Utilisation d'OpenRefine pour le nettoyage des données

Description

OpenRefine est un logiciel libre de nettoyage et de mise en forme de données. Il est similaire dans son apparence à un tableur mais fonctionne en fait comme une base de données. (https://fr.wikipedia.org/wiki/OpenRefine)

Mise en route

Page de téléchargement: https://openrefine.org/download

- Dézipper le dossier et lancer l'exécutable.
- Quand on ouvre OpenRefine, un terminal et une page web de votre navigateur s'ouvrent.
- Pas de crainte, le fichier reste en local.

Ressources utiles

Ecole d'hivers SHS 2023 (1h24) (L'Alliance, Canada)

Description: Tutoriel sous forme de cours donné à des étudiant-e-s. Très bien expliqué depuis la base, en français.

A retenir:

- mise en route
  - [14:40] enlever les espaces vides avant et après le texte
- la possibilité de faire des facettes texte
  - [27:00] reconnaître des valeurs similaires - corriger les erreurs de saisie
  - [41:33] mettre en forme des cellules en enlevant par exemple des [ou des "
  - edit cells >transform > value.replace("[","")
- facette texte avancée pour permettre d'avoir des infos sur chaque item d'une même colonne

    - cas: (pomme;banane;poire) dans les cellules -> je veux savoir le nombre de fois où chaque mot apparaît dans le tableau.
        - [44:26] Facet > Custom text facet >value.split(";")
- revenir en arrière / reproduire ses actions
    - [01:07]pour voir l'historique des actions, revenir en arrière si nécessaire, pouvoir exporter ses actions pour les reproduire sur un autre fichier
    - [01:12] exporter ses données, sauvegarder son travail

Vidéo tutoriel Tuto@Mate , 2019(2h05)

Wiki sur Github , mars 2015)

- regrouper les données
    - [28:58] option "grouper" pour transformer toutes les manières dont une même donnée apparaît.
- restructurer les données, transposer les colonnes
    - [32:23]
- utilisation des expressions régulières
    - [41:13]
    - exemple (GED): enlever les groupes "Adm3 - " dans une PMR Sharegate
        - éditer les cellules > transformer > value.replace(/Adm3 -.*/,"0")
        - résultat: remplace tous les groupes qui commencent par "Adm3 -" par 0
- réconciliation des données: aller chercher des données dans le domaine public pour enrichir ses données
    - [43:53]

Manipulations à connaître

Effacer toutes les données d'une colonne:

`""`

Cette expression remplace chaque cellule de la colonne par une chaîne vide, effaçant ainsi toutes les données

Changer le format d'une colonne contenant des dates

Si votre date est au format `D.M.YYYY` (comme `1.1.2023` pour le 1er janvier 2023), vous pouvez utiliser l'expression suivante pour la transformer en `M/d/yyyy` :

`value.toDate("d.M.yyyy").toString("M/d/yyyy")`

Cette expression fait deux choses :

- `value.toDate("d.M.yyyy")` : Convertit la chaîne de caractères en un objet date en utilisant le format actuel.

- `.toString("M/d/yyyy")` : Convertit l'objet date en une chaîne de caractères dans le nouveau format souhaité.

Astuces

Utiliser Copilot ou Mistral AI pour trouver les expressions dont on a besoin sur GREL

Créé 2025-02-07 07:42:40 UTC par Laurence Gauvin
Mis à jour 2025-02-19 13:59:09 UTC par Laurence Gauvin