To run this model

1-Go to "web\_scraping\_links.ipynb" then run the code to collect ("Job Name", "Date", "Link") in this code each site will be dealt with separately

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص

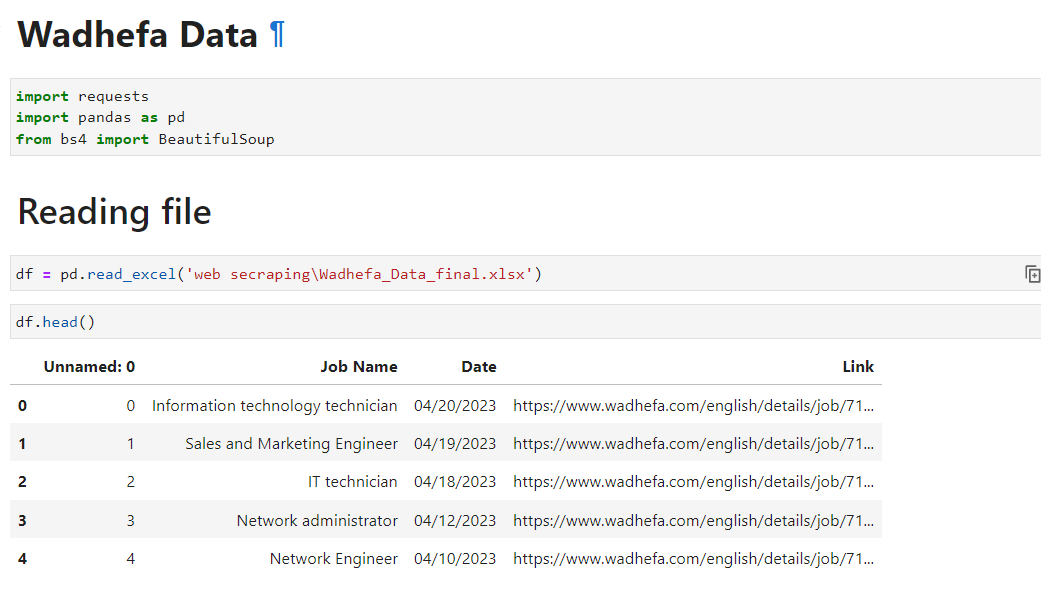
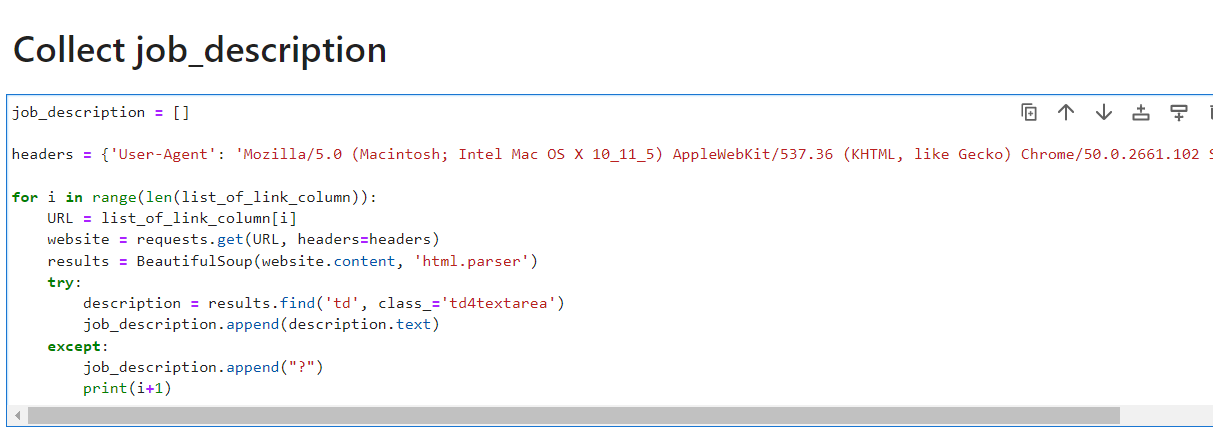
تم إنشاء الوصف تلقائياً

In this code, you only need to pay attention to the number of functions in the code and the location, so as not to cause collection time problems.

|  |  |
| --- | --- |
|  |  |
|  |  |

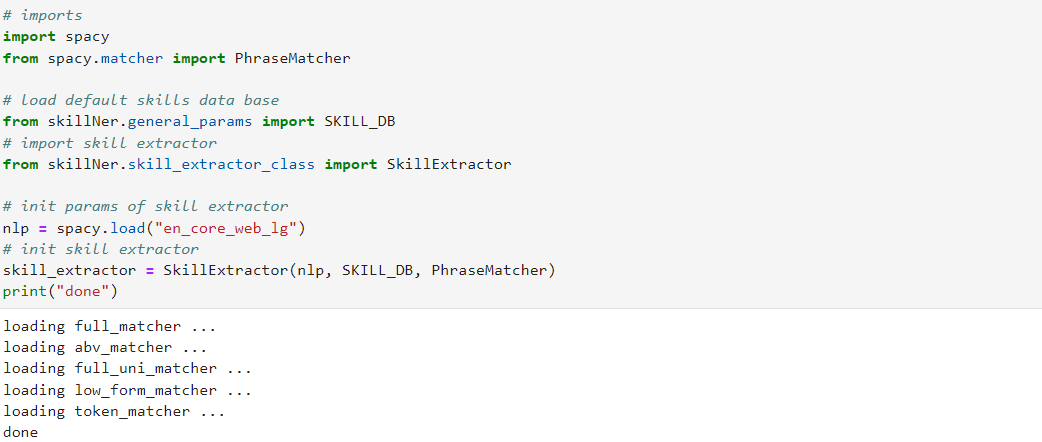
After complete run this code will be create two file ("Bayt\_Data\_final.xlsx", " Wadhefa\_Data\_final.xlsx") this file moved tow next step

2- Analysis code " Job\_description.ipynb" in this code first read file coming from web scraping then collect job description, and cleaning dataset.



Import SKILLNER



After that reading (skills.xlsx) file

Skills file content all keyword for IT skills (collecting from <https://lightcast.io/open-skills/categories>)

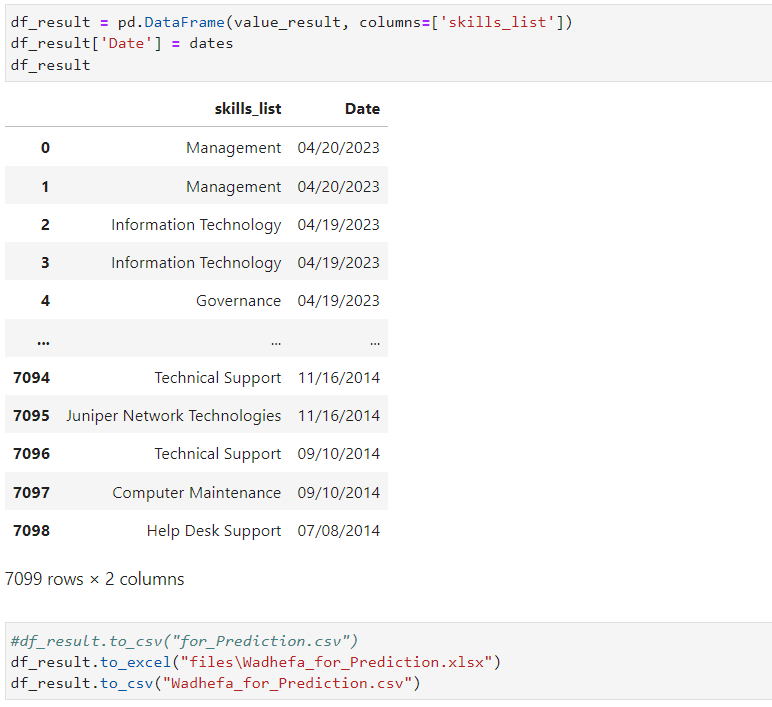
In case need more skills put more keywords can edit of excel fileصورة تحتوي على نص

تم إنشاء الوصف تلقائياً

Matching between job description and stop words.



Skills list and create ("Wadhefa\_for\_Prediction.csv", "Bayt\_for\_Prediction.csv") files for moved to next step (each site create separately file).



Put skill type.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

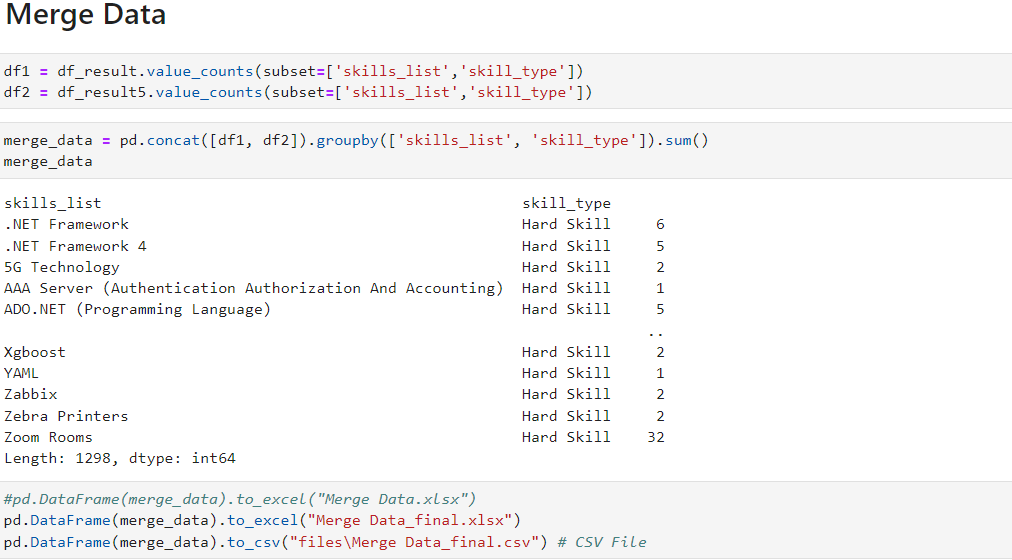
Skills list

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

And after end Wadhefa.com data go to Byte.com data and applying same steps (The code is built to handle each site separately)

After that Marge result with other and create "Merge Data\_final.xlsx" file content final result analysis (Result are skills required today in the market)



Graph

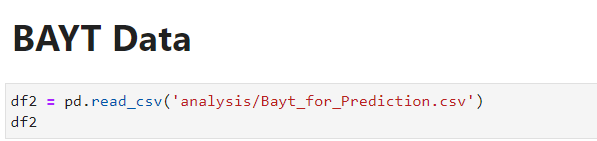
صورة تحتوي على مخطط بياني

تم إنشاء الوصف تلقائياً

3- Pre-processing before machine learning read each file ("Wadhefa\_for\_Prediction.csv", "Bayt\_for\_Prediction.csv").

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً



Delete column not necessary to predict and concatenate all row.

Then separation data into two array soft\_skills = [], hard\_skills = [] (depending on soft skills keyword)

[In order to obtain better results in prediction because if they are not separated, soft skills always take the highest rate because they are the most visible compared to hard skills and affect the prediction results].

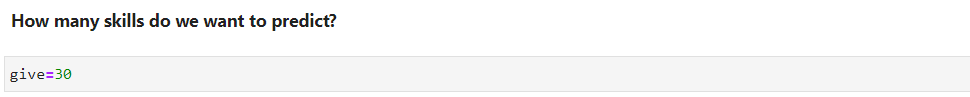
صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

Added the counter, determined the lowest and highest value, calculated the rate, added the label coding tool, and created a final prediction file ("hard\_final.csv", " soft\_final.csv") are moveing to next step. It contains ("skills\_list", "count", "rate", "skills\_list\_catogrised")



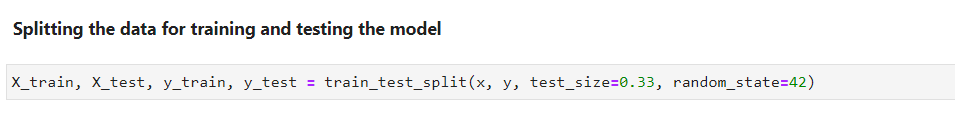
4- Model Machine learning " Model.ipynb" in this code first time determined 'How many skills do we want to predict?'



Then read finals files ("hard\_final.csv", " soft\_final.csv") in each time read one file, after that delete unnecessary columns and fit X and Y



Then Splitting the data for training and testing the model



Then calculated score and D2 absolute error score for each model.



Then create file content list of predicted skills.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

This CSV file create moved to next step.

5- Recommendation for professional certificates " recommendation.ipynb "

This code has two mine array ["main\_skill\_category "," main\_certificates "] each one has many small array.

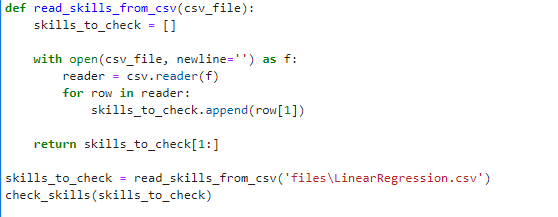
The idea of the recommendations is that the Information Technology (IT) Sub-Categories in Information Technology are grouped according to a classification EMIS database, small arrays are Sub-Categories name in "main\_skill\_category " is contains skills name keyword in each one, in "main\_certificates" contains professional certificates in each one.



صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

Then read file have result prediction.



It matches skills, prediction outputs, and arrays to find which ones are within which Sub-Categories to select appropriate professional certificates.

