Rapport Projet - Ingénierie des Langues

Mohand Oukhemanou et Vincent Vilfeu

29 Mai 2023

Projet de Scrapping - Pokédex

Introduction

Dans le cadre de l'EC Ingénierie des Langues, nous sommes amenés à devoir réaliser un projet de scrapping afin d'extraire des données textuelles sur un ou plusieurs sites.
Nous avons alors décidé de créer un corpus reposant sur le principe du Pokédex, c'est-à-dire de créer une base de données comportant les pokémons des 4 premières générations ainsi que des informations à leur sujet.
Pour cela nous avons utilisé les bibliothèques BeautifulSoup et Requests pour la partie scrapping, les bibliothèques Pandas et CSV pour celle sur la base de données, les bibliothèques Functools et Operator pour travailler sur les listes et la bibliothèque Re(gex) pour travailler sur les données textuelles.
Dans la suite du rapport, nous expliquerons brièvement le fonctionnement de certaines des bibliothèques que nous avons utilisées, puis nous détaillerons le fonctionnement de notre code pour le scrapping des pages.

Bibliothèques

BeautfulSoup & Requests

La bibliothèque BeautifulSoup est une bibliothèque permettant de faire de l'analyse syntaxique de page HTML ou XML.
La bibliothèque Requests est une bibliothèque permettant de réaliser des requêtes HTTP de manière simple.
Ainsi en utilisant les deux ensembles, cela permet de faire du parsing HTML de façon simpliste. Pour cela, il faut également bien analyser le code source de la page en utilisant l'inspecteur (via le raccourci ctrl+shift+i). Dans le but de comprendre sous quelles formes se trouvent nos données dans le code source, pour bien les extraire.

Pandas & CSV

La bibliothèque Pandas est une bibliothèque permettant de manipuler et faire de l'analyse de données. Elle nous sert à stocker les données que l'on extrait des pages web.
La bibliothèque CSV est une bibliothèque permettant de lire et d'écrire des données au format CSV. Elle nous sert à mettre sous la forme de fichier csv (puis xmls) nos données.

Functools & Operator

Les bibliothèques Functools et Operator nous permettent de travailler sur les listes, notamment pour aplatir les listes que nous utilisons. Pour cela on utilise la fonction *reduce()* de Functools et la fonction *concat()* d'Operator.

Re(gex)

La bibliothèque Re est une bibliothèque permettant de travailler sur les expressions régulières. Elle nous sert à modifier les données extraites afin de rendre notre base de données plus lisible.

Nos Fonctions

class Pokemon

La *class Pokemon* contient les fonctions propres à un pokémon (soit à une donnée). Elle contient notre constructeur (avec toutes les caractéristiques que l'on veut récupérer sur les pages web), la fonction d'affichage d'un pokémon (et de ses caractéristiques), la fonction d'affectation des détails d'un pokémon et enfin la fonction d'affectation des stats d'un pokémon.

get_page(url, nom_poke)

Conclusion

En conclusion, au travers de notre projet et de notre rapport, nous avons pu voir différents types de clustering (que ce soit ceux basés sur les centroïdes, ceux sur la densité, ceux sur la distribution ou encore les hiérarchiques), avec pour chacun des résultats de partitionnement similaires.
Ainsi quant à notre hypothèse de départ qui était "L'espèce influe sur les caractéristiques physiques des pingouins.", nous pouvons la confirmer, puisque nos résultats de clustering sont très similaires à la dispersion en fonction de l'espèce.
À contrario, nous pouvons infirmer la seconde hypothèse qui était "L'habitat influe sur les caractéristiques physiques des pingouins.", puisque qu'aucun de nos résultats nous permet d'aboutir à une telle conclusion.
Aussi, pour aller plus loin, nous pourrions essayer de voir s'il y a des différences entre les pingouins mâles et femelles appartenant à la même espèce.