**Veille sur le Web Scraping**

Objectif de la veille : réunir des informations sur un sujet spécifique dans le but de faire une présentation orale devant la classe.

Les 5 projets :

Chaque projet contient des pistes à explorer, cependant, vous êtes libres d’aborder des points supplémentaires si vous les jugez pertinents.

* ***Web scraping* : définitions, applications et cadre légale :** 2 personnes.
  + **Objectif : Comprendre l’intérêt du web scraping et ses limites**
  + Pistes à explorer :
    - Qu’est-ce que le *web scraping* ? ;
    - Fixer le cadre légale ;
    - Citer des exemples d’utilisation commerciale ;
    - Expliquer la différence entre *scraping* et *crawling ;*
    - Définir les 2 grandes étapes du scraping : *fetching* content et *parsing* content ;
    - Définir spider.
* **HTML & CSS** : 3 personnes.
  + **Objectif : Comprendre la structure du langage, son utilisation et ses bonnes pratiques**
  + Pistes à explorer :
    - Définir HTML et CSS, relation, versions actuelle ;
    - Quelles sont les différences entre HTML et python ? Leurs utilisations respectives ? ;
    - Les bonnes et mauvaise pratiques de HTML ;
    - Les bonnes et mauvaise pratiques de CSS ;
    - Définir la notation XPATH ;
    - Quelles sont les différentes manières de déclarer une couleur dans CSS ? Expliquer leur fonctionnImageement ;
    - Définir le CSS *Specificity.*
* **Les outils du web scraping :** 2 personnes.
  + **Objectif : faire l’état des lieux des outils en python pour effectuer du web scraping**
  + Pistes à explorer :
    - Lister et expliquer (succinctement) les principaux packages de client http en python (Urllib, requests, etc.) ;
    - Focus sur Requests (utilité, avantages/inconvénients, exemple) ;
    - Focus sur BeautifulSoup (utilité, avantages/inconvénients, exemple) ;
    - Focus sur Scrapy (utilité, avantages/inconvénients, exemple).
* **Le stockage en BDD :** 2 personnes.
  + **Objectif : Comprendre les différentes possibilités pour exporter des données**
  + Pistes à explorer :
    - Définir les formats CSV, JSON, XML;
    - La différence entre le SQL et le NoSQL;
    - Qu’est-ce que MongoDB ? Ses avantages ? Son application ?;
    - Définir le Big Data : les 3 V, les enjeux;
    - Qu’est-ce que le cloud ? Ses principaux acteurs.
* **Le fonctionnement du web :** 2 personnes.
  + **Objectif : Avoir une connaissance globale du fonctionnement du web**
  + Pistes à explorer :
    - Définir Internet et le World Wide Web ;
    - Qu’est-ce que le HTTP ? Le HTTPS ? ;
    - Définir l’architecture Client/Serveur ;
    - Développer la notion de *request* (GET, POST, etc.) et *response* (404, 200, etc.) ;
    - Définir le Document Object Model ;
    - Protocole TCP/IP, SSL ;
    - Les ports par défaut.

La présentation :

* **Objectifs de cet exercice :**
  + Savoir transmettre vos connaissances et vulgariser un sujet ;
  + Être capable d’écouter et de retenir des informations ;
  + Produire un support de présentation efficace ;
  + Respecter une contrainte temporelle.
* **A propos du support visuel :**
  + Créer une mise en page propre ;
  + Ne pas hésiter à inclure des visuels (graphiques, schémas, etc.) ;
  + Inclure le numéro de diapo ;
  + Mettre un titre aux visuels.
* **Plan à respecter :**
  + Introduction (1 slide) ;
  + Sommaire (1 slide) ;
  + {Partie libre en fonction du sujet} ;
  + Conclusion (1 slide) ;
  + Difficultés rencontrées / Pistes d’approfondissement (1 slide).
* **Livrable :**
  + Support de présentation (powerpoint).
* **Format de la présentation :**
  + 10 minutes de présentation ;
  + 5 minutes de questions-réponses.