Data Cleaning using Openrefine

Video Link : https://www.youtube.com/watch?v=cX_2MkShlJk

Openrefine : OpenRefine is an open-source desktop application for data cleanup and transformation to other formats, an activity commonly known as data wrangling.

> How to cluster a column:
>
> > Drop down menu(Available in the Column) → Facet → Text Facet → Cluster

We can also edit the name of the new cluster formed in this way.

Some details about the Openrefine algorithm:

- **Key collision** is the default clustering algorithm. It is also the most stringent algorithm. It removes the special characters from the text then converts the whole string it into lowercase & then clusters it.
- **Nearest neighbors** [Levenshtein distance] is based on Levenshtein distance(*Number of edits that needs to be done between two strings*)
- **Nearest neighbors** [ppm] if any of the substring matches between 2 strings, it clusters those 2 strings into one.