Examen

Visseho Adjiwanou, PhD.

09 June 2024

Exercice 1: individuel

Les résultats obtenus à partir d'une recherche avec les mots-clés suivants: 'data science africa' sur crossref:

https://api.crossref.org/works?query=%27data%20science%20africa%27

Remarque: a. Si vous cochez la case Pretty-print, il vous donne l'arborescence avec une vue plus claire.

b. Crossref (anciennement stylisé CrossRef) est une organisation à but non lucratif fournissant une infrastructure numérique ouverte pour la communauté mondiale de la recherche universitaire. Elle enregistre et connecte de manière unique et persistante les connaissances via des métadonnées ouvertes et des identifiants pour tous les objets de recherche, tels que les subventions et les articles. En tant que plus grande agence d'enregistrement de Digital Object Identifier (DOI) de la Fondation Internationale DOI, Crossref compte 19 000 membres de 150 pays, représentant des éditeurs, des bibliothèques, des institutions de recherche et des bailleurs de fonds. Lancée au début des années 2000, elle a débuté comme un effort coopératif entre éditeurs pour permettre le lien de citation persistant entre les plateformes dans les revues académiques en ligne. En juillet 2023, Crossref identifie et connecte 150 millions de dossiers de métadonnées sur des objets de recherche, rendant ces données librement disponibles pour réutilisation sans restriction. Elle facilite en moyenne 1,1 milliard de résolutions de DOI (clics sur un lien DOI) chaque mois et gère 1 milliard de requêtes de métadonnées par mois.

1. Déterminez la taille de l'ensemble de données renvoyé par la recherche.
2. Convertissez le fichier JSON en une base de données avec les informations importantes en utilisant R de deux manières :

2.1. Codage manuel : Écrivez votre propre code en R pour analyser et transformer les données JSON.

2.2. Utilisation d'un package approprié : Utilisez un package R pour vous aider dans la conversion.

3. Collecter et analyser les résumés des articles: À partir des DOI ou des URL des articles, accédez aux sites respectifs et collectez les résumés des articles. Ajoutez cette information à votre base de données initiale. Vous pouvez collecter uniquement les données de 10 articles.
4. Réalisez une petite analyse de texte à partir des résumés collectés.

Exercice 2: Individuel

Le site web des Nations Unies sur les SDG (sustainable development goals), en Français objectifs du développement durable (ODD) vous donne les informations sur les 17 objectifs.

le lien du site est le suivant: https://sdgs.un.org/fr/goals

1. Scraper ce site et collecter les informations dans une base de données sur :

- le titre de l'objectif
- le nombre de cibles
- le nombre d'évènements
- le nombre de publications
- le nombre d'initiatives

Par ailleurs, vous avez toutes les publications concernant ces objectifs. Voici le lien du site: https://sdgs.un.org/publications

2. Votre objectif est de scraper ce site et de présenter dans une seconde base de données claire les titres de ces publications. Remarquez qu'il y a jusqu'à 113 pages. Cette base de données doit comporter deux colonnes, la colonne 1 pour l'identification, allant de 1 au nombre de publication, la deuxième, le titre.
3. Maintenant, à partir des titres, essayer de prédire l'objectif en question. Mettez cette information dans la base de données (avec la nouvelle variable : objectifs)

Exercice 3 : Individuel

ce site https://books.toscrape.com/index.html est une librairie fictive qui propose des milliers de livres à scraper.

1. Scraper-le et présenter le résultat dans une base de données qui indique en plus le type de livre (Travel, Mystery).
2. Pour chaque livre, collecter les informations sur la description du livre et ajouter-les dans la base de données.
3. Présenter les 20 mots les plus cités dans chaque collection. Est-ce que ces mots reflètent

Exercice 4 : En équipe

Si vous souhaitez vous lancer dans les forums, je vous conseille de retrousser vos manches et de visiter Reddit. Le site suit un format d'URL spécifique afin que les utilisateurs puissent publier des images, des vidéos, des liens et du contenu similaire. Vous pouvez extraire n'importe quel commentaire ou image ayant reçu le plus de votes positifs, identifier les mots-clés les plus récurrents dans un subreddit ou analyser le sentiment public derrière une information que vous trouvez intéressante.

Le scraping Web d'un forum peut vous conduire à une idée commerciale réussie et, en même temps, vous pratiquerez certaines bases comme l'extraction de liens, d'images, de noms d'utilisateur et de commentaires.

Cependant, le scraping n'est pas si simple après la refonte de Reddit – le site Web est quelque peu compliqué. C'est pourquoi je suggère d'utiliser l'ancienne mise en page sur old.reddit.com.

1. Que fait le site?
2. Est-ce que ce site est utilisé au Burkina, ou en Afrique subsaharienne?
3. Donner un exemple de discussion concernant l'Afrique subsaharienne.
4. Scraper ce site old.reddit.com

Exercice 5: Par équipe

Ce site web (https://www.digitalgendergaps.org/) a été généré à partir des informations collectées sur Facebook. Le API du site permet de télécharger un certain nombre d'information sur le site. L'objectif de cet exercice est de suivre la méthodologie décrite ici pour collecter d'autres données sur Facebook et de les présenter sous forme de page web à partir de Shinyapp (https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/index.html).

1. Décrivez comment les données ont été générées
2. Décrivez comment les indicateurs ont été calculés. Est-ce que ces indicateurs peuvent-ils être présenté selon le milieu de résidence?
3. A votre tour de choisir un autre indicateur que les données Facebook peuvent vous donner (sur la migration par exemple, les déplacements de population au sein de la CEDEAO. . .) et présenter-le sous forme de site web dynamique.

Voici quelques références qui peuvent vus aider:

https://www.demographic-research.org/articles/volume/43/27/

1.