TP22 – Nettoyage de données

Correction des fautes d'orthographe

Ressource : SlateGunDeaths.csv

Ouvrez le fichier dans OpenRefine, choisir le séparateur ‘,’ et metter les nombres en couleur verte et finalement créer le projet (choisir un nom, TP3 par exemple)

1. Classement par ville
    a. City -> Facet -> Text Facet. Cela donne 2071 villes différentes
2. Coupez les espaces (voir City ‘Albuquerque’ avec et sans espace)
    a. City -> Edit cells -> Transform -> value.trim (). cela donne 2063 villes
3. Majuscules et minuscules (voir State ‘OH’, ‘oh’)
    a. State -> Facet -> Text Facet
    b. State -> Edit cells -> Common Transform -> to uppercase
4. Clustering (voir: City ‘Oklahoma city’)
    a. City -> Edit cells -> Cluster and Edit. Par défaut Method = ‘Key collision’ et Keying Function = ‘Fingerprint’.
    b. Choisir les valeurs que vous souhaitez fusionner
    c. Combiner d’autres methodes (nearest neigbor) et d’autres Keying Functions ('Key collision', ‘ngram Fingerprint’, ‘nearest neigbors’, ‘Levenshtein’) et fusionner.

5. Analyser les données
    a. Creer une facette pour City. Pour classer les victimes par ville, dans City facet, clicker sur count. Parce qu'il compte le nombre de lignes par ville, vous obtenez le nombre de victimes par ville
    b. Cliquer sur une valeur (ou une ville) affichera les données de cette valeur. Exemple city ‘Columbus’. On peut remarquer que la ville Columubus est dans differents etats (OH, GA, IN). Ce probleme peut etre regler par la concaténation.

6. Concaténation
    a. Lier State à City pour filtrer, cliquez sur reset dans facet de City. Puis, City -> Edit Cells -> Transform ->Expression : value + "," + cells ["colonne"].value