CS 235 Data Mining Techniques

Suchitra Pithavath

Student-ID: 862121654

Assignment Phase1

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:

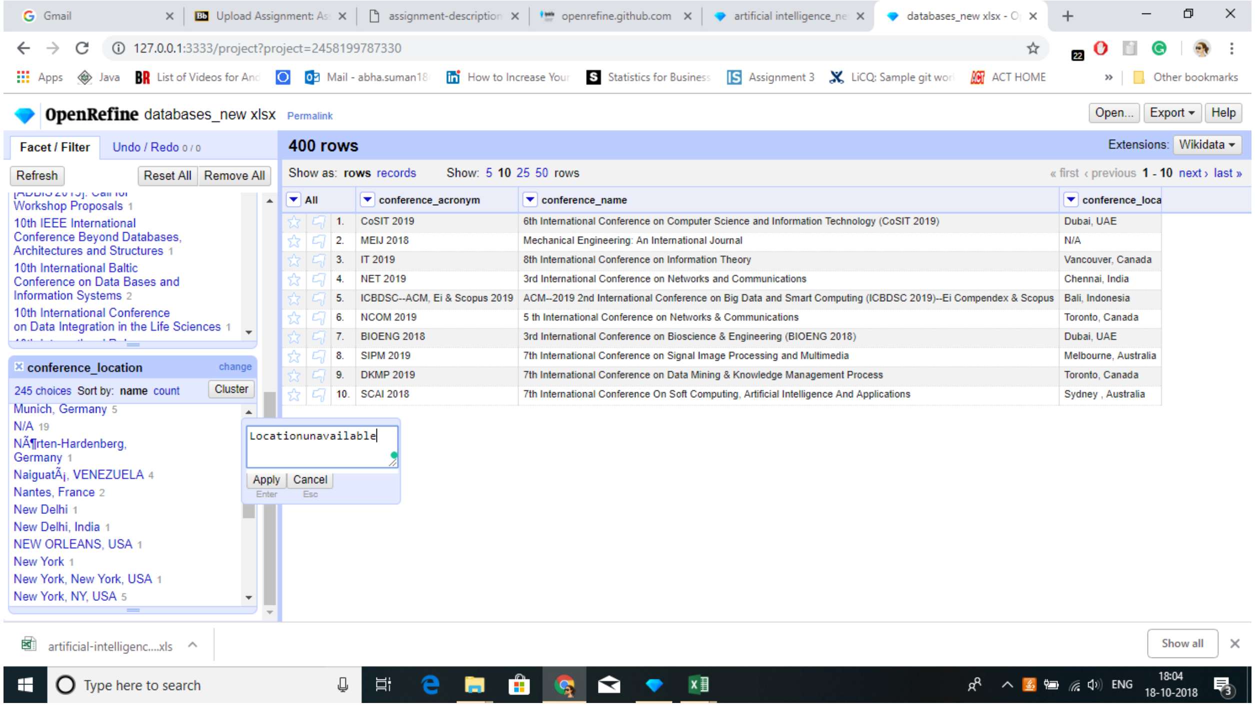
It was challenging to extract the location value .because the table row data with location value had only one attribute align="left." Differentiating from other row data of the HTML table was difficult. Using Jsoup API selected the first row among three-row data of the table having "align=left" attribute.

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:


OpenRefine databases_new xlsx
Permalink
Open... Export Help
Facet / Filter Undo / Redo 0 / 0
Refresh Reset All Remove All
Extensions: Wikidata

400 rows

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All	conference_acronym	conference_name	conference_loca
1.	CoSIT 2019	6th International Conference on Computer Science and Information Technology (CoSIT 2019)	Dubai, UAE
2.	MEIJ 2018	Mechanical Engineering: An International Journal	N/A
3.	IT 2019	8th International Conference on Information Theory	Vancouver, Canada
4.	NET 2019	3rd International Conference on Networks and Communications	Chennai, India
5.	ICBDSC--ACM, Ei & Scopus 2019	ACM--2019 2nd International Conference on Big Data and Smart Computing (ICBDSC 2019)--Ei Compendex & Scopus	Bali, Indonesia
6.	NCOM 2019	5 th International Conference on Networks & Communications	Toronto, Canada
7.	BIOENG 2018	3rd International Conference on Bioscience & Engineering (BIOENG 2018)	Dubai, UAE
8.	SIPM 2019	7th International Conference on Signal Image Processing and Multimedia	Melbourne, Australia
9.	DKMP 2019	7th International Conference on Data Mining & Knowledge Management Process	Toronto, Canada
10.	SCAI 2018	7th International Conference On Soft Computing, Artificial Intelligence And Applications	Sydney , Australia

Workshop Proposals 1
10th IEEE International Conference Beyond Databases, Architectures and Structures 1
10th International Baltic Conference on Data Bases and Information Systems 2
10th International Conference on Data Integration in the Life Sciences 1

conference_location change
245 choices Sort by: name count Cluster

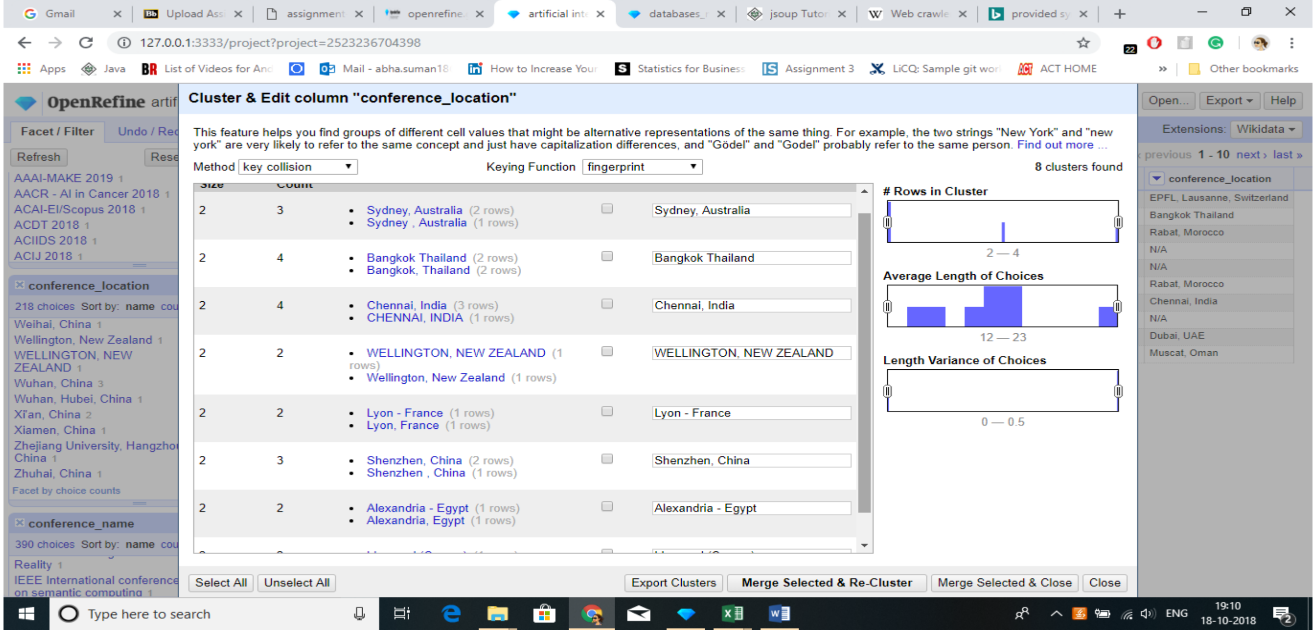
Munich, Germany 5
N/A 19
NÃ¶rten-Hardenberg, Germany 1
NaiguatÃ¡, VENEZUELA 4
Nantes, France 2
New Delhi 1
New Delhi, India 1
NEW ORLEANS, USA 1
New York 1
New York, New York, USA 1
New York, NY, USA 5

Locationunavailable

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.



Cluster & Edit column "conference_location"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision Keying Function fingerprint 8 clusters found

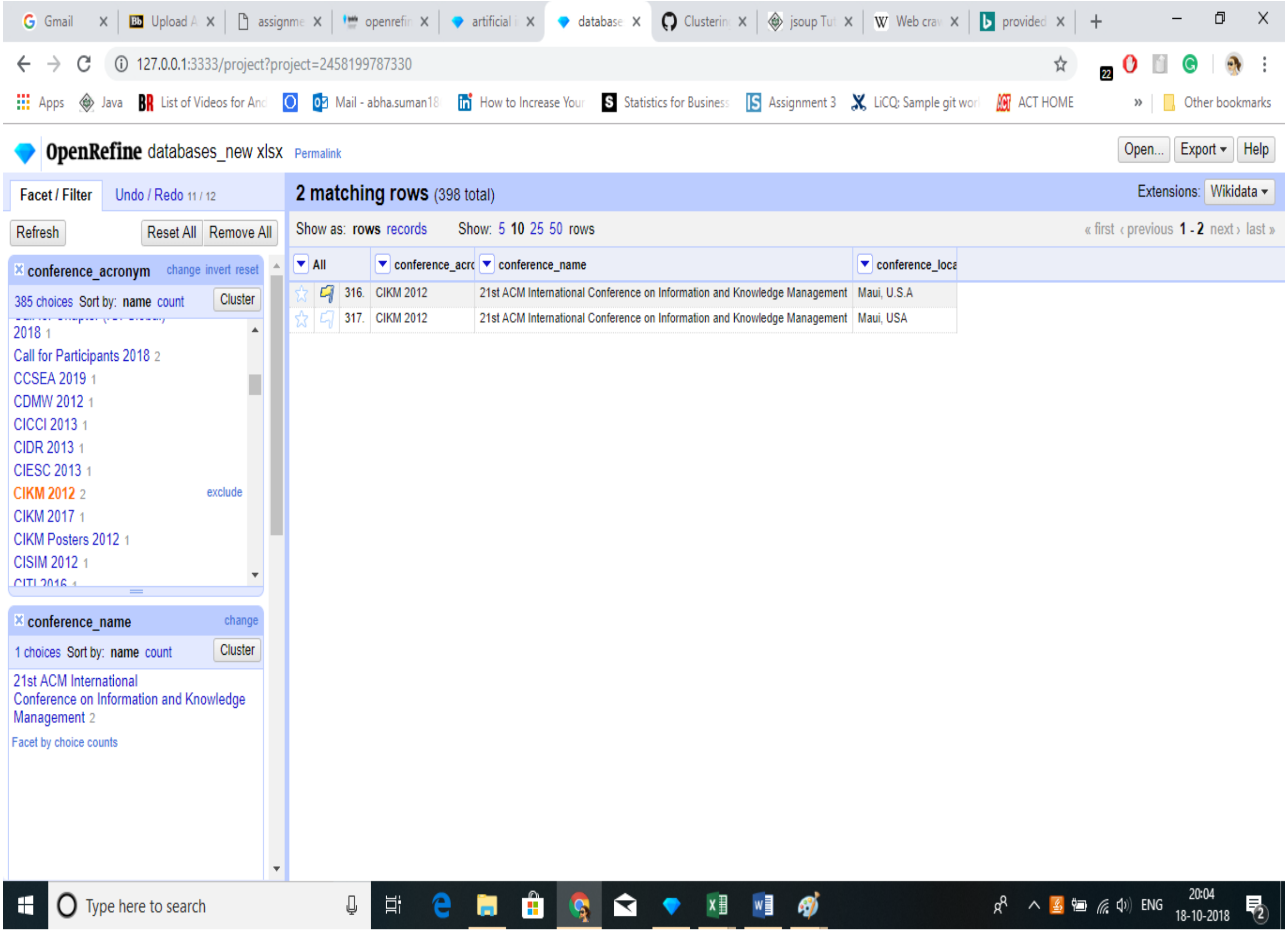
Size	Count	Values in Cluster	Merge?	New Cell Value
2	3	Sydney, Australia (2 rows) Sydney , Australia (1 rows)	□	Sydney, Australia
2	4	Bangkok Thailand (2 rows) Bangkok, Thailand (2 rows)	□	Bangkok Thailand
2	4	Chennai, India (3 rows) CHENNAI, INDIA (1 rows)	□	Chennai, India
2	2	WELLINGTON, NEW ZEALAND (1 rows) Wellington, New Zealand (1 rows)	□	WELLINGTON, NEW ZEALAND
2	2	Lyon - France (1 rows) Lyon, France (1 rows)	□	Lyon - France
2	3	Shenzhen, China (2 rows) Shenzhen , China (1 rows)	□	Shenzhen, China
2	2	Alexandria - Egypt (1 rows) Alexandria, Egypt (1 rows)	□	Alexandria - Egypt

# Rows in Cluster
2 — 4
Average Length of Choices
12 — 23
Length Variance of Choices
0 — 0.5
Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close
Type here to search
ENG
19:10
18-10-2018

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.

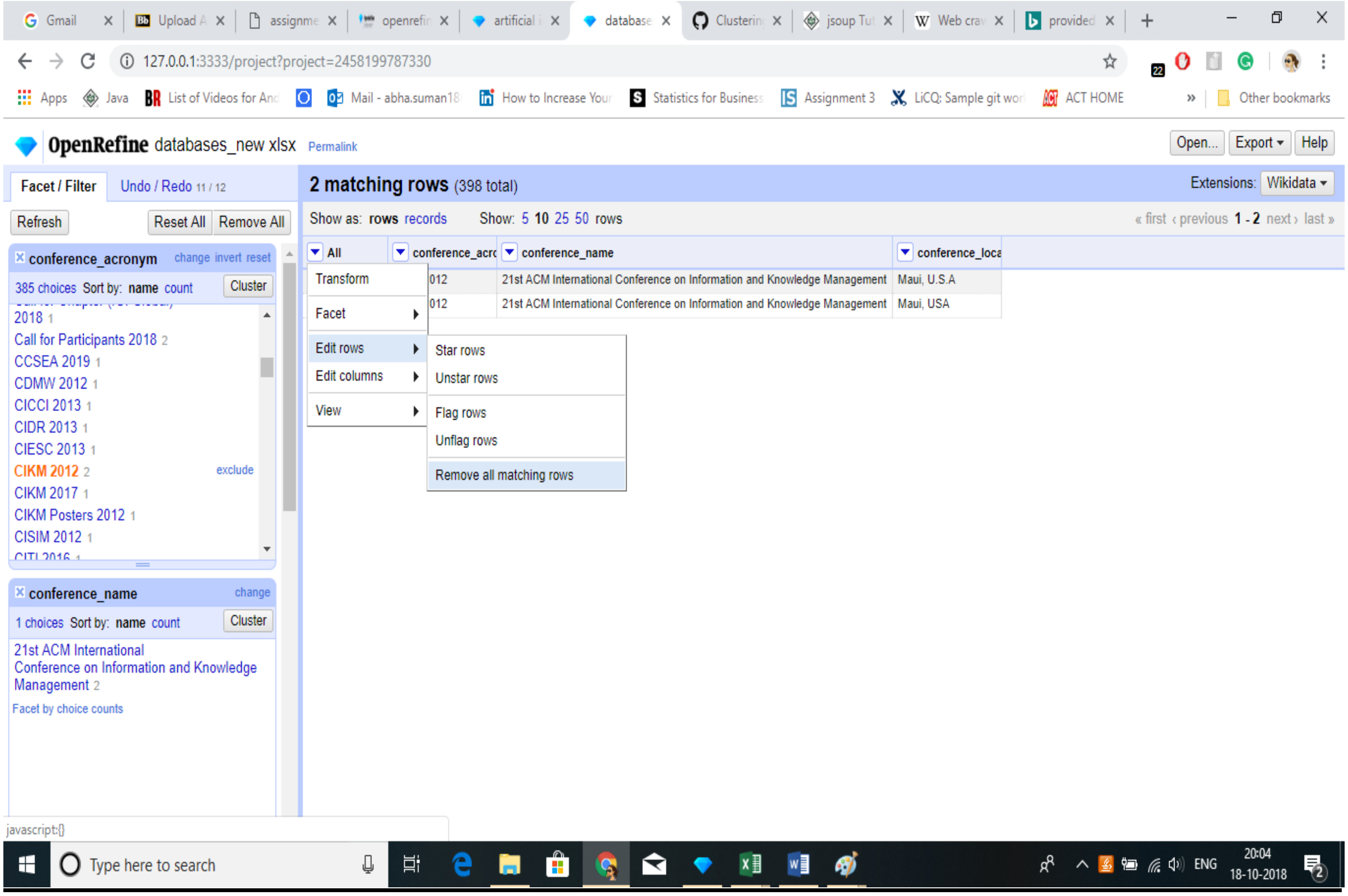

127.0.0.1:3333/project?project=2458199787330
OpenRefine databases_new xlsx Permalink
Open... Export Help
Facet / Filter Undo / Redo 11 / 12
Refresh Reset All Remove All
conference_acronym change invert reset
385 choices Sort by: name count Cluster
2018 1
Call for Participants 2018 2
CCSEA 2019 1
CDMW 2012 1
CICCI 2013 1
CIDR 2013 1
CIESC 2013 1
CIKM 2012 2 exclude
CIKM 2017 1
CIKM Posters 2012 1
CISIM 2012 1
CITI 2016 1
conference_name change
1 choices Sort by: name count Cluster
21st ACM International Conference on Information and Knowledge Management 2
Facet by choice counts
Extensions: Wikidata
Show as: rows records Show: 5 10 25 50 rows
« first ‹ previous 1 - 2 next › last »
Type here to search
ENG 20:04 18-10-2018

2 matching rows (398 total)

All	conference_acro	conference_name	conference_loca
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:



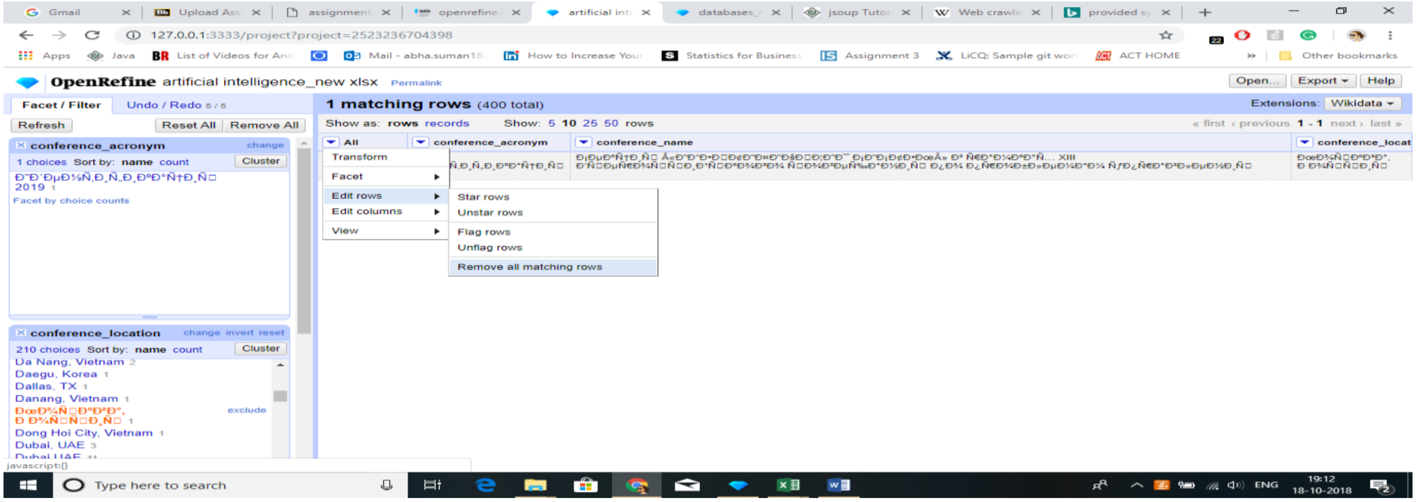
2 matching rows (398 total)

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 2 next › last »

All	conference_acrc	conference_name	conference_loca
	012	21st ACM International Conference on Information and Knowledge Management	Maui, U.S.A
	012	21st ACM International Conference on Information and Knowledge Management	Maui, USA

Screenshot-5

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



1 matching rows (400 total)

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 1 next › last »

All	conference_acronym	conference_name	conference_locat
	[illegible]	[illegible] XIII [illegible]	[illegible]

References:

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

OpenRefine