

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:

The screenshot shows the OpenRefine web interface. At the top, there's a browser tab for 'artificial-intelligence-ne' and a URL '127.0.0.1:3333/project?project=2458199787330'. The main area displays a table with 400 rows. The columns are 'conference_acronym', 'conference_name', and 'conference_location'. A dropdown menu for 'conference_location' is open, showing a list of locations and 'Location unavailable' selected. The table shows various conference records, including CoSIT 2019, MEUJ 2018, IT 2019, NET 2019, ICBDS-C-ACM, EI & Scopus 2019, NCOM 2019, BIOENG 2018, SIPM 2019, DKMP 2019, and SCAI 2018.

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.

The screenshot shows the 'Cluster & Edit column' dialog in OpenRefine. The dialog is titled 'Cluster & Edit column "conference_location"'. It shows a list of clusters with their size and count. The clusters are: Sydney, Australia (2 rows), Bangkok Thailand (2 rows), Chennai, India (3 rows), WELLINGTON, NEW ZEALAND (1 rows), Lyon - France (1 rows), Shenzhen, China (2 rows), Alexandria - Egypt (1 rows), and Alexandria, Egypt (1 rows). A histogram on the right shows the distribution of cluster sizes, with a peak at 2 rows. The dialog also includes a 'Method' dropdown set to 'key collision' and a 'Keying Function' dropdown set to 'fingerprint'. The '8 clusters found' message is displayed at the top right of the dialog.

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

Screenshot-3

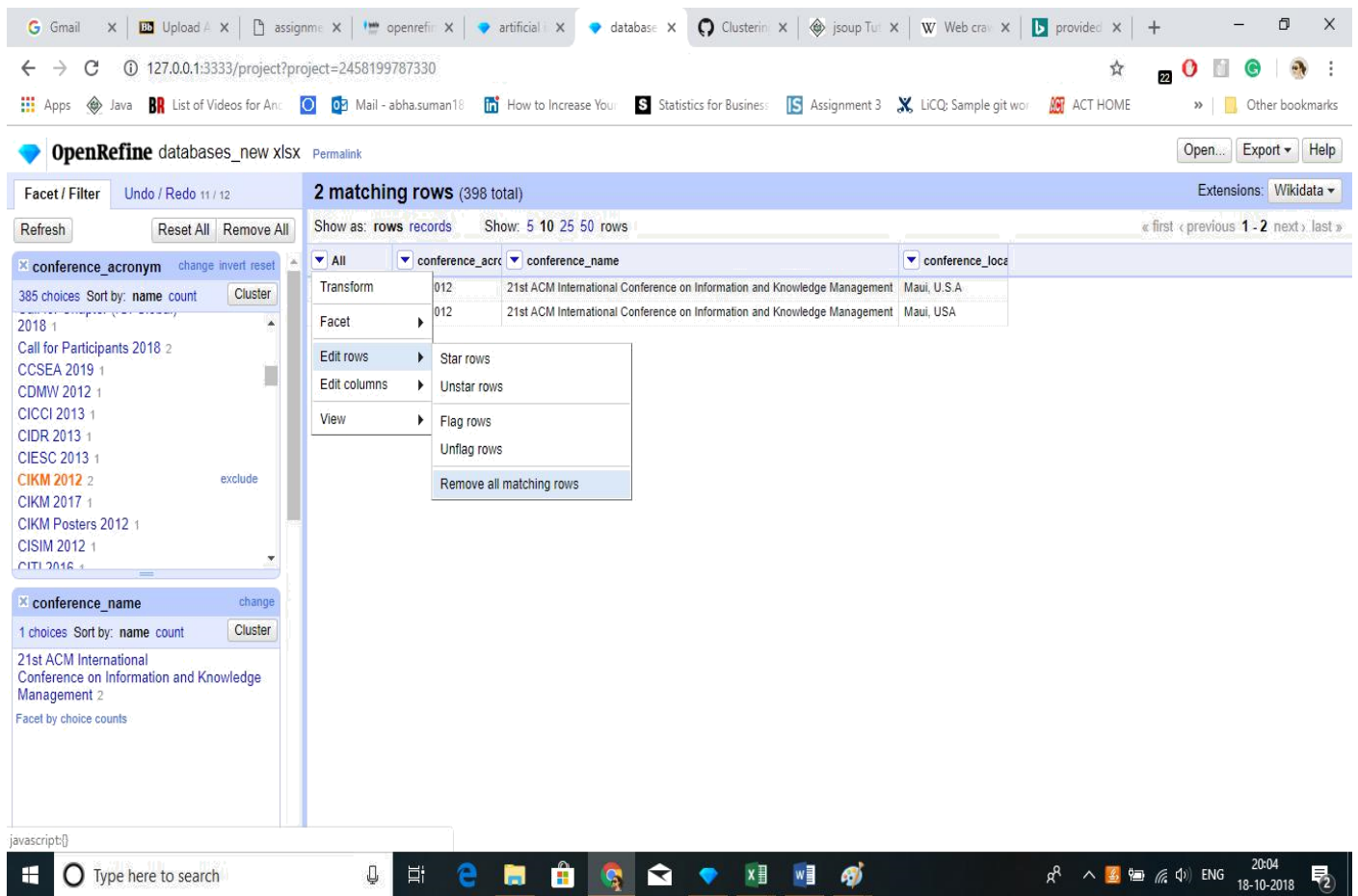
Facet is applied to acronym column to find the duplicate records where the same conference is listed multiple times. The duplicate rows are flagged and removed.

The screenshot shows the OpenRefine web interface. The browser address bar indicates the URL `127.0.0.1:3333/project?project=2458199787330`. The OpenRefine title bar shows the project name `OpenRefine databases_new.xlsx`. The **Facet / Filter** panel on the left shows a facet on the `conference_acronym` column. The facet list includes various conference acronyms, with `CIKM 2012` highlighted in orange and showing a count of 2. The **Facet by choice counts** section shows the selected choice `21st ACM International Conference on Information and Knowledge Management` with a count of 2. The main table displays 2 matching rows for `CIKM 2012`, both of which are identical, indicating duplicate records. The table columns are `conference_acronym`, `conference_name`, and `conference_location`. The table shows two rows with identical data: `CIKM 2012`, `21st ACM International Conference on Information and Knowledge Management`, and `Maui, U.S.A`. The table also shows the total number of rows (398) and the number of rows displayed (2).

	conference_acronym	conference_name	conference_location
316	CIKM 2012	21st ACM International Conference on Information and Knowledge Management	Maui, U.S.A
317	CIKM 2012	21st ACM International Conference on Information and Knowledge Management	Maui, USA

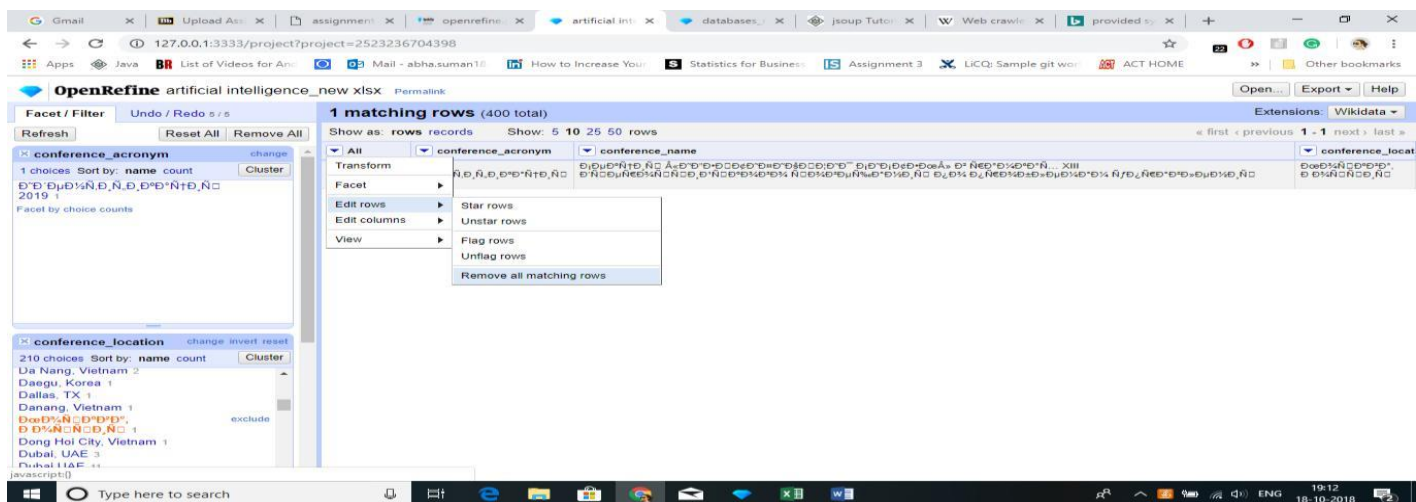
Screenshot-4

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