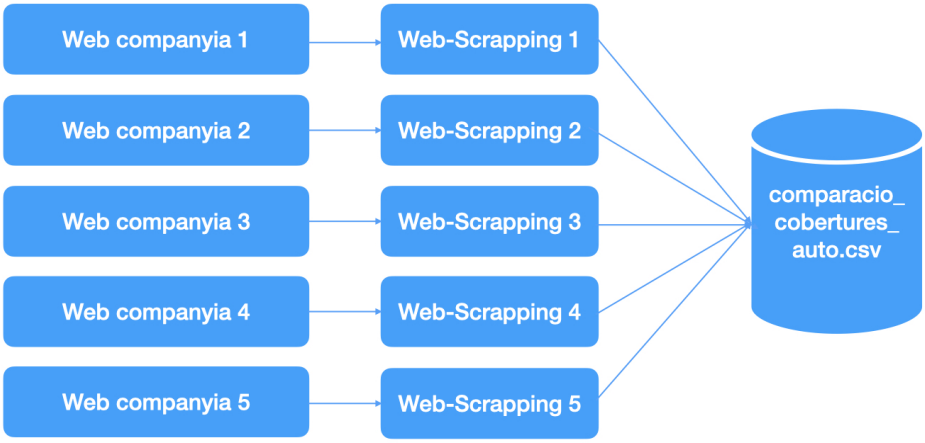
> Comparació de cobertures d'assegurances d'auto entre diferents companyies

3. Descripció del dataset

El dataset conté informació sobre diferents cobertures i garanties que cobreix cada paquet d'assegurances d'automòbil ofert per cada companyia d'assegurances.

Les dades s'obtenen a partir de les pàgines web de les diferents companyies i es preprocessen per tal d'homogeneitzar el format. Així es permet fer una anàlisi directa de les dades, tot i que els noms d'algunes cobertures podrien representar el mateix però anomenar-se diferent. En aquest cas no hem trobat una manera d'automatitzar la normalització dels camps de cobertura i paquet.

4. Representació gràfica



5. Continguts

El dataset conté informació de les cobertures d'auto extretes a una determinada data. La informació s'ha recollit executant un script de Python on es realitza web-scrapping sobre les webs esmentades en l'apartat 1.

En aquest cas, l'extracció es va realitzar en el mes de març de 2022. Com s'ha mostrat a l'apartat 4, l'script consulta al comparador de cobertures de cada pàgina web per tal d'extreure la taula corresponent de cobertures per cada paquet d'assegurances d'auto.

Finalment es transforma cada taula amb la finalitat de tenir les dades tan normalitzades com sigui possible.

Els camps del dataset són els següents:

Nom	Format	Descripció	Exemple
data	Date	Dia d'extracció de la informació	2022-03-25
companyia	String	Nom de la companyia	Catalana Occidente
producte	String	Producte	Auto
paquet	String	Tipus de paquet del producte	Terceros básico
garantia	String	Tipus de garantia del producte	Responsabilidad civil obligatoria
cobertura	String	Descripció del grau de cobertura	Incluido

6. Agraïments

Agraïm a les diferents webs des d'on hem extret la info:

- Catalana Occident
- Direct Seguros
- MMT Seguros
- RACC
- Seguros Bilbao

En aquest cas, donat que es tracta de dades públiques, aquest projecte no entra en conflicte amb cap normativa de protecció de dades personals.

Mirant les condicions web de cada lloc, tampoc hi ha cap menció específica al web-scrapping.

Finalment, en consultar els fitxers de `robots.txt` no em trobat que el directori on es troba el comparador de cobertures estigui exclós.

Posant com a exemple el fitxer `robots.txt` de Catalana Occident:

```
User-agent: *
Disallow: /*.file$
Disallow: /blog/wp-admin/
Disallow: /sites/Satellite*
Sitemap: https://www.seguroscatalanaoccidente.com/sitemap_index.xml
```

Simplement està excloent directoris diferents del directori d'interès:

> https://www.seguroscatalanaoccidente.com/seguros-coche-comparativa

7. Inspiració

Aquest projecte sorgeix del fet de que un dels integrants (Jonathan Mir), treballa a una companyia d'assegurances, on dintre dels departaments de negoci i producte es requereix continuament fer una anàlisi comparativa dels productes que ofereix la companyia versus la competència.

8. Llicència

Segons la privacitat de les dades que tenim, la llicència que elegim és la *Released Under CC0: Public Domain License*, aquesta llicència és emprada en llocs de domini públic, on donen plena llibertat a usar la informació aportada, veiem que les dades extretes són de domini públic, ja que qualsevol persona interessada a buscar un segur de cotxes podrà accedir a totes aquestes dades, sense cap impediment, a més, l'accés a aquestes no està controlat per cap procés de seguretat, pel fet que no ens demanen ni la validació mitjançant captcha, ni és necessari estar registrat.

9. Codi

El codi es troba en el següent repositori de GitHub

10. Dataset link

El dataset generat es pot descarregar des de Zenodo

11. Video link

El video es pot descarregar en el següent enllaç:

Video link

Taula de contribucions

Contribucions	**Signatura**
Investigació prèvia	JMF, DCG
Redacció de les respostes	JMF, DCG
Desenvolupament del codi	JMF, DCG