# Data Mining Techniques

## Data Crawling

Building a crawler to extract the Data mining, Machine Learning, Databases and AI conferences and their location from WIKICFP page using java.

Approach to solving the issue:

Provided Java code is used to build the web crawler&JSOUP library is used to crawl the HTML content of the page.It is a java library which helps in extracting, manipulating and parsing DOM data very conveniently using API. Java program is to extract acronym, name, location where conference is held and write this information to the tab separated txt file.HTML Table data "<td>" attributes like "rowspan", colspan, "align" etc. is used to extract the information and stored in an integer array,later content of the array is iterated and written in to the text file. Comments are included in the code which clearly explains the process of parsing and extraction of HTML data.

Challenges faced:

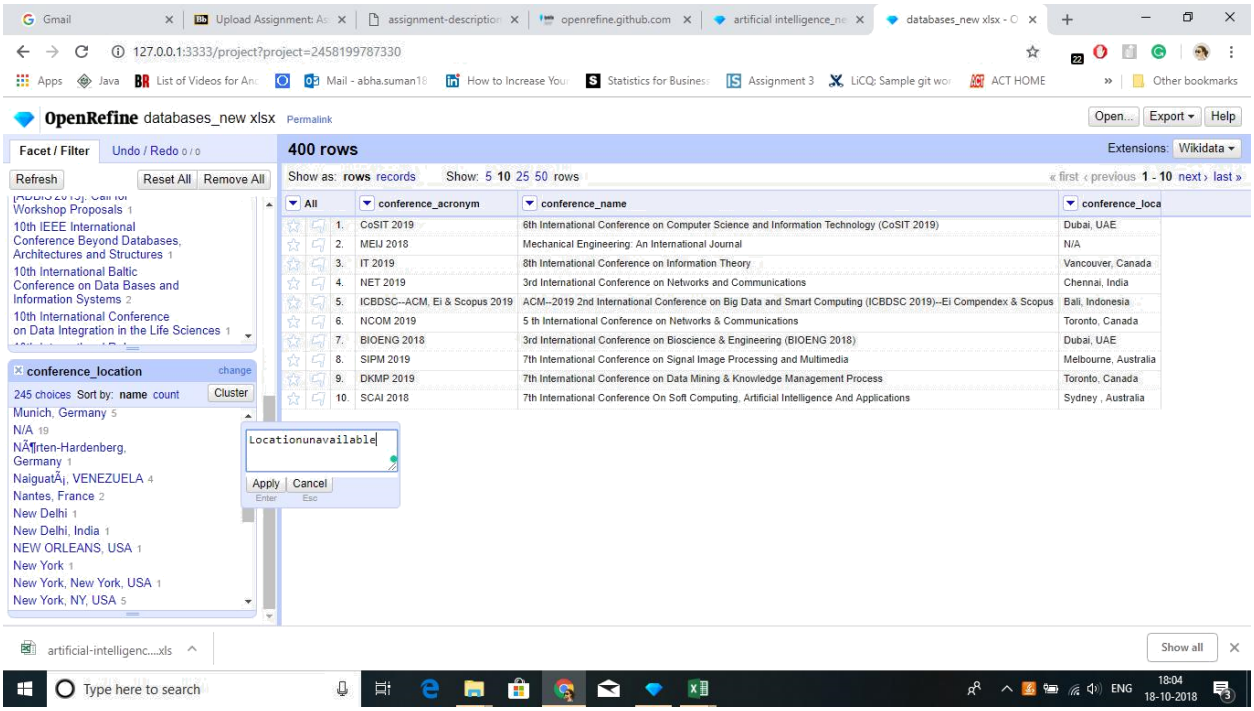
It was challenging to extract the location value .because the table row data with location value had only one attribute align="left." Differentiating from other row data of the HTML table was difficult. Using Jsoup API selected the first row among three-row data of the table having "align=left" attribute.

Data Cleaning:

Tab separated txt file from Java program is converted in to ".xlsx" and uploaded in to Open Refine. Project is created in Open Refine to clean and cluster the four categories data.

Screenshot -1

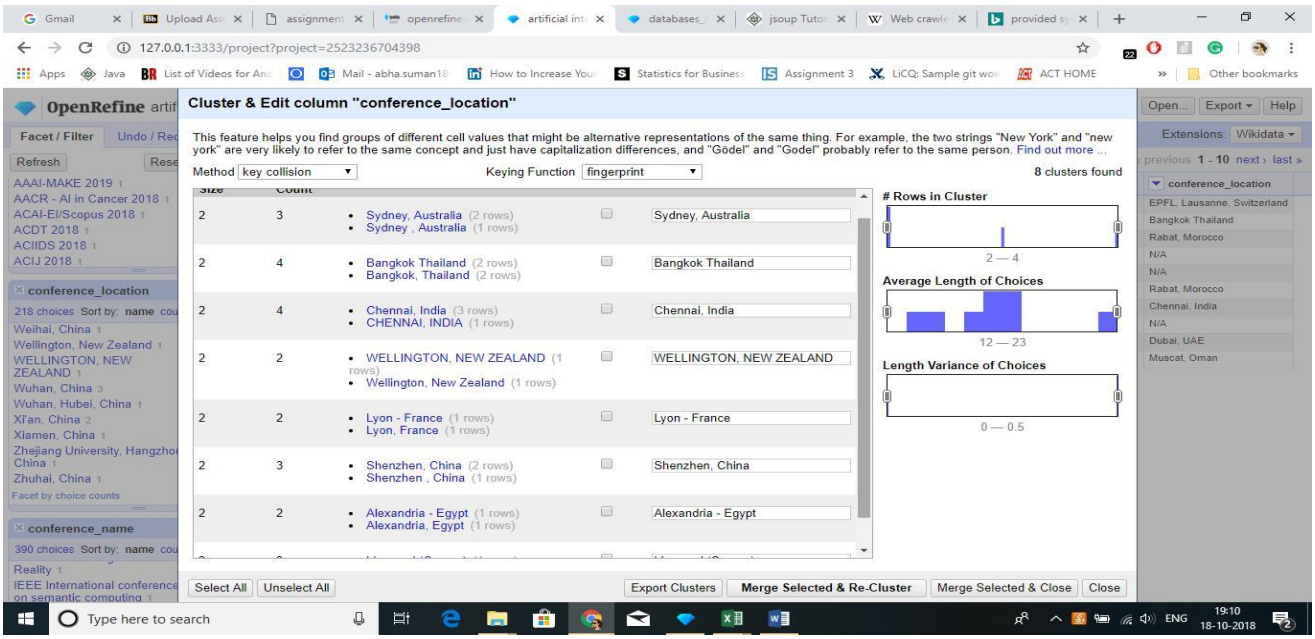
Assignment is to find the location where the conference is taking place. The data which is collected in txt file has improper values like N/A .using OpenRefine N/A is replaced with meaning value such as "LocationUnavailable" .below is the screenshot of same:




Here 19 records has location values as N/A, using OpenRefine ,value is replaced with “Location Unavailable”

Screenshot-2

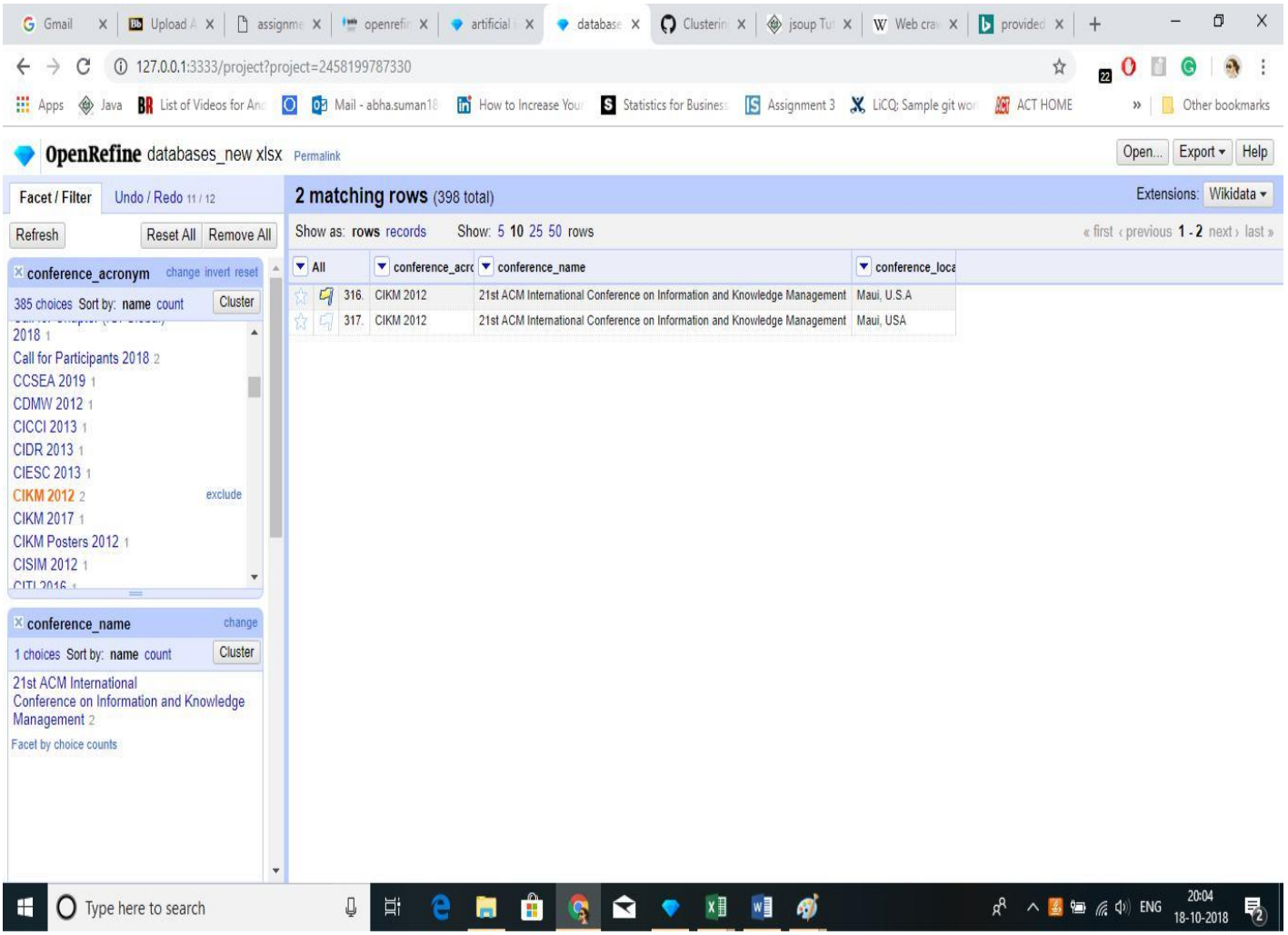
OpenRefine helps in clustering or grouping the different cells values which is alternative representation of same thing.




Here for example Sydney, Australia can be clustered and merged using openRefine to give uniform name to the records.

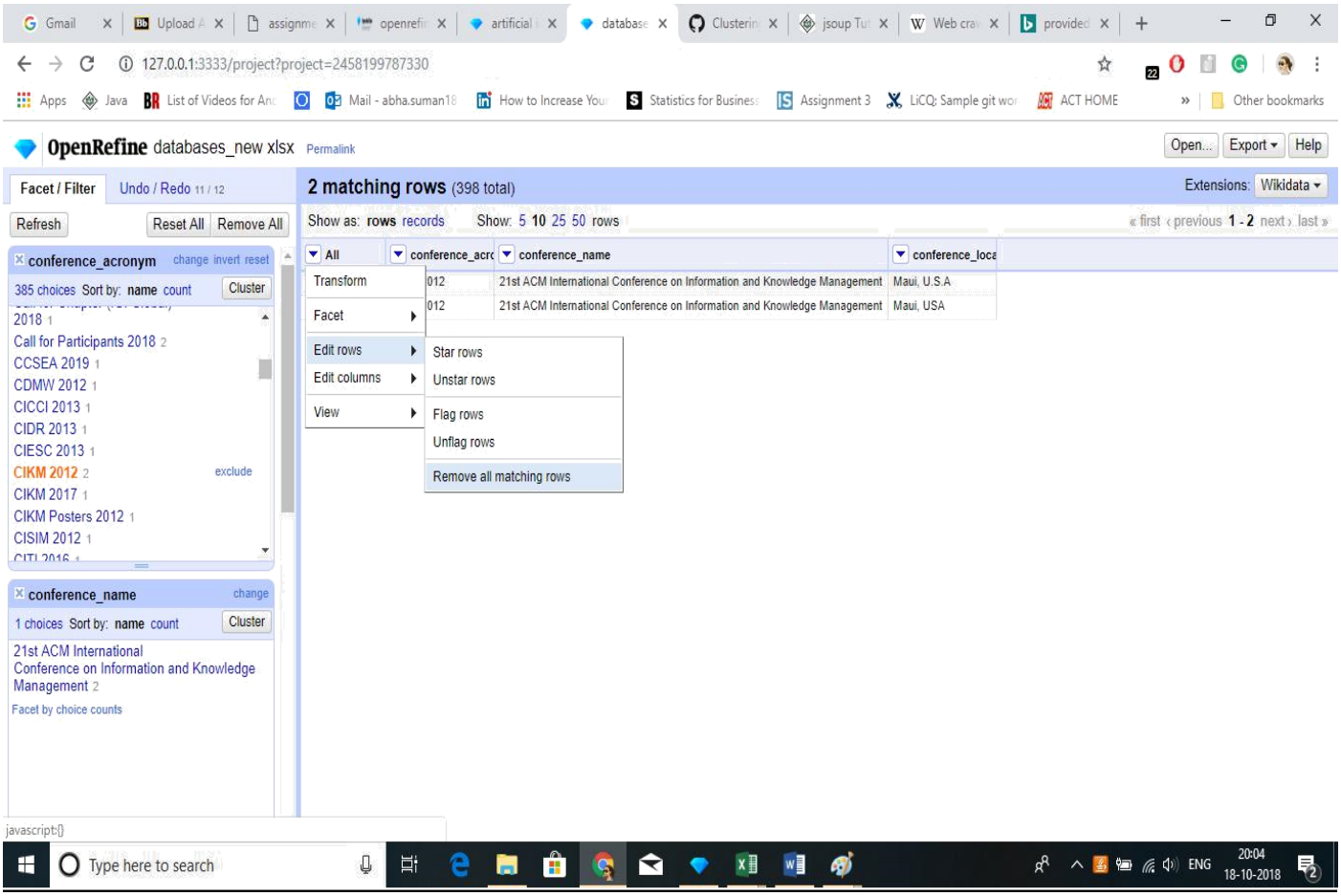
Screenshot-3

Facet is applied to acronym column to find the duplicate records where the same conference is listed multiple times. The duplicate rows are flagged and removed.



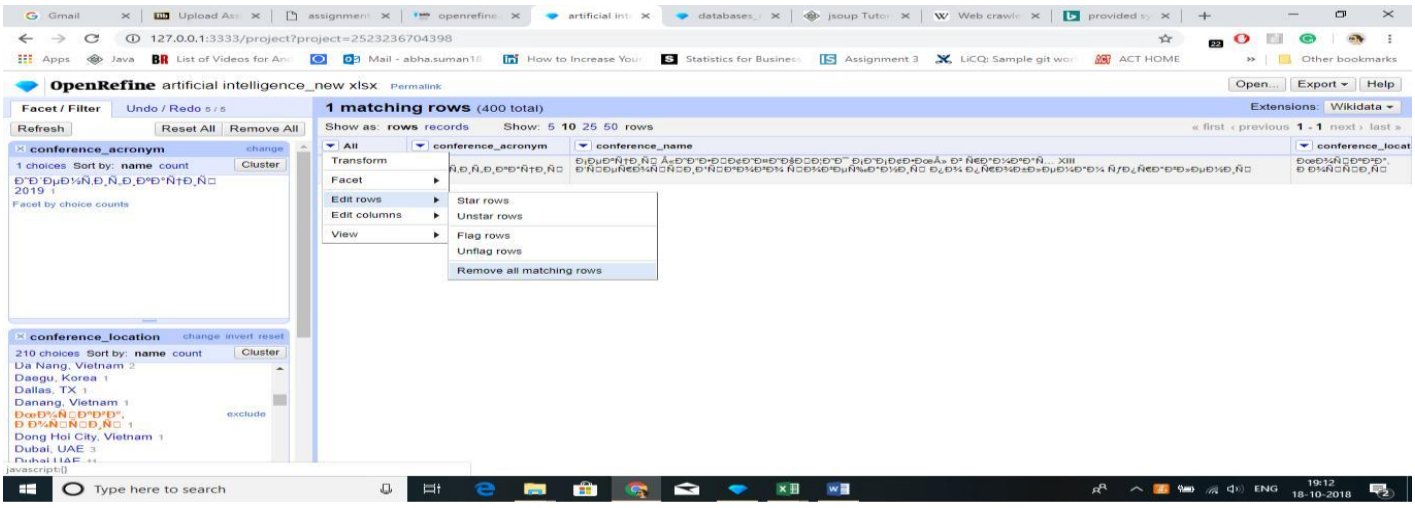

Screenshot-4

Lists the duplicate rows with acronym =” CIKM 2012” .duplicate record is flagged and removed as below:




## Screenshot-5

Junk or garbage values from the records can be removed by flagging the rows .the below screenshot explains the same:



References:

https://www.tutorialspoint.com/jsoup/index.htm

OpenRefine