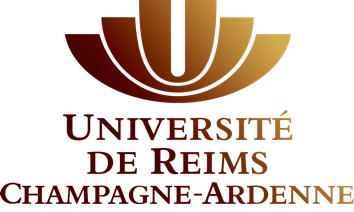
LIEPO Brice-Kevin - ROA SERRANO Walter

Résumé

Projet réalisé dans le cadre du cours INFO0801  
Modélisation connaissances   
& humanités numériques



theme : Realisation d’un jeu de données

Web Scraping – API’s

Table des matières

[I. INTRODUCTION 2](#_Toc32834426)

[II. CONTEXTE 3](#_Toc32834427)

[III. ENVIRONNEMENT DE TRAVAIL 4](#_Toc32834428)

[IV. BESOINS ET OBJECTIFS DU PROJET 4](#_Toc32834429)

[V. DÉVELOPEMENT TECHNIQUE 4](#_Toc32834430)

[VI. PRÉSENTATION DE DONNÉES 5](#_Toc32834431)

[VII. BILAN DU PROJET 6](#_Toc32834432)

# INTRODUCTION

Dans de nombreux domaines comme l’industrie, le secteur de la finance ou la vente au détail l’extraction de données est le processus d’exploration de vastes ensembles de données visant à trouver des renseignements pertinents qui pourraient être utilisés dans un but précis.

Une fois que les données ont été collectées et stockées, l’étape suivante consiste à donner du sens aux données. Autrement le processus est inutile. L’extraction de données est utilisée à de multiples fins, en fonction de l’entreprise et de ses besoins.

Le web représente une de plus importantes sources d’information où il est possible l’acquisition de données dans les contenues des sites ou des pages web publiés dont l’utilisation de web scraping permettre d’exploiter la structuration des langages des balises du web et aussi les APIs qui ont augmenté la capacité et la vitesse d’accès à l’information.

Dans le cadre de notre première année de master en informatique parcours Intelligence Artificielle à la faculté des Science naturels et exactes de Reims, il nous est proposé de mettre en pratique nos connaissances et nos compétences professionnels au travers d’un projet ayant pour objectif principal l’utilisation des ces deux méthodes pour l’extraction et la construction d’un jeu de données.

# CONTEXTE

C’est quoi le Web Scraping ?

Le Web Scraping est l’un des moyens le plus robuste et fiables d’acquérir le contenu d’un page web publiée sur internet, cette technique consiste en convertir des données non structuré (balises HTML) vers un format structuré facilement accessible et utilisable : par exemple la construction de bases de données.

Une autre façon d’obtenir des informations via requêtes http c’est à travers des APIs.

Une API permet de rendre disponibles les données ou les fonctionnalités d’une application existante afin que d’autres applications les utilisent.

Utiliser une API permet donc d’utiliser un programme existant plutôt que de le redévelopper pour avoir accès à des informations. Autrement dit, elle rend disponibles des fonctionnalités ou des données. Pour les utiliser, la plupart des API requièrent d’une clé (API key). Cette clé permet à l’API de vous identifier comme étant un utilisateur ayant les droits nécessaires pour se servir de l’API.

Une des principales fonctionnalités qu’on retrouve quand on utilise une API est l’exposition de données. Cela signifie que les contenues d’un site (pages, articles) sont accessibles grâce à des end points, dans un format de données structurées. Concrètement, en se rendant à une certaine URL vous aurez les données d’un site au format JSON.

# ENVIRONNEMENT DE TRAVAIL

LOGICIEL R

R est un langage de programmation pour statistiques et graphiques par ordinateur dont le but est de pouvoir traiter, organiser et analyser des volumes importants de données de manière rapide et flexible afin de pouvoir y appliquer des tests statistiques et se représenter ces données graphiquement à l’aide d’une grande variété de libraires disponibles. R permet d’importer différentes librairies qui mettent à disposition des puissantes fonctionnalités qui facilitent l’extraction des données.

PACKAGES ET LIBRARIES INSTALLÉS

Tidyverse : C’est une collection d’extensions conçues pour aborder un grand nombre d’opérations courantes dans R comme :

* Visualisation
* Manipulation des tableaux de données
* Import / Export de données
* Manipulation de variables
* Extraction de données du web
* Programmation

L’objectif de ces extensions est de fournir des fonctions avec une syntaxe cohérente, qui fonctionnent bien ensemble, et qui retournent des résultats prévisibles.

Xml2 : Ce paquet facilite le travail au moment de travailler avec HTML ou XML, on l’utilise pour la lecture de l’url contenant le fichier HTML dont le web scraping a été réalisé.

Rvest : Est un paquet qui permet de chercher le contenu d’une page web, pour le rendre exploitable par R ensuite d’extraire ce contenu à l’aide de la syntaxe XPath ou des sélecteurs CSS. Surtout on a utilisé les sélecteurs CSS pour l’extraction de données provenant de différents sites web qu’on a utilisé pour extraire les pays et les capitaux.

Jsonlite : Le paquet offre des outils flexibles, robustes et de haute performance pour travailler avec JSON dans R il est particulièrement puissant pour créer des pipelines et interagir avec des APIs Web. Il a été utilisé pour générer rapidement les fichier JSON provenant de l’API.

OPENWEATHERMAP : Il est un service en ligne qui fournit un API qui rendre des données météorologiques en temps réel aux développeurs de services web et d’applications mobiles. Elle fournit des endpoints JSON d’utilisation gratuite avec lequel pour notre projet, on a récupéré des informations telles que les points cardinaux, la température, l’humidité et la description du temps.

# BESOINS ET OBJECTIFS DU PROJET

OBJECTIFS TECHNIQUES

OBJECTIFS ACADEMIQUES (LA RÉALISATION DU JEU DE DONNÉES POUR VISUALISATION DE DONNÉES)

# DÉVELOPEMENT TECHNIQUE

LA STRATÉGIE

LES FONCTIONS

LES JOINTURES

RÉALISATION DE TIBBLE (CONVERTIR LA DATAFRAME FINAL EN TIBBLE COMME IL A DIT)

# PRÉSENTATION DE DONNÉES

# BILAN DU PROJET

Ce projet a été réalisé dans le cadre du cours modélisation connaissances & humanités numériques. Celui-ci convoque des connaissances variées afin de mettre en œuvre les techniques d’extraction du contenu de sites web, via un script ou un programme dans le but de le transformer pour permettre son utilisation dans un autre contexte. On l’a utilisé pour la construction d’un jeu de données à l’utiliser au futur pour la visualisation de ces données au cours du prochaine module académique. Il a été l’occasion pour nous, étudiants, d’avoir une première expérience pratique et ainsi se familiariser avec le domaine du web scraping, comme nous l’avons vu, sont infinies et qui sert à améliorer la compétitivité des entreprises. On a utilisé ses techniques tout en respectant la propriété intellectuelle des bases de données car il a été un projet strictement académique.