Fouille de Textes et Scraping

(M2 Informatique La Rochelle/Niort)

2021-2022

TD2 Web Scrapping sur Données Hétérogènes

Gaël Lejeune et Jean-Baptiste Tanguy, Sorbonne Université

Vous rendrez (un rendu par groupe) votre programme d'extraction (pas tous les Html, un échantillon suffira), les fichiers Json générés et le programme pour requêter.

Exercice 1 : Objectifs

La Fédération Française des Echecs organise un certain nombre de compétitions. On vous contacte pour faire des statistiques sur les tournois individuels.

Les informations sur un tournoi sont rangées dans des fiches accessibles par l'url "`http://echecs.asso.fr/FicheTournoi.aspx?Ref=X`" où X est un identifiant arbitraire de base de données obtenu par auto-incrément. Un exemple : "`http://echecs.asso.fr/FicheTournoi.aspx?Ref=28123`

Nous allons chercher à extraire les informations sur ces tournois et à les structurer en JSON.

Pour simplifier nous allons travailler sur un échantillon en rpenant tous les tournois dont l'ID est compris entre 30.000 et 30.500 ($30000 <= x <= 30500$). Pour chaque identifiant XXX de l'intervalle nous allons donc récupérer :

1. Le descriptif du tournoi: `http://echecs.asso.fr/FicheTournoi.aspx?Ref=XXX`
2. Les statistiques de participation: `http://echecs.asso.fr/Resultats.aspx?URL=Tournois/Id/XXX/XXX&Action=Stats`
3. La liste des participants: `http://echecs.asso.fr/Resultats.aspx?URL=Tournois/Id/XXX/XXX&Action=Ls`
4. La grille des résultats: `http://echecs.asso.fr/Resultats.aspx?URL=Tournois/Id/XXX/XXX&Action=Ga`

On veut pouvoir donner facilement des informations telles que :

- Tous les tournois auquel un joueur a participé
- Un classement des joueurs par rapport à leur nombre de tournois gagnés (terminés à la première place selon la grille de résultats)
- Un classement des joueurs qui jouent le plus sur une période
- Les clubs les plus actifs dans les tournois

Exercice 2 : Extraction d'Informations

Vous allez créer un programme qui va parser les données récupérées ci-dessus va extraire et structurer dans un fichier JSON:

1. Dans un fichier `tournois.json`: Pour chaque tournoi, toutes les informations du descriptif : Nom, Localisation, Dates, Prise en compte... [1]

2. Dans un fichier `stats.json`, pour chaque ID de tournoi, les statistiques (réparttion par titre, par catégorie d'âge ...)

3. Dans un fichier `participants.json`, pour chaque participant de chaque tournoi : Nom, Classement Elo (4 chiffres), Type de classement (la lettre figurant à côté F, N ou E), catégorie, sexe, liste des participations en tournois ... Identifiez le participant de manière unique par une simple concaténation de son nom et de son prénom (vous verrez que c'est une approximation).

4. Dans un fichier `resultats.json`, pour chaque ID de tournoi, la liste des joueurs et pour chaque joueur sa place finale, son nombre de points et la liste de ses résultats (adversaire rencontré, couleur, résultat)

Exercice 3 : Requêtage

Prévoyez maintenant une manière simple de requêter vos données pour obtenir les informations recherchées plus haut (les tournois auxquels un joueur a participé ...).

Puis :

- Passez à l'échelle en scrappant tous les tournois depuis $ID = 1$
- Donnez des statistiques de participation (nombre de tournois, de joueurs, de parties jouées) par année civile puis par saison sportive (du 1/09 au 31/08)

[1] Sous la forme d'un tableau associatif (dictionnaire en Python) où chaque tournoi est identifié par son ID auquel on associe un autre tableau associatif où les clés sont les noms des informations