Rapport projet langage de script

M1 - Sem1 2018-2019 INALCO

Par :
ORMAECHEA, Lucia
MONACO, Andrea Francesco
GUEYE, Ousseynou

Enseignant :
PLANCQ, Clément

I. **Mise en contexte :**

L'éducation est une des portes d'entrée du monde professionnel. Dans un monde de plus en plus globalisé, où la demande du marché évolue constamment, le choix éducatif est essentiel.

Étant donnée le taux de chômage élevé parmi les jeunes, nous avons voulu établir un tableau regroupant des données sur l'éducation, sur la démographie et l'emploi. De même, nous avons concentré nos recherches sur la France.

II.**Les données de base :**

Nous avons choisi trois jeux de données liées que nous avons décidé de confronter.

Les voici :

- Le jeu de données intitulé « Insertion professionnelle des diplômés de Master en universités et établissement assimilés[1] » (qu'on appellera *JdD1*), portant sur l'employabilité des jeunes, extrait de la *Plateforme ouverte des données publiques françaises (https://www.data.gouv.fr/)*.

- Le jeu de données nommé « Statistiques sur les effectifs d'étudiants inscrits par établissement public sous tutelle du ministère en charge de l'Enseignement supérieur[2] » (qu'on appellera *JdD2*), auquel on accédera par API, extrait du site OpenData, qui comporte des données numériques reliées à

---

[1] https://www.data.gouv.fr/fr/datasets/insertion-professionnelle-des-diplomes-de-master-en-universites-et-etablissements-assimil-0/?fbclid=IwAR0wpSSCUdzGpLDd6J9jyHlMMLkKjhvhBiajtuLiG94_X_uF6U6Ow9oeLJw

[2] https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-statistiques-sur-les-effectifs-d-etudiants-inscrits-par-etablissement/api/?sort=-rentree&fbclid=IwAR2rCif1TXwYmPxOwLsqXuUNQNjgZUz_EFvV9QUrel0Hga7zYXEAY-iyM_I

l'enseignement supérieur (https://data.enseignementsup-recherche.gouv.fr/).

- Le jeu de données appelé « Estimation de la population au 1er janvier[3] » (qu'on appellera *JdD3*), qui contient des informations démographiques (estimation de population par sexe, âge…) de chaque département de la France.

Vu que ce sont des données récoltées depuis l'année 1975, le fichier .xls contenant les données constitue tout un ensemble d'onglets chacun correspondant à une seule année (et un fichier .csv).

III.Méthode :

A. Les discussions sur le projet :

Dans un premier temps, on a envisagé de faire une analyse globale de l'éducation. Dans ce cadre, les plateformes API du PISA (Programme International pour le Suivi des Acquis des élèves) et du ISU (Institut de Statistique de l'Unesco) nous semblaient intéressantes.

Toutefois, nous nous sommes rapidement rendu compte que la différence entre les objectifs de ces instituts/programmes rendait très difficile la possibilité de croisement entre les jeux de données.

C'est ainsi que nous avons décidé de restreindre notre champ géographique (à la France), ainsi que la plage (2010-2015).

[3] https://www.insee.fr/fr/statistiques/1893198?fbclid=IwAR2vl7uAkWlyOQ9H-eMliTED5dYJMI7QVX61P3TSi01m1DSYXeCGXfSPZAo

B. Les difficultés rencontrées :

Notre premier problème fut à la récolte de données. En effet, même s'il existe une infinité de jeux de données en ligne, c'est ce qui fait que l'on s'y perd très rapidement. Parfois, nous pensions avoir nos bonnes sources, pour nous rendre compte qu'il n'y a pas vraiment matières à synthèse.

Par exemple, nous sommes parfois tombés sur des jeux comportant des données très intéressantes (du point de vue migratoire, des universités…), mais où la division départementale était absente. Ce fut notre principale contrainte de ce côté.

Le second problème fut celui de la sélection des données pertinentes. Nous avions l'embarras du choix.

Troisièmement, il nous fallait trouver un système de clés efficaces pour filtrer nos données, et ensuite pouvoir créer notre fichier final. Nous avons procédé ainsi :

- Entre le *JdD1* et le *JdD2* nous avons utilisé comme clef le tuple ("numéro de l'établissement", 'année').

- Entre le *JdD2* et le *JdD3* nous avons utilisé le ("nom du département") comme clé.

Enfin, le dernier problème, ou plutôt une observation concerne la partie codage. En effet, ce sont trois cerveaux pensant différemment qui doivent développer un seul et même script. Donc il a fallu à chaque relire nos codes et trouver un moyen de les harmoniser.

C. Filtrage de données :

On a procédé à faire un filtrage de données dans le but de ne garder que ce qui est commun entre les trois bases de données.

a) *Filtrage du JdD1* :

Dans sa version originale, ce jeu contenait une grande quantité de données, aussi bien en termes de lignes que de colonnes, ce qui a rendu nécessaire une opération de filtrage.

Nous avons donc décidé d'exploiter les données salariales des nouveaux diplômés, les domaines disciplinaires, ainsi que le taux d'insertion professionnelle et de chômage régionale.

Comme dit précédemment, nous avons utilisé comme clef entre ce jeu et JdD2 le tuple ("numéro de l'établissement", 'année').

La sélection des attributs a été effectué par le biais du module 'csv' de Python. Le script relatif au filtrage a été ensuite transformé en fonction pour être enfin inclus dans le module 'collecte_donnees' contenant les fonctions de filtrage de tous les autres jeux de données.

b) *Filtrage du JdD2* :

Pour ce jeu de données, vu que nous y avons accédé par API, il suffisait de ne questionner que les champs qui nous intéressaient.

Par la suite ce jeu de données fut notre fichier central, car c'était le seul à partager des données avec tous.

c) *Filtrage du JdD3* :

Pour ce jeu, nous avons choisi de nous concentrer sur la population de 20 à 39, tous sexes confondus.

Pourquoi cette tranche d'âge ? Elle nous semblait pertinente car, rappelons-le, nous voulions étudier l'employabilité des gens qui ont récemment terminé leurs études de master.

Vu que les informations associées à chaque année se trouvaient dans des fichiers .csv différents, nous avons dû parcourir les 6 bases de données concernées dans le but de les rassembler dans un seul document. Cette procédure a été initialement un peu compliquée à concevoir, mais on y est finalement arrivé à l'aide de notre camarade Ousseynou.

d) Implémentation :

Pour l'implémentation :

- Dans un premier temps, chacun a travaillé sur le bout de code qui interagit avec ses données (souvent en créant 1 ou 2 fonctions).
- Ensuite, nous avons pensé à une mise en commun selon l'étape du processus (les fonctions de génération, celle de sélection …).
- L'étape précédente a naturellement donné lieu à la création de modules (dont les noms sont assez explicites).
- Enfin, il ne restait plus qu'à mettre en place le main.py qui s'occupe des appels aux fonctions.

IV.Les données finales:

A l'aide de nos trois jeux de données, nous avons constitué une table résultante composée de données de différente nature (à savoir, des informations démographiques, éducatives et professionnelles).

Notre but, depuis le début de ce projet, était de confronter les données extraites et vérifier s'il existe (ou non) un rapport entre elles.

Pour ce faire, nous pourrions formuler des requêtes concrètes et ensuite les implémenter sur notre base. Voyons, donc, quelques exemples:

- Extraire le pourcentage d'hommes et de femmes sur chaque discipline. Vérifier si les pourcentages sont semblables dans chaque département ou s'il existe des disparités.

- Trier les données par département et vérifier si le taux d'insertion professionnelle monte, descend ou se maintient au fil de ces 6 années.

- Vérifier si les salaires bruts diffèrent en fonction de la discipline et/ou du département.

- Trier les données par année, les hiérarchiser en fonction de la stabilité d'emploi et les mettre en rapport avec le département et la discipline concernée.

V. Conclusion :

S'il y a bien une chose que nous aurions appris, c'est que la récolte et l'analyse des données sont deux choses bien différentes. Ainsi, nous avons déployé beaucoup d'efforts pour la première partie, en espérant que ce tableau puisse modestement servir aux analyste.