Formation

Collecter des données sur le web : Python pour le webscraping

Partant du constat souvent répété de l'accumulation croissante des données produites chaque année, dont une part importante circule sur le web, cette formation vise à introduire des notions élémentaires de **structuration, formatage et traitements de corpus de données**, étapes préalables et néanmoins essentielles à tout travail d'analyse.

Il s'agira de travailler particulièrement à partir des **données du web** puisque celui-ci constitue aujourd'hui une source d'information très riche et utile dans le cadre de travaux de recherche, qu'il s'agisse d'interroger les dynamiques des sociétés contemporaines (à travers l'étude de corpus de données extraits des réseaux sociaux, par exemple), la production de savoirs et l'émergence de controverses (à travers l'étude des échanges sur les pages Wikipédia, par exemple), la formation d'opinions publiques (à travers l'étude de corpus de presse), etc. Les outils de collecte et de traitement de données sont également utiles pour tout travail à partir de corpus de données volumineux, trop importants pour être parcourus de manière cursive ou traités "à la main".

Ce que cette formation n'est pas : Il ne s'agit pas d'une formation d'analyse de données textuelles, ni d'analyse quantitative. Elle s'arrêtera à présenter les différents environnements techniques nécessaires à comprendre pour pouvoir **collecter et exploiter** des données issues du web. Ces compétences pourront être déclinées pour traiter de grands jeux de données textuelles extraites d'autres supports (lots de fichiers PDF, texte brut, etc.).

Objectifs

- Comprendre l'environnement technique du web, les principes de bases et la façon, notamment, dont ses formats spécifiques permettent des traitements utiles à la sélection et au filtrage de données en amont d'opérations de collecte
- Comprendre les formats de fichiers structurés et les façons d'interagir avec eux
- (structures de données)
- Comprendre les briques élémentaires de l'algorithmie pour construire des scripts utiles pour la collecte et le nettoyage de données
- Interroger des pages web et récupérer leur contenu filtré/nettoyé
- Mettre au clair les aspects éthiques et légaux du webscraping

Intervenants

Romain Mularczyk, ingénieur d'études en gestion de données (Univ. Lyon 2, MSH Lyon St-Etienne)
Agathe Déan, statisticienne (CNRS, MSH Lyon St-Etienne)

Public

Ces ateliers sont ouverts à tous les personnels (chercheurs, enseignants-chercheurs, ingénieurs et techniciens, doctorants) membres des laboratoires associés à la MSH Lyon St-Etienne.

Prérequis

Ce cycle de formation s'adresse à un public "grand débutant", n'ayant aucune expérience ou connaissance particulière des technologies du web ni de programmation informatique.

Méthode

Cette formation s'appuie sur une approche qui met en avant la mise en pratique des concepts étudiés et l'expérimentation.

Elle privilégie pour ce faire un **outil de notebook** propre à l'environnement **Python, Jupyter**.
Cet outil libre et gratuit permet d'imbriquer du code Python exécutable et du texte permettant de décrire les étapes du code ou d'imbriquer des éléments de réflexion et d'analyse à l'aspect plus technique du seul code. L'avantage d'un tel outil réside dans l'interactivité qu'il permet : celui-ci est composé de cellules individuelles qui peuvent contenir des petites briques de code actionnables individuellement. Il permet ainsi un raisonnement incrémental qui facilite la découverte de la programmation et la construction pas à pas d'un raisonnement formel.

Format

Un cycle de 6 à 7 sessions d'une journée entière, permettant de prendre le temps, en amont (matinée), de poser les concepts essentiels, et de passer à la pratique dans un second temps (après-midi) en travaillant à la fois sur des exercices élémentaires, puis en élaborant à mesure des sessions, son propre projet personnel pour permettre d'appliquer les concepts à différents cas pratiques.

Pour pouvoir accompagner chaque participant, il s'agirait de restreindre le groupe à un maximum d'une dizaine de personnes.

Dates, lieu

Dates :
11 janv 2021 : Le web
18 janv 2021 : Introduction à Python 3
25 janv 2021 : Python 3 (suite)
1er fév 2021 : Rappels et introduction au webscraping
8 fév 2021 : Application au webscraping
15 fév 2021 : Concepts avancés et conclusion
22 fév 2021 : Pratique

Lieu : en raison de la situation sanitaire actuelle, les sessions seront organisées en distanciel/visioconférence.

Inscription

L'inscription à ce parcours de formation est gratuite mais obligatoire (10 personnes maximum).
Merci de vous inscrire **avant le 4 janvier 2021**, en remplissant le formulaire en ligne :
https://enquetes.msh-lse.fr/index.php/718827/lang-fr

Contacts

Contenu des sessions, organisation :
Romain Mularczyk (MSH Lyon St-Etienne) : romain.mularczyk@msh-lse.fr
Agathe Déan (MSH Lyon St-Etienne) : agathe.dean@msh-lse.fr

Programme

Session 1 - Le web

Lundi 11 janvier 2021 (10h-18h)

- Introduction aux technologies du web
  - Documents structurés et leur format (CSV, HTML, XML, etc.)
  - Langages à balises (HTML/XML)
  - Structures arborescentes et leur parcours (DOM)
  - Environnement client-serveur et requêtes HTTP
- Pratique autour de la structuration d'un document HTML

Session 2 - Introduction à Python 3

Lundi 18 janvier 2021 (10h-18h)

- Introduction à la programmation avec Python 3
    - Variables
    - Types de données
    - Boucles
    - Structures conditionnelles
    - Fonctions
- Pratique à partir d'exercices sur les opérations de base en Python

Session 3 - Python 3 (suite)

Lundi 25 janvier 2021 (10h-18h)

- Bases de la programmation orientée objet
- Suite et méthodes de sélection et de nettoyage
    - Ouvrir, écrire dans des fichiers
    - Expressions régulières
- Pratique à partir d'exercices et applications éventuelles à des projets personnels

Session 4 - Rappels et introduction au webscraping

Lundi 1er février 2021 (10h-18h)

- Rappels des notions de programmation orientée objet (boucles, structures) et pratique
- Retour sur les expressions régulières
- Introduction au webscraping

Session 5 - Application au webscraping

Lundi 8 février 2021 (10h-18h)

- Construire des requêtes HTTP avec Python
    - Bibliothèque `requests`
- Scraping et parsing d'une page web
    - Bibliothèque `BeautifulSoup`
- Pratique à partir d'exercices sur l'extraction de données de Wikipédia et de Twitter et application à des projets personnels

Session 6 - Concepts avancés et conclusion

Lundi 15 février 2021 (10h-18h)

En fonction du programme couvert, cette dernière session pourra donner lieu à la présentation de concepts avancés ou d'approfondissement des concepts précédents. Il se conclut dans tous les cas sur les aspects éthiques et légaux des usages du webscraping.

- Interroger une API
    - Parcourir et comprendre la documentation d'une API
    - Autres exemples d'utilisation de la bibliothèque `requests`
- Concepts avancés
    - Requêtes asynchrones (`Selenium` pour le scraping de données générées via JavaScript)
- Découverte du framework `Scrapy`
- Aspects éthiques et légaux du webscraping

Session 7 – Pratique

Lundi 22 février 2021 (10h-18h)

Application aux projets personnels des participants.