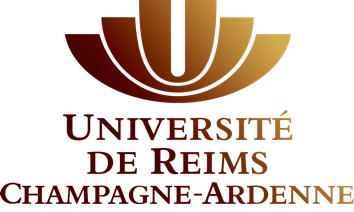
LIEPO Brice-Kevin - ROA SERRANO Walter

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](#_Toc32874772)

[II. OBJECTIFS DU PROJET 3](#_Toc32874773)

[III. PRÉSENTATION DES DONNÉES 4](#_Toc32874774)

[IV. OUTILS UTILISÉS 6](#_Toc32874775)

[1) R STUDIO 6](#_Toc32874776)

[2) OPENWEATHER 6](#_Toc32874777)

[3) Les différents sites utilisés pour le web scrapping 7](#_Toc32874778)

[4) GitHub 7](#_Toc32874779)

[V. BILAN 8](#_Toc32874780)

# 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. C’est aussi un processus de récupération de plusieurs types de données à partir de différentes sources, dont certaines peuvent être mal structurées. L'extraction des données permet de consolider, traiter et affiner les données, puis de les stocker dans un emplacement bien défini.

L'extraction des données est la première étape du processus ETL (Extract, Transform, Load) qui désigne une séquence d'opérations portant sur les données, permettant de consolider et regrouper différents types de données provenant de différentes sources sous un format commun. Ce processus comprend trois étapes :

* L’Extraction ou Extract : Identifie les données pertinentes, puis les prépare pour traitement ou transformation ;
* La Transformation ou Transform : Les données sont triées, structurées et nettoyées, les entrées en double sont supprimées de manière à conserver une seule instance, les valeurs manquantes sont supprimées ou enrichies et des vérifications sont effectuées de manière à obtenir des données cohérentes, exploitables et fiables.
* Le Chargement ou Load : Une fois transformées, les données sont de haute qualité, et elles sont conservée dans un emplacement analyse.

Ce processus est de plus en plus utilisé dans la plupart des secteurs d'activité avec des objectifs variables. Une fois que les données ont été collectées et stockées, l’étape suivante consiste à leur donner du sens.

Le web représente une des plus importantes sources d’informations où il est possible d’acquérir des données dans les contenues des sites par l’utilisation du Web Scraping, permettant d’exploiter la structuration par balises du web pour en extraire du contenu, et par l’utilisation des APIs qui augmente 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 de Reims, il nous est proposé de mettre en pratique nos connaissances et nos compétences professionnelles au travers d’un projet ayant pour objectif principal la construction d’un jeu de données utilisable au moyen des différentes méthodes et techniques énoncées ci-dessus.

# OBJECTIFS DU PROJET

Ce projet consiste à mettre en place des techniques d’extraction des données dont le web scraping et l’utilisation des APIs pour construire un jeu de données contenant un minimum de 100 individus décrits par 10 caractéristiques au minimum dans le cadre de ce projet on a utilisé 198 capitaux des pays du monde qui avaient les informations les plus significatives dont il y aura 17 variables à observer lesquelles seront des différents indicateurs comme par exemple l’espérance de vie, la mortalité, la natalité, les tourisme entre autres qui ont été extraites avec la technique du web scraping et des différents conditions météorologiques qu’on a récupéré en utilisant l’API par le service en ligne Openweathermap.

Ce jeu de données sera utilisé dans le cadre du prochain module INFO0808 Visualisation des données où nous allons l’utiliser comme ressource pour visualiser et analyser les résultats.

LA STRATÉGIE

Le projet est composé par un seul fichier d’extension .R dont qui contiendra tout le code nécessaire pour l’extraction et la construction du jeu de données.

Il a consisté à écrire des différentes fonctionnalités. Tout d’abord, on a écrit la fonction qui permettra collecter les données sur le site, faire le web scraping et récupérer des pays et les informations fournies grâce à les paramètres dans l’url du site utilisé et extraire les différents variables qu’on a choisi avec son pays correspondant en construisant une dataframe, ensuite on a écrit le script qui permettra de récupérer le pays et son contient correspondant, cela permettra d’une part la possibilité de faire la jointure en utilisant le pays comme la valeur des colonnes dans chaque table et d’autre part pouvoir construire une nouvelle dataframe avec toutes les variables souhaitées mais aussi avec le continent qui nous allons utiliser comme factor du jeu de données puis on a renommé le colonnes avec la valeur de chaque variable.

En second, il est réalisé un nouvel web scraping dans un autre site web où on a récupéré encore le pays mais cette fois en construisant une dataframe avec sa capitale correspondante qui permettra ensuite faire une nouvelle jointure à partir du pays comme colonne équivalente. Ici on a trouvé le premier souci qui a besoin de nettoyage des données car les pays n’avaient pas le même nom donc pour régler ce problème on a écrit deux fonctions, une pour supprimer les caractères accentues et des espaces superflus et une autre pour remplacer le nom des pays et capitaux pour faire concorder les données parmi les deux data frames auparavant créés. Avec cette fonction on a fait le remplacement du nom de certains pays et des capitaux pour assurer la concordance lors de la jointure.

Ensuite, une fois qu’il y a concordance entre les noms de pays te des capitaux on a executé la jointure entre les deux data frame réalises lors du web scraping réalisé sur les deux différents sites web.

Puis, en finissant la première partie de nettoyage on a supprimé les lignes de la data frame contenant les pays qui ne possédant pas assez d’informations significatives.

La deuxième étape du développement a été l’utilisation de l’API Openweathermap donc, on a écrit une fonction qui permettra tout d’abord construire l’url contenant les paramètres et la clé nécessaire pour l’authentification qu’on a obtenu de manière gratuite.

Pour solliciter les informations nécessaires l’url a été construit en ajoutant la capitale de chaque pays comme le paramètre utilisé dans notre requête http puis la clé (API key). Dans la fonction on s’est servi du paquet Jsonlite pour obtenir les informations sous le format JSON.

Ici on a retrouvé encore un souci à nettoyer, cette fois, pour pouvoir réussir à faire correctement les requêtes http on a besoin de vérifier à chaque fois si le paramètre demandé ne contient pas la valeur 0 donc, en ajoutant la valeur NA nous permettra de réaliser avec succès toutes les requêtes sans fausser les résultats.

Une petite astuce a été d’utiliser une fonction que permettra faire les requêtes dans certains intervalles du temps pour ne pas charger le serveur au moment de l’exécution du script.

Pour le retour de la fonction on a récupéré toutes les informations souhaitées qui ont été fourni par l’API.

Ensuite, on a créé une data frame qui contiendra les informations retournées par l’appelle de la fonction qui récupère les données de l’API.

Puis, on a appliqué cette fonction sur l’ensemble des capitales présentes dans le résultat de la jointure réalisé dans les data frames construits avec du web scraping.

Le dernier nettoyage qu’on a réalisé a été la suppression de la première ligne utilisée pour initialiser la data frame contenant les informations retournées par l’API.

Ensuite on a renommé les colonnes avec les noms correspondants des variables.

Finalement on a fait la dernière jointure entre les deux data frames restants, le renommage des différentes lignes de la data frame finale et la suppression de la colonne pays pour éliminer le doublon.

RÉALISATION DE TIBBLE (CONVERTIR LA DATAFRAME FINAL EN TIBBLE COMME IL A DIT ça on verra après si ce possible)

# PRÉSENTATION DES DONNÉES

Après avoir collecté et nettoyé différent type d’information, nous avons obtenu un jeu de données contenant 194 individus qui sont les différents pays et 17 variables. Parmi les différentes variables contenues dans notre jeu de données, 7 proviennent des informations récupérées avec l’API par utilisation de la capitale du pays et 10 proviennent des informations récupérées par web scrapping. Les différentes informations contenues dans les variables récupérées par web scrapping ont été récoltées entre 2016 et 2017.

Les 17 différentes variables sont les suivantes :

|  |  |
| --- | --- |
| **NOM DE LA VARIABLE** | **DESCRIPTION** |
| Capitals | Variable de type chaine de caractère, qui représente la capitale du pays. C’est le contenu de cette variable qui sera utilisé comme paramètre pour l’exécution de la fonction mise à disposition par l’API. |
| Continent | Variable de type facteur à cinq niveaux, représentant le contient auquel le pays appartient. |
| Esperance\_vie | Variable de type numérique, représentant l’espérance de vie dans le pays. |
| Mortalite\_inf | Variable de type numérique, représentant le taux de mortalité infantile dans le pays. |
| Indice\_perf\_env | Variable de type numérique, représentant l’indice de performance environnemental du pays. |
| Mortalite | Variable de type numérique, représentant le taux de mortalité global du pays. |
| tourisme | Variable de type numérique, représentant le nombre de touriste enregistré au cours de cette période. |
| pib-par-habitant | Variable de type numérique, représentant le PIB par habitant. |
| Natalite | Variable de type numérique, représentant le taux de natalité dans le pays. |
| Superficie | Variable de type numérique, représentant la superficie du pays. |
| Longitude | Variable de type numérique, représentant la longitude de la capitale du pays. |
| Temp\_actu | Variable de type numérique, représentant la température actuelle de la capitale du pays. |
| Temp\_min | Variable de type numérique, représentant la température minimale du jour de la capitale du pays. |
| Temp\_max | Variable de type numérique, représentant la température maximale du jour de la capitale du pays. |
| Humidity | Variable de type numérique, représentant l’humidité du jour de la capitale du pays. |
| Type\_temps | Variable de type facteur à quinze niveaux, représentant le type de temps. |

# OUTILS UTILISÉS

## R STUDIO

R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données. Il permet de 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 telles que : Tidyverse (visualisation et manipulation des tableaux de données, importation et exportation de données, manipulation de variables, extraction de données du web …), Xml2, Rvest ou encore Jsonlite.



## OPENWEATHER

OPENWEATHER est un site qui propose un service permettant de fournir des informations sur la météo et les prévisions actuelles dans une ville. Ce site dispose d’une API permettant d’obtenir des données météorologiques en temps réel telles que : la température actuelle, la température maximale, la température minimale, l’humidité, la direction du vent… Dans le but d’utiliser l’API nous avons demandé l’obtention d’une clé gratuite nous permettant d’effectuer 60 requêtes par minutes et à l’option « Current weather API » parmi toutes les options disponible sur le site. L’utilisation de l’API se fait comme suit :

* L’url : <http://api.openweathermap.org/data/2.5/weather>?
* Les paramètres utilisés dans notre projet à ajouter à l’url:
* q : désigne dans notre cas le nom de la capitale (q=Paris)
* units : désigne l’unité de la valeur (units=metric)
* appid : désigne la clé api utilisée

(appid=9ada210033e2363be58a9fac5b682c4f)

* lang : Désigne la langue (lang=fr)

L’APIs propose bon nombre de paramètres et de méthodes permettant d’accéder à l’information. Nous avons choisi la méthode qui n’inclue que les paramètres cité ci-dessus.



## Les différents sites utilisés pour le web scrapping

Les sites utilisés pour le web scrapping contiennent plusieurs informations sur bon nombre de pays. Ce sont :

* <https://www.populationdata.net/palmares>
* <https://jeretiens.net/tous-les-pays-du-monde-et-leur-capitale/>

## GitHub

GitHub est un service web d'hébergement et de gestion de développement de logiciels, utilisant le logiciel de gestion de versions Git. Il un contrôle d'accès et des fonctionnalités destinées à la collaboration comme le suivi des bugs, les demandes de fonctionnalités, la gestion de tâches et une sorte de documentation pour chaque projet.

IMAGE GIT



# BILAN

Ce projet a été réalisé dans le cadre du cours de modélisation de connaissances & humanités numériques. Celui-ci exige des connaissances variées afin de mettre en œuvre les techniques d’extraction de contenu de sites web, au moyen d’un script ou un programme dans le but de le transformer pour permettre son utilisation dans un autre contexte. Ces différentes connaissances ont été utilisées pour la construction d’un jeu de données, qui sera utilisé pour faire de la visualisation de données.

Nous avons, au cours de ce projet, rencontré bon nombre de difficultés liées en grande partie à l’utilisation de l’API. En effet, certaines erreurs telles que des erreurs de requêtes, de permissions ou de réponses nous ont posé problème et ont nécessité un travail de recherche sur les différents types d’erreurs http, appelés encore les statuts http et codes d'erreur concernant le JSON. Nous avons donc découvert que l’erreur 400 fait référence à une mauvaise requête, l’erreur 403 à une restriction au niveau des permissions et l’erreur 404 à une réponse vide, c’est-à-dire que l’élément recherché n’existe pas.

En définitive, il a été l’occasion pour nous, étudiants, d’avoir une première expérience pratique et ainsi se familiariser avec le domaine du web scraping, l’utilisation des APIs ou encore le choix des ressources adéquates pour la réalisation d’un projet.