TP25 – Nettoyage de données

Text Arabe et Clustering

Ressource : arabText.xlsx

1. **Scraping à partir du web**
   - **a.** Voir le tableau de donnée sur la page: http://pcbs.gov.ps/Portals/_Rainbow/Documents/jerica.htm?lipi=urn%3Ali%3Apage%3Ad_flagship3_pulse_read%3BOLebhT5oTKGzb3rqUyS0BQ%3D%3D
   - **b.** Scrapez le tableau dans un fichier Excel ou Google Spreadsheet

2. **Clustering à l'aide de OpenRefine**
   a. Ouvrer le fichier arabText.xlsx dans OpenRefine et créer un projet.
   b. Aller dans l'outil du clustering avec : colonne "اسم التجمع" – edit cells- cluster and edit.
   c. Choisisser une methode de clustering "nearest neighbor" et modifier les distances et le "Radius" et le "block chars" pour ajuster les résultats puis fusionner.