RER – RegEx et Scraping

Contexte :

Appréhender les notions de RegEx et de web Scraping 😉

Problématique(s) :

Quels sont les avantages et les inconvénients des regex dans le web Scraping ?

Comment manipuler le concept de RegEx dans le nettoyage des données ?

Comment Scraper un site ?

Mots clés :

webScraping : (Harvesting, grattage) Technique automatisée qui permet d'extraire des données aux format exploitable depuis des pages web

RegEx : Les **expressions régulières** (de l'anglais *regular expressions*) sont des chaînes de caractères sur la base de règles syntaxiques permettant de décrire des **séquences de caractères**. Elles font de fait partie des langages rationnels – un sous-groupe des langages formels d'une grande importance notamment en informatique, et plus spécifiquement dans le développement de logiciels.

Parsing : Analyse syntaxique fait de chercher des données spécifiques à l'aide d'un parser HTML / XHTML (lxml, html.parser, Html5lib)

Parser	avantages	désavantages
html.parser	Inclus dans python (pas de bibliothèque externe à installer) Assez rapide Indulgent	Moins rapide que lxml Moins indulgent que html5lib
Parser html de la bibliohtèque lxml	Très rapide indulgent	Dépendance de C
Parser xml de la bibliohtèque lxml	Très rapide Le seul parser XML supporté par beautiful soup	Dépendance de C
Html5lib	Très indulgent Parse les pages web de la même manière qu'un navigateur web Crée du html5 valide	Très lent Bibliothèque externe

Beautiful Soup : C'est une bibliothèque Python d'analyse syntaxique de documents HTML et XML. Elle produit un arbre syntaxique qui peut être utilisé pour chercher des éléments ou les modifier. Il transforme un document HTML complexe en un arbre d'objets Python. Il convertit aussi automatiquement le document en Unicode, de sorte à nous éviter l'étape d'encodage. Elle a besoin de parser (lxml ou Html5lib).

Scrapy : est un framework Python open-source pour extraire les données des sites Web. Il est rapide car les requêtes sont effectuées de manière asynchrone et extensible.

Crawler : (bot d'indexation) est un programme informatique qui explore automatiquement le web en suivant les liens hypertextes pour collecter des informations à des fins d'indexation ou d'analyse.

Transmission synchrone / asynchrone :

Le mode de transmission désigne le nombre d'unités élémentaires d'informations (bits) pouvant être simultanément transmises par le canal de communication.

Base de comparaison	Transmission synchrone	Transmission asynchrone
Sens	Envoie des données sous forme de blocs ou de cadres	Envoie un octet ou un caractère à la fois
Vitesse de transmission	Vite	Lent
Coût	Coûteux	Économique
Intervalle de temps	Constant	Au hasard
Écart entre les données	Absent	Présent
Exemples	Salons de discussion, vidéoconférence, conversations téléphoniques, etc.	Lettres, courriels, forums, etc.

Twisted : c'est un framework python qui permet de créer facilement des serveurs smtp, http, proxy et ssh. Twisted est asynchrone et « event-driven » et permet aux applications de répondre à différentes connexions sans utiliser de threads (processus parallèles). Il est utilisé par Scrapy.

Selenium : Framework qui permet le scraping il est plus généraliste et permet d'automatiser la navigation web, tester les application web et prend en charge différents langages. Il permet de lire le contenu dynamique. (JS)

Splash : Navigateur web implémenté en python qui utilise Twisted qui permet de faire du scraping de javascript et js rendering service

Requests : Module python qui simplifie les requêtes http vers les serveurs web

PySpider : System de webCrawling en python, il offre une interface web pour cela.

Hypothèses :

1. Adrien : On prononce RejEx et pas RegEx --> Faux

1bis. Adrien : On peut cliquer sur un bouton d'un site --> Vrai

2.Osman : Le web scraping fait partie de l'OSINT --> Vrai

3.Solenn : Le scraping peut être utilisé pour récolter les données sur les sites non répertoriés --> Vrai ou privés --> Faux

4.Jean Paul SOSSAH : Le scraping peut inclure des requêtes SQL --> Faux

5. Aude : Les RegEx permettent de faire du parsing --> Vrai

6. Etienne : le web Scraping peut être bloqué au-delà d'un certain nombre de requêtes par IP --> Vrai

7. Adeline 1 : Un bon web Scraping peut intégrer les données dans une BDD --> Vrai

8. Adeline 2 : Il est préférable de faire du scraping pour faire un site web / E-commerce --> Vrai

9. Briand : Les techniques de Scraping peuvent être utilisé pour collecter des données en temps réel --> Vrai

10. Briand : Le scraping peut être utilisé pour tester la sécurité, la vulnérabilité d'un site --> Faux

11.Tetyana : Les techniques anti-scraping empêche le SEO / SEA (Faux mais elles peuvent parfois affecter négativement ces pratiques en limitant l'accès aux données de référencement)

12.Axel : Les extensions de type : AdBlock peuvent empêcher le web scraping --> Faux

13.Seydou : Le scraping peut être utilisé de manière non éthique --> Vrai

14.Seydou : Le RegEx permet de bloquer l'injection SQL --> Vrai

15. Loïc : Impossible de faire du Scraping sans les RegEx --> Faux

16. Loïc : La RegEx ne s'utilise pas uniquement pour du scaping mais entre autres pour mettre en place une couche de sécurité sur un site --> Vrai

Plan d'action :

- Explorer les ressources
- Définir les mots clés
- Répondre aux problématiques
- Faire les workshops
- Rendre les livrables : RER et workshops

Critère	Scrapy	BeautifulSoup
Fonctionnalités	- Extraction de données structurées et non structurées buivi de liens (Crawling) Intégration de pipelines de traitement de données Gestion de session et d'authentification	- Extraction de données structurées et non structurées - Manipulation de l'arbre DOM - Analyses HTML/XML Gestion de session et d'authentification
Facilité d'utilisation	Scrapy nécessite une certaine courbe d'apprentissage en raison de son architecture complexe et de ses fonctionnalités avancées.	BeautifulSoup est plus facile à prendre en main car elle utilise une syntaxe simple et directe pour accéder aux éléments HTML.
Performance	Scrapy est conçu pour le scraping de sites Web à grande échelle et peut gérer de manière efficace des milliers de pages en peu de temps.	BeautifulSoup peut être plus lent que Scrapy en raison de sa structure de traitement en série.
Prise en charge de JavaScript	Scrapy ne prend pas en charge JavaScript.	BeautifulSoup ne prend pas en charge JavaScript.
Flexibilité	Scrapy est très flexible et peut être personnalisé pour s'adapter à une grande variété de projets.	BeautifulSoup est moins flexible et est mieux adapté pour des projets plus simples.