Guide utilisateur du Projer R

Stage Web Scraping – Données immobilières de Mexico

Le projet R réalisé lors de ce stage est regroupé au sein d'un dossier, comprenant plusieurs scripts R, qui permettent le scraping des sites, le nettoyage des données, le traitement des données et la génération des cartes

Pré-requis

Un IDE en langage R comme RStudio est nécessaire pour faire fonctionner les différents scripts et varier les paramètres à afficher. Il est également nécessaire d'installer les librairies R suivantes :

- dplyr : https://cran.r-project.org/web/packages/dplyr/index.html
- tidyr : https://cran.r-project.org/web/packages/tidyr/index.html
- rvest : https://cran.r-project.org/web/packages/rvest/index.html
- xml2 : https://cran.r-project.org/web/packages/xml2/index.html
- httr : https://cran.r-project.org/web/packages/httr/index.html
- sf : https://cran.r-project.org/web/packages/sf/index.html
- ggplot2 : https://cran.r-project.org/web/packages/ggplot2/index.html
- stringr : https://cran.r-project.org/web/packages/stringr/index.html

Contenu du dossier

Le dossier du projet se présente sous la forme suivante :

- 5 scripts R permettant le scraping, le nettoyage et les traitements des données
- Les ficher .csv sous la forme, qui contiennent les données brutes scrapées
- Le dossier *Donnees_nettoyees*, qui contient les données nettoyées et mises en forme au format .csv
- Le dossier *Donnees_traitees*, qui contient les données traitées et validées géographiquement, au format .csv
- Le dossier *SHP_Traites*, qui contient les mêmes données au format .shp
- Les dossiers *Municipios*, *AGEB* et *Colonias*, qui contiennent les géométries des différents découpages géographiques de la ZMVM (Municipes, AGEB et Colonies)

Processus de collecte et de traitement des données

Web Scraping

La partie Web Scraping est réalisée à partir des scripts R *Scrap_metroscubicos.R*, *Scrap_inmuebles24.R* et *Scrap_tecnocasa.R*, qui permettent de scrapper respectivement les sites Mercado Libre (ou Metros Cubicos, les deux sites ayant la même architecture et les mêmes données), Inmuebles24 et Tecnocasa.

Si l'utilisateur souhaite simplifier le scraping en réduisant le nombre d'annonces à scraper, il est possible de modifier la liste *l_type* qui contient les types de logements (maison, appartements, …) à scraper, et le dataframe *df_**nom du site a scraper*, qui contient les municipes ciblés.

Une fois scrapées, les données sont écrites sur un fichier .csv, nommé *nom du site**.csv*.

Nettoyage des données

Le nettoyage des données est effectué avec le script *Nettoyage.R*. La fonction *nettoyage (site, file)* permet d'effectuer le processus. Elle contient deux paramètres à remplir :

- Le paramètre *site* indique le nom du site web depuis lequel les données ont été scrapées ("mercadolibre", "metroscubicos", "inmuebles24", "tecnocasa" ou "metroscubicos2015")
- Le paramètre *file* indique l'emplacement du fichier contenant les données brutes (chemin d'accès + nom du fichier)

Une fois le nettoyage effectué, les données sont enregistrées au sein d'un fichier .csv, situé dans le dossier *Donnees_netoyees*, et nommé *nom du site**_clean.csv*.

Traitement et affichage des données

Le traitement (suppression des annonces indésirables et validation géographique) s'effectue depuis le script *Traitement.R*.

La fonction *traitement (df, mode, tolerance_dist, source, keep_false, echelle)* regroupe l'intégralité des fonctions de traitement. Ses paramètres sont les suivants :

- df : un dataframe contenant les données nettoyées. Il peut être créé en amont par l'utilisateur, avec la fonction *read.csv (**chemin d'accès fichier .csv des données nettoyées*, *encoding = "UTF-8", stringsAsFactors = FALSE)*
- mode : mode de paiement pour le bien ("V" pour les ventes, "R" pour les locations)
- tolerance_dist : distance, en m, de tolérance des annonces autour des municipes ou des colonies pour la validation géographique
- source : nom du site web d'origine des données (voir 0)
- keep_false : vaut TRUE si l'utilisateur veut conserver les données non validées géographiquement, FALSE sinon.
- echelle : échelle de la validation géographique ("MUN" pour les municipes, "COL" pour les colonies)

Une fois le traitement effectué, cette fonction retourne un dataframe contenant toutes les données valides, qui peut être exploité pour l'affichage des données. Elle écrit également un fichier *source*_*clean*_*mode*.*csv* situé dans le dossier *Donnees_traitees*, ainsi qu'un fichier *source*_*mode**clean.shp* situé dans le dossier *SHP_Traites*.