# PRAC 1

Laura Gassó Montserrat - Oriol Toll Roca

**Context:** Degut a la inquietud dels estudiants al finalitzar els seus estudis a la UOC en relació a la oferta laboral, s'ha creat un programa que ajuda als estudiants a buscar ofertes de treball a Linkedin en funció de paraules clau i d'una ubicació. L'objectiu d'aquest cas pràctic és aconseguir informació de forma automàtica, que ajudi als estudiants a trobar les ofertes de feina més interessants per a cada cas en particular. En aquest cas, es buscaran ofertes "d'Analista de Dades" a "Barcelona".

**Títol: Cerca d'ofertes de treball a Linkedin**

**Descripció del dataset:** Tal com expressa el títol, es crearà un dataset amb la informació que es creu més rellevant sobre les diferents ofertes de treball que apareixen a la web de Linkedin en funció d'un perfil i una cerca determinada. A tall d'exemple, a continuació es llista la possible informació que podria contenir el dataset: el nom de l'oferta, el nom de la companyia, l'ubicació, el temps que la oferta ha estat activa a la plataforma, si és a jornada completa, etc.

La majoria de les dades són categòriques.

Com les dades no han passat un procés de neteja, poden existir inconsistències i que el format d'elles no sigui el més adient per futurs anàlisis. És el cas de la variable "temps_oferta", que representa el temps que porta l'oferta de treball activa a la plataforma. Al no estar tractada, les seves unitats varien en funció de l'oferta i poden estar en mesos, setmanes o dies.

Un dels paràmetres que pot triar l'usuari és el nombre de pàgines que es volen extreure de la búsqueda. Així doncs, és decisió de l'usuari el nombre d'ofertes que vol descarregar sobre la mateixa busqueda (paraules clau - ubicació). Només s'ha de tenir en compte que cada pàgina pot contenir fins a un màxim de 25 ofertes. En funció d'aquest paràmetre el dataset tindrà una mida o una altre. El format final del dataset és un fitxer CSV que facilitarà el seu posterior tractament així com la visualització de les ofertes més interessants.

**Representació gràfica:**


**Contingut:**

| Variables | Descripció | Exemple |
|---|---|---|
| Nom Oferta | Nom de l'oferta al portal de Linkedin. | Data Scientist |
| Empresa | Empresa que publica l'oferta. | Mediktor |
| Ubicació | Ubicació de l'oferta de feina. | Barcelona, Cataluña, España |
| Temps Oferta | Temps que la oferta ha estat publicada a Linkedin. | hace 3 semanas |
| Jornada | Tipus de jornada laboral de la oferta: completa, temps parcial, etc. | Jornada completa |

Cada oferta laboral té un temps de publicació a Linkedin. El període de temps de les dades, és diferent per a cada oferta i per tant, a cada instància del set de dades li correspondrà un període diferent igual al valor de la variable Temps Oferta.

El contingut del set de dades es crea a partir d'un codi amb python que realitza web scraping a partir de la llibreria Selenium.

El codi consta dels següents passos:

A grans trets, el que fa el codi creat és fer una búsqueda de les ofertes de treball que s'ajusten més al perfil amb el que hem entrat a Linkedin en funció de les paraules clau utilitzades. Un cop trobades les ofertes de treball, el codi anirà iterant per sobre les ofertes de la primera pàgina agafant la informació que li hem demanat, un cop acabi les ofertes de la primera pàgina, passarà a la següent Fins a arribar al nombre de pàgines que l'usuari hagi escollit.

Com a next steps:

- Un cop realitzada l'obtenció de la informació de la web de Linkedin, aquesta s'hauria de processar per tal de consolidar les unitats temporals, neteja de caràcters en el text que poden dificultar la seva lectura, etc.
- Es podria crear un apartat de filtres per ordenar el llistat segons les preferències de l'usuari.

**Agraïments:** El propietari del lloc web i del conjunt de dades extretes de la pàgina és Linkedin, que forma part del grup Microsoft.

En el fitxer de robots el propietari expressa que l'ús de robots o altres mitjans automatitzats per a accedir a LinkedIn sense el permís exprés de LinkedIn està estrictament prohibit. Per tant es contacta amb el propietari a través del contacte especificat a robots "whitelist-crawl@linkedin.com" per demanar permís. I s'accepten el termes i condicions que es poden trobar a http://www.linkedin.com/legal/crawling-terms

Existeixen diferents anàlisis de les ofertes de treball de LinkedIn, fent una cerca ràpida a Google, en podem trobar diversos com:

- https://medium.com/datos-y-ciencia/una-forma-interactiva-para-buscar-y-analizar-ofertas-de-trabajo-en-la-web-ef9327b0a8d3
- https://jecas.es/web-scraping-para-caracteristicas-de-empresas/

**Inspiració:** La idea de fer un bot que reculli la informació de les diferents ofertes, sorgeix de la necessitat d'evitar buscar aquesta informació de manera manual i així estalviar temps a l'usuari. Dins de Linkedin, per poder crear una API que reculli aquesta informació, és necessari un compte de desenvolupador de Linkedin. Per aquest motiu, s'implementa un codi Python per fer-ho de manera gratuïta.

Es proposa fer una cerca de les ofertes de treball utilitzant una paraula clau i una ubicació. És interessant recollir informació que permeti un filtratge de les ofertes més interessants per a l'usuari. Així per exemple, es pot agafar informació com el temps que porten aquestes ofertes a Linkedin, el tipus d'horari que demanen, l'experiència necessària per realitzar la feina, etc.

Els anàlisis esmentats anteriorment cobreixen exactament l'objectiu que es persegueix en l'exercici d'aquesta pràctica. No obstant això, tot i que el codi sigui públic i s'hagués pogut aprofitar, els paquets que utilitzen són request i Beautiful Soup. Per a la realització d'aquest exercici, es volia utilitzar Selenium per aprofundir més i complementar els continguts de l'assignatura.



**Codi:** El codi de Python es pot trobar en el següent repositori GIT. "https://github.com/laugamo/web-scraping-uoc".

**Dataset:** CSV publicat a Zenodo. Enllaç del DOI: "DOI: 10.5281/zenodo.6413079".

| Contribucions | Signatura |
|---|---|
| Investigació prèvia | Oriol Toll, Laura Gassó |
| Redacció de les respostes | Oriol Toll, Laura Gassó |
| Desenvolupament del codi | Oriol Toll, Laura Gassó |