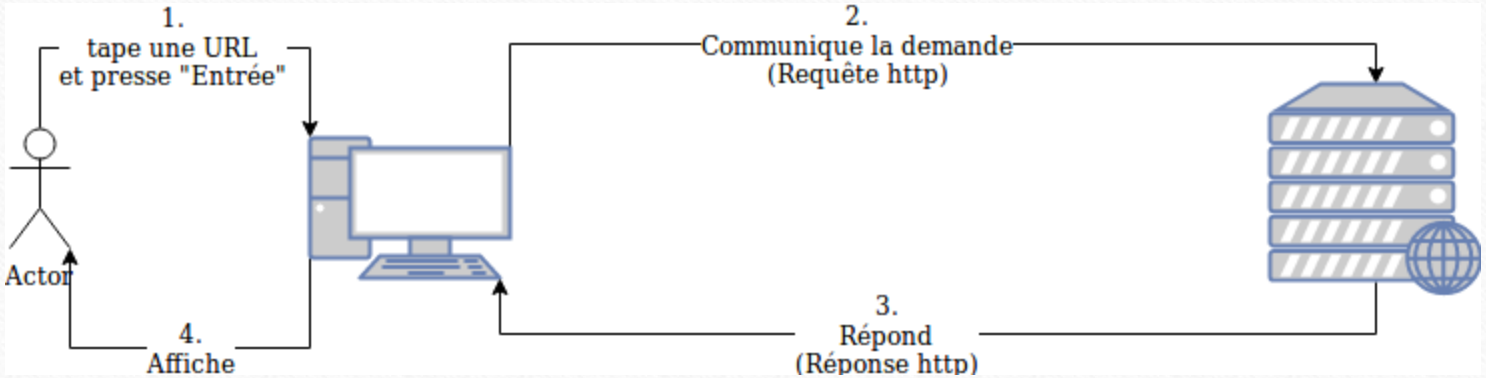
Chapitre 4 :
Initiation au *scraping*

Virgile Reignier

Master 2 TNAH – 2023-2024

École nationale des chartes - PSL

Requête HTTP



Composition d'une réponse :

- Code HTTP (200, 404, …)
- Contenu (page web, ressource, …)
- Headers (métadonnée décrivant la taille, l'encodage, …)

HTTP et python

- Observation de la page web http://www.memoire-ardeche.com/cahiers/129.htm
- Python peut envoyer des requêtes et recevoir des réponses
- Un package dédié : https://fr.python-requests.org/en/latest/

Exercices :

- Afficher sur son terminal le code de la page https://www.chartes.psl.eu/

HTML et python

- Observation du code source de la page http://www.memoire-ardeche.com/cahiers/129.htm
- Python peut analyser les langages à balises
- Un package dédié : https://www.crummy.com/software/BeautifulSoup/

Exercices :

- Affichez la description du formulaire en bas de page
- Fin de chapitre : à partir de http://www.memoire-ardeche.com/cahiers/table.htm, produire un fichier csv résumant, pour chaque cahier de l'association : le numéro, la date, le titre, l'url et le titre de la page correspondant à ce numéro.