

08/12/2018



Projet Python
(IN3I19 – 3I-IN5)

Sarah CE-OUGNA

I- Introduction

Pour ce projet, j'ai choisi de représenter des données concernant les personnages de Comics (i.e. Superman, Batman, etc...).

J'ai dans un premier temps créé un fichier .csv que j'ai rempli de données récupérées automatiquement depuis « superheroapi.com ». Cette API présente des informations trouvées sur le site « Superhero Database ».

J'ai mis ce fichier csv dans le sous-dossier « data/preloaded » de mon programme, afin d'avoir une source de données sûre.

Il contient ainsi les informations (parfois incomplètes) de 728 personnages, dont : un ID, un nom, un nom de « civil » (i.e. Clark Kent pour Superman), l'éditeur (par exemple, Marvel ou DC), un lieu de naissance, l'URL d'une photo, et des colonnes représentant leurs scores dans six catégories : Intelligence, Force, Vitesse, Endurance, Pouvoir, et Combat.

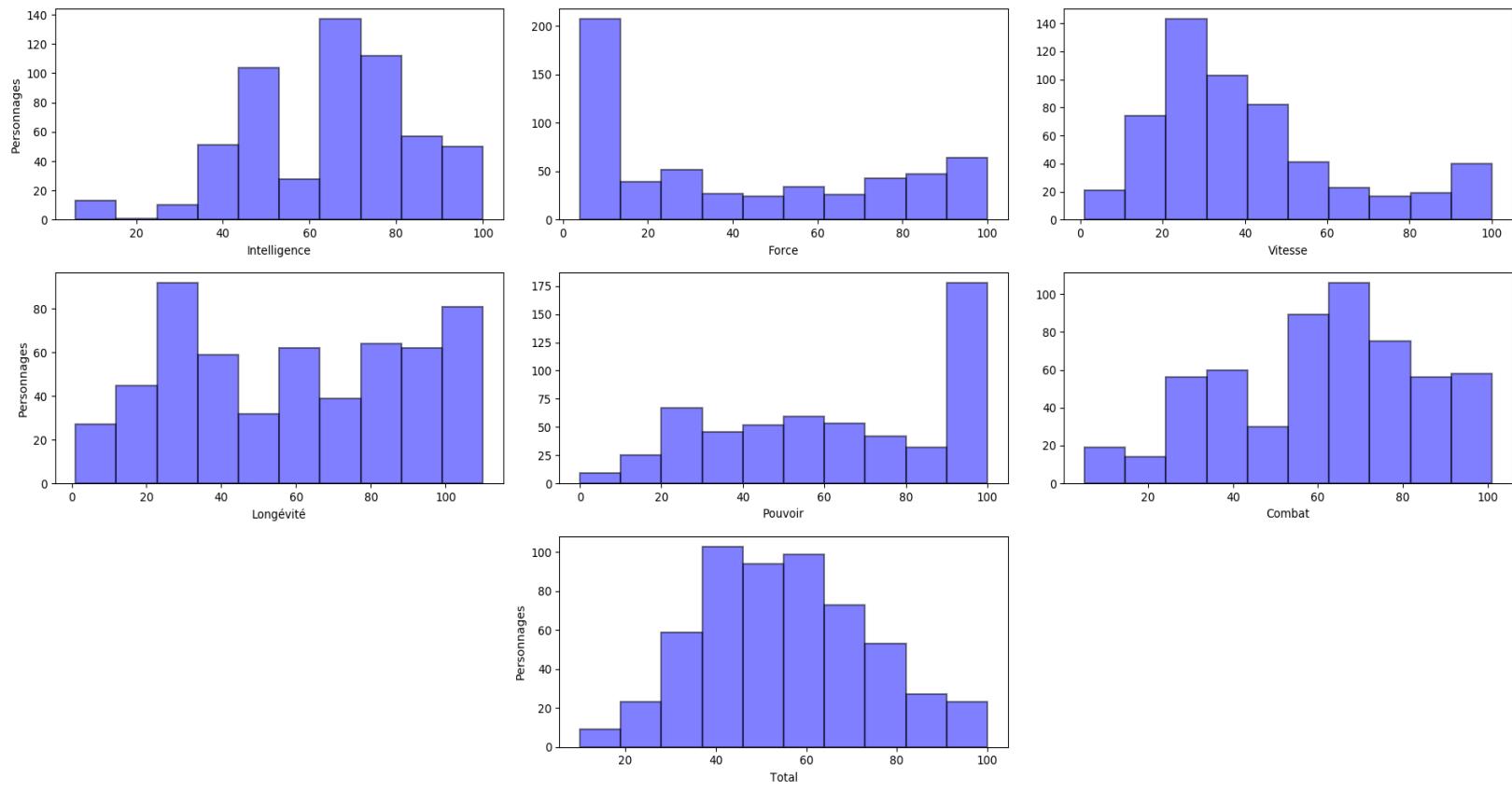
Ce sont sur ces données que se basent les histogrammes. Pour la géolocalisation, la colonne du lieu de naissance a été utilisée.

II- Histogrammes

Sept histogrammes sont présentés dans l'application, correspondant aux 6 catégories citées ci-dessus, ainsi qu'une de plus : Total.

Ce dernier histogramme présente le total des six scores fournis, afin d'avoir un score global pour chaque personnage.

Les données de 563 personnages ont pu être utilisées pour faire ces histogrammes, sur les 728 de départ. En effet, les lignes où les scores n'étaient pas renseignés ont été ignorées grâce à un pré-traitement des données.



Grâce à ces histogrammes, on peut ainsi voir que, contre toute attente, les scores ne sont pas aussi élevés que l'on pourrait s'y attendre. Prenons l'histogramme de force par exemple, il y a beaucoup d'individus dont la force est située entre 0 et 10 sur 100. La vitesse est elle aussi moindre que ce que l'on pourrait espérer.

Cela est dû au fait que les scores sont basés sur un grand nombre de personnages, dont beaucoup sont « humains », comme c'est le cas pour Batman. Ainsi, les scores sont plutôt déséquilibrés, avec les « humains » possédant beaucoup de 0 dans les catégories comme la force et la vitesse.

Les scores d'endurance, intelligence et combat sont assez irréguliers, alors que les scores pour la catégorie « pouvoir » sont plutôt élevés.

Toutefois, le dernier histogramme « total » présente une courbe gaussienne qui correspond plus à ce à quoi on s'attendait.

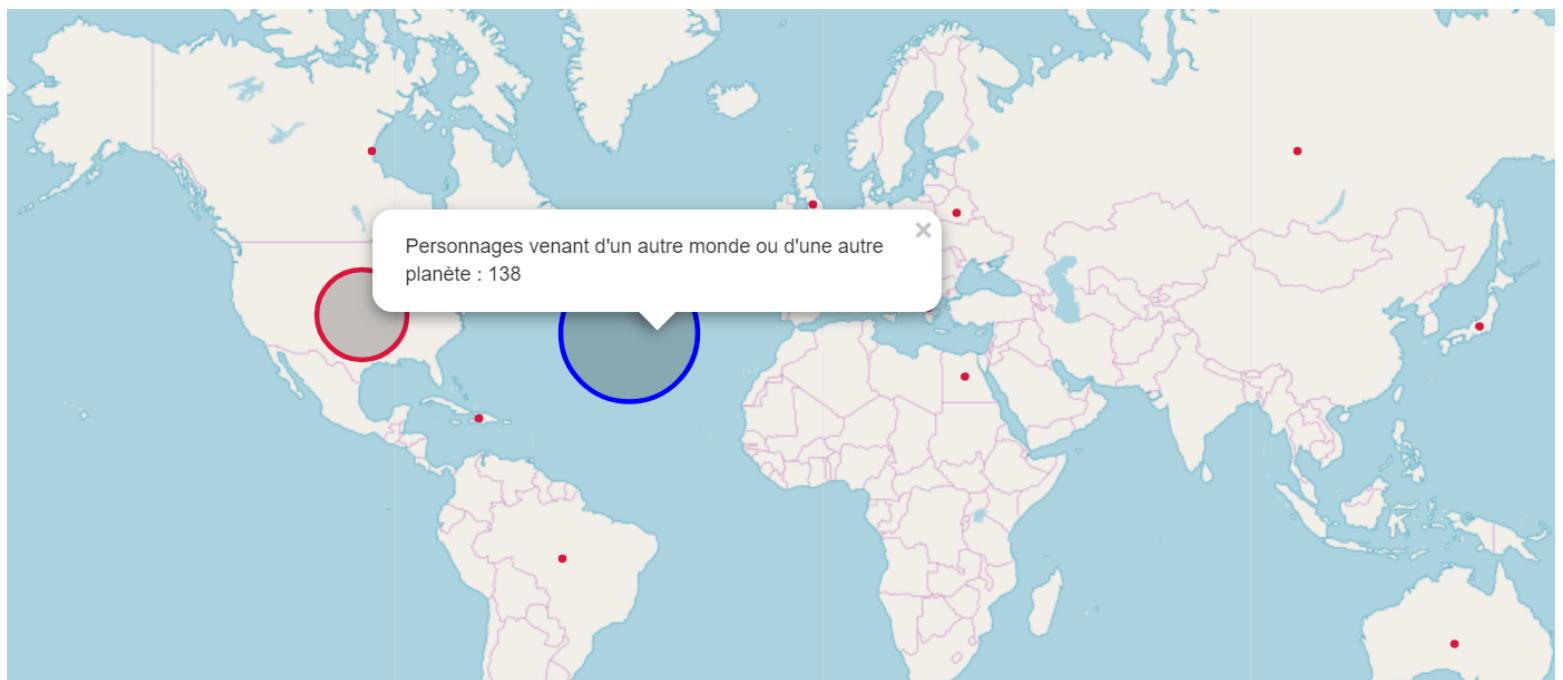
III- Cartes

Trois cartes ont été réalisées pour ce projet : une présentant les lieux de naissance de tous les personnages par des cercles individuels ; une autre contenant des cercles plus

ou moins large sur les pays selon le nombre de personnages y étant nés, et enfin une dernière carte présentant les personnages nés aux Etats-Unis.

Seulement 295 lignes sur 728 possédaient des informations sur le lieu de naissance. Grâce à un pré-traitement, j'ai pu identifier automatiquement 169 villes, et j'ai été forcée d'en rajouter 45 autres manuellement, car elles n'étaient pas reconnues. Il restaient encore 138 endroits non-identifiés, qui correspondent à des lieux imaginaires ou tout du moins, pas sur Terre.

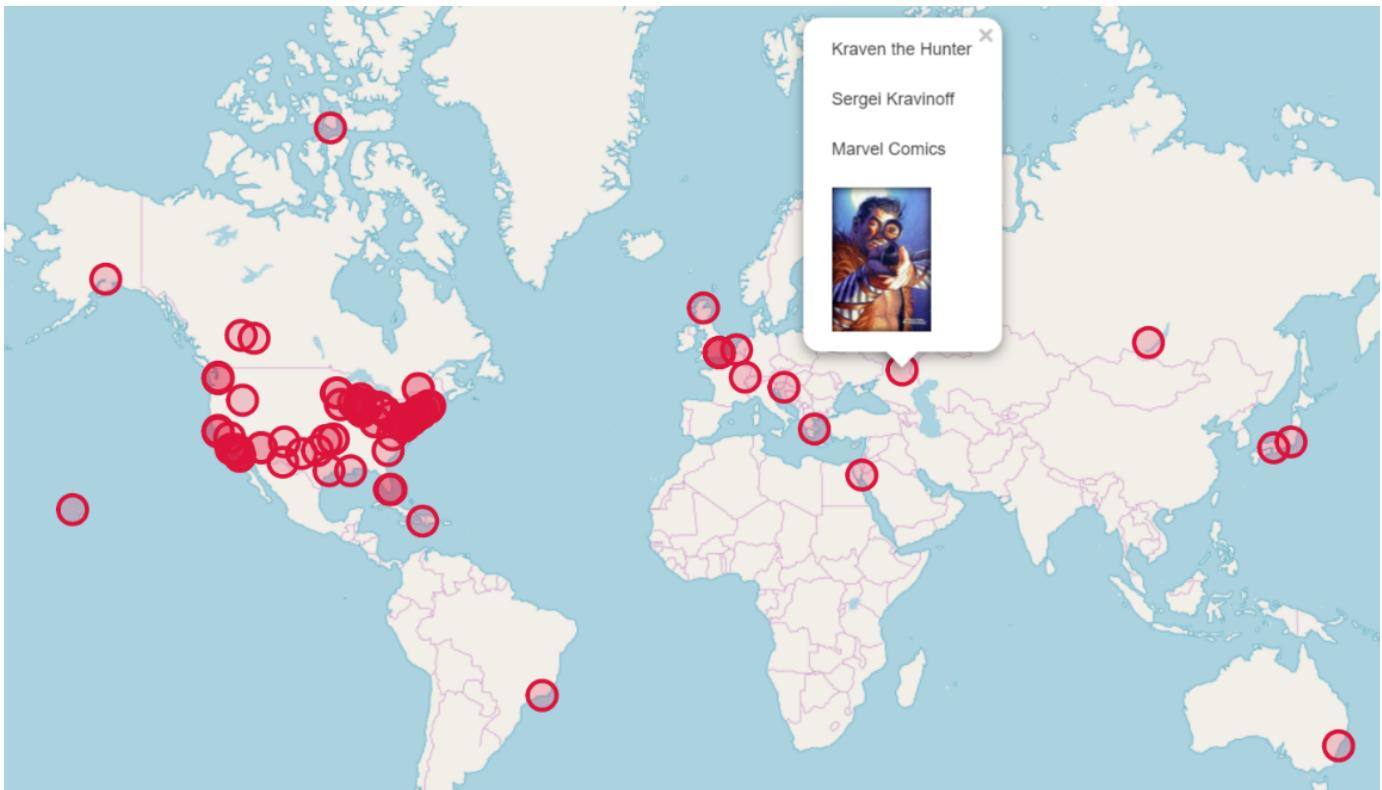
Ainsi, la carte représentant les fréquences par pays possède un cercle bleu, qui représente les personnages nés sur une autre planète ou dans un autre monde.



Les autres cercles, rouges, représentent des lieux réels, et on peut voir que les Etats-Unis sont, comme on pouvait s'y attendre, les seuls à être largement représentés.

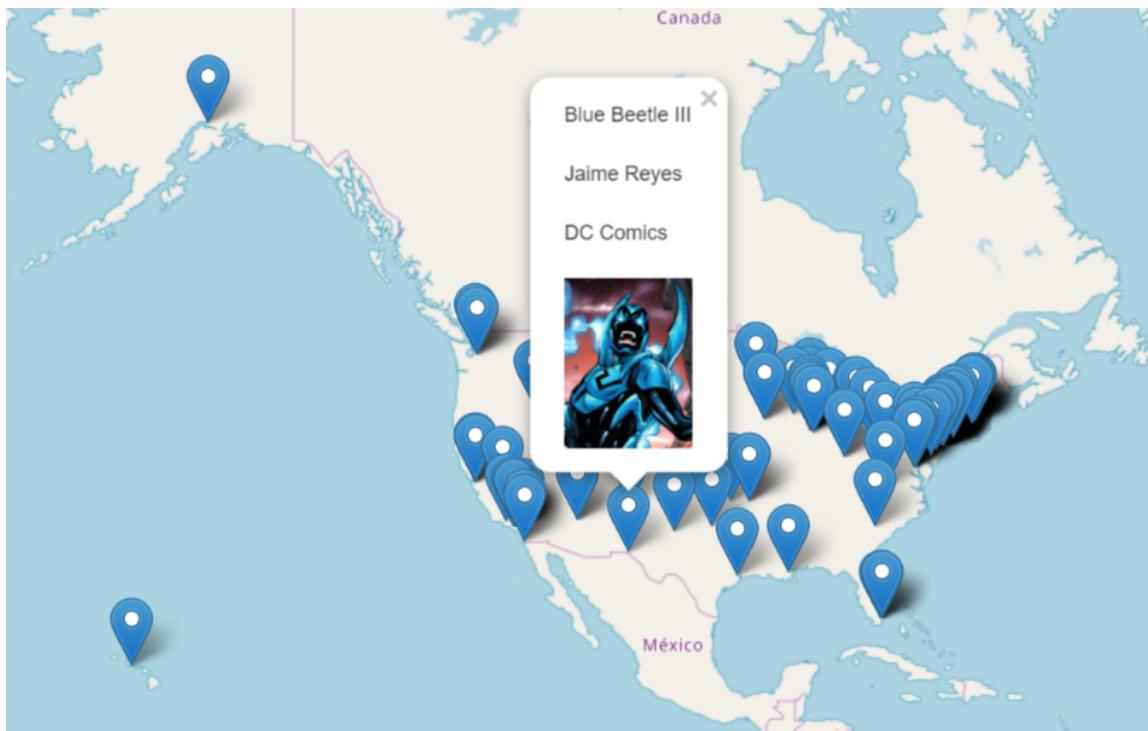
On peut même constater qu'il y a presque autant de personnages nés aux Etats-Unis que dans des endroits imaginaires. Si on clique sur un des cercles, cela affiche le total de héros né dans le pays correspondant (ou imaginaire pour ce qui est du cercle bleu).

Les deux autres cartes affichent les coordonnées individuelles des personnages :



Cette carte nous prouve bien que la grande majorité des personnages nés sur Terre sont nés aux Etats-Unis. Beaucoup des cercles se superposent dans les grandes villes américaines comme New York ou Chicago, car leurs coordonnées sont les mêmes.

De plus, si on clique sur un des cercles ou une des icônes, cela affichera le nom, l'éditeur et l'image du personnage.



IV- Fonctionnement du programme

Le programme peut être lancé avec le paramètre « --download » pour indiquer si l'on veut que les données soient téléchargées automatiquement.

Attention cependant, la récupération automatique des données prend près de 20 minutes ! En effet, chaque personnage se trouve à une URL différente, définie par son id.

```
730/731 retrieved!  
731/731 retrieved!  
Données récupérées ! Cela a pris: 1169.4794510036977 secondes, soit 19.491324183394962 minutes...
```

(On peut remarquer que le compte affiche 731 lignes, au lieu de 728. C'est en effet parce que les données d'ID 132, 173, 368 renvoient un json vide. Ces lignes ont donc été sautées.)

Par contre, si la commande est lancée sans le paramètre « --download », le programme se base sur les données présentes dans le fichier Superhero_data.csv, que j'avais préalablement rempli, et qui est situé dans le sous-dossier « /data/preloaded ».

Si une erreur se produit pendant le téléchargement (par exemple : non-réponse du serveur, qui s'est produite fréquemment), le programme se basera sur le fichier mentionné dans le paragraphe précédent.

```
136/731  
137/731  
Une erreur s'est produite pendant le téléchargement des données...  
Nous allons donc utiliser un fichier .csv pré-téléchargé
```

L'application possède une interface faite avec Tkinter, qui permet d'afficher les histogrammes directement, et/ou d'ouvrir les cartes dans un navigateur.

Avant de lancer l'interface, le pré-traitement des données est fait afin que l'affichage soit plus rapide. Ce pré-traitement consiste à ignorer les lignes ne possédant pas de données ou des données « null » pour les scores et les lieux de naissance, ainsi qu'à obtenir les coordonnées des lieux de naissance connus.

Les fichiers de code ont été séparés en trois : l'interface (Interface.py), les fonctions concernant le traitement des données (dataFunctions.py) et le fichier main qui les appelle (main.py).

V- Difficultés rencontrées

La plus grande difficulté rencontrée dans la réalisation de ce programme a été le pré-traitement des données pour la géolocalisation. En effet, les lieux de naissance étaient

rédigés de façon très irrégulière, avec parfois le nom de la ville, parfois le nom du pays, et parfois le nom d'un lieu imaginaire.

Il m'a donc fallu analyser longuement les données, puis en extraire les coordonnées automatiquement dans un premier temps, puis manuellement pour les données restantes (fichier manual_coordinates.csv). Cela m'a pris beaucoup de temps.

VI- Conclusion

Ce projet m'a beaucoup plu du fait que j'ai pu choisir un sujet que j'apprécie beaucoup. J'ai été étonnée par le nombre de personnages que j'ai pu récupérer (728 !) et par la quantité de données que l'on peut trouver à leur propos sur internet.

De plus, j'ai été surprise par le peu de personnages étant nés ailleurs qu'aux Etats-Unis. Je m'y attendais, mais voir le résultat en chiffres m'a quand même étonnée.