

## Assignment Phase 1 Report

### Data Mining

For the data mining portion, I used the provided Java code to build my crawler. The crawler is located in the code.txt file that I uploaded. I use the Java jsoup library which is a HTML parser. By looking for the “tbody” element in the html code, I can find the table the data is located in. After that, I extracted each row by finding all the “tr” elements within the table and then all the “td” elements in each row that gives me the data for each column that I needed. The data is then written into four different .txt files (one for each type of conference) where each column is tab-separated and each row is separated by the newline character.

### Data Cleaning

For the data cleaning portion, I utilized OpenRefine. I included a few screenshots attached to the end of this report of the process. The first thing I did was realign some of the rows that were shifted due to an extra entry of “expired CFPs”. After the first column was shifted, I went on to search for all locations that were missing (which were marked with N/A) and deleted these rows. Then I took advantage of the clustering feature of OpenRefine (Fig 1) and was able to cluster some acronyms and location names (not so much conference names). After that I used the “Facet” feature to go through all the acronym names and edited the ones that had extras that were not actually part of the acronyms (Fig 2). These included “EI”, “IEEE”, “Scopus”, etc. The clustering was then used to clean the acronyms some more and then I searched for conference acronyms that were duplicates. This was because any duplicates are

the same conference that happened in the same year but they might have had different announcements, deadlines, or type. We don't want to count these types of conferences twice during the analysis portion, or the results will be come skewed to how many duplicate conferences happened on the site we crawled through. After that, I did a final run though by doing a facet on the dates (or the duration of the conference) to see if any conference with duplicate durations were the same but called differently and found a few of these which were flagged and removed. Finally, I removed the unneeded columns and was left with the three columns (acronym, name, location) for the final data (Fig 3).

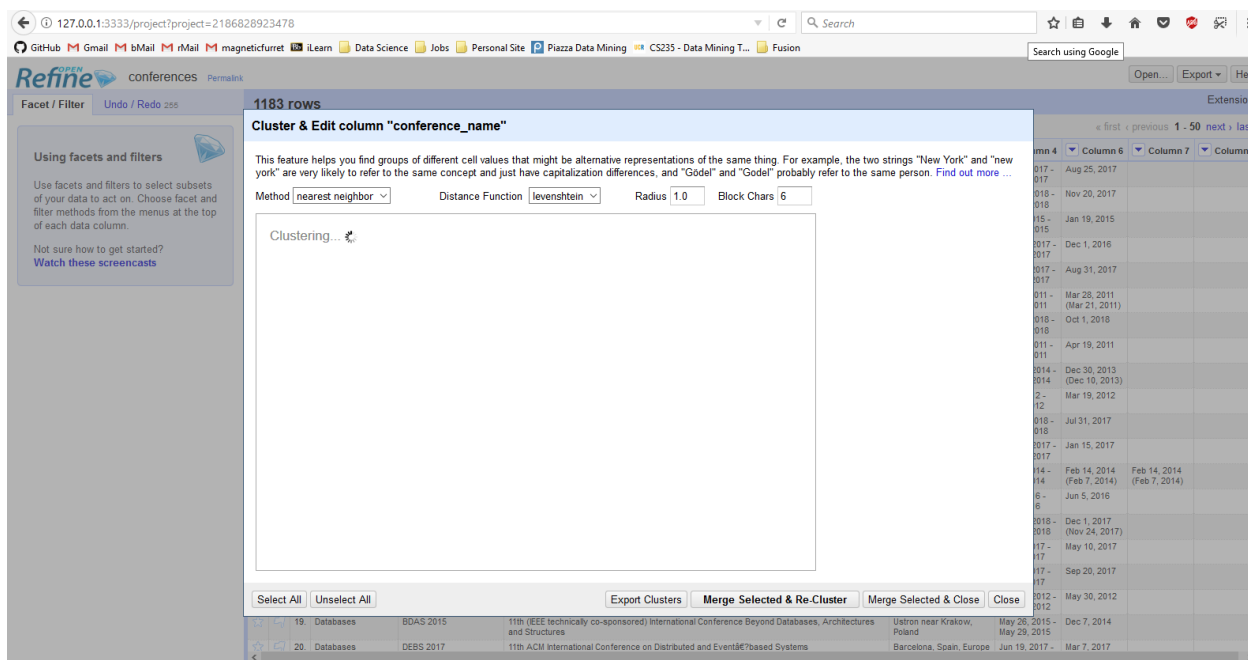


Fig 1: Here I clustered all the conference acronyms, conference names, and conference locations.

Refine<sup>OPEN</sup> conferences Permalink

Blank down 25 cells in column conference\_acronym Undo

Open... Export... Hi

Facet / Filter Undo / Redo 496

1146 rows

Show as: rows records Show: 5 10 25 50 rows

« first « previous 1 - 50 next » la

		conference_type	conference_acr	conference_name	conference_location	Column 4	Column 6	Column 7	Column 8
1.	AI	AAAI 2018	AAAI-18 Workshop on artificial intelligence applied to assistive technologies and smart environments	New Orleans, USA	Feb 2, 2018 - Feb 3, 2018	Oct 20, 2017			
2.	AI		The Thirty-Second AAAI Conference on Artificial Intelligence	New Orleans, USA	Feb 2, 2018 - Feb 7, 2018	Sep 11, 2017 (Sep 8, 2017)			
3.	AI	AJIA 2017	12th International Symposium Advances in Artificial Intelligence and Applications	Prague, Czech Republic	Sep 3, 2017 - Sep 6, 2017	May 16, 2017			
4.	AI	AAMAS 2018	International Conference on Autonomous Agents and Multiagent Systems (AAMAS-18)	Stockholm	Jul 10, 2018 - Jul 15, 2018	Nov 14, 2017 (Nov 10, 2017)			
5.	ML	ACCV 2016	ACCV'16 Workshop on Multi-view Lip-reading & Audio-visual Challenges	Taipei, Taiwan	Nov 20, 2016 - Nov 20, 2016	Aug 20, 2016			
6.	Databases	ACIDS 2012	The 4th Asian Conference on Intelligent Information and Database Systems	Kaohsiung, Taiwan	Mar 19, 2012 - Mar 21, 2012	Sep 15, 2011	Sep 15, 2011		
7.	Databases	ACIDS 2013	Special Session on Intelligent Recommended Systems	Kuala Lumpur, Malaysia	Mar 18, 2013 - Mar 20, 2013	Nov 5, 2012			
8.	AI	ACIDS 2018	10th Asian Conference on Intelligent Information and Database Systems	Dong Hoi City, Vietnam	Mar 19, 2018 - Mar 21, 2018	Oct 1, 2018			
9.	Data Mining	ACIS 2016	The Fifth Asian Conference on Information Systems	Krabi, Thailand	Oct 27, 2016 - Oct 29, 2016	Jun 20, 2016			
10.	AI	ACL 2017	The Third International Conference on Arabic Computational Linguistics	Dubai, UAE	Nov 5, 2017 - Nov 6, 2017	Jun 20, 2017			
11.	AI	ACL 2018	56th Annual Meeting of the Association for Computational Linguistics	Melbourne, Australia	Jul 15, 2018 - Jul 20, 2018	Feb 22, 2018			
12.	Data Mining	ACML 2016	8th Asian Conference on Machine Learning	Hamilton, New Zealand	Nov 16, 2016 - Nov 18, 2016	Aug 15, 2016			
13.	AI	ACML 2017	The 9th Asian Conference on Machine Learning	Seoul, Korea	Nov 15, 2017 - Nov 17, 2017	Aug 5, 2017			
14.	AI	ACMLC 2017	2017 Asia Conference on Machine Learning and Computing (ACMLC 2017)—EI Compendex, Scopus	Singapore, Singapore	Dec 8, 2017 - Dec 10, 2017	Sep 30, 2017			
15.	Databases	ACTIVE 2017	First International Workshop on Data Management on Virtualized Active Systems (in Conjunction with the IEEE ICDE 2017)	San Diego, CA	Apr 22, 2017 - Apr 22, 2017	Dec 23, 2016			
16.	Data Mining	ACUMEN 2017	Data Science for Human Performance in Social Networks	ICDM'17, New Orleans, USA	Nov 18, 2017 - Nov 18, 2017	Aug 7, 2017			
17.	Databases	ADBS 2011	Fifteenth East-European Conference on Advances in Databases and Information Systems	Vienna, Austria	Sep 19, 2011 - Sep 23, 2011	Apr 5, 2011			
18.	Databases	ADBS 2015	[ADBS'2015]: Call for Workshop Proposals	Polier, France	Sep 8, 2015 - Sep 11, 2015	Jan 19, 2015			

Fig 2: By sorting all the rows alphabetically and then using the “Blank down” option, I can find any duplicate conference\_acronyms. Then using the “Facet” feature I can find all the blank acronyms and delete them.

Refine<sup>OPEN</sup> conferences Permalink

Facet / Filter Undo / Redo 501

1121 rows

Show as: rows records Show: 5 10 25 50 rows

« first « previous 1 - 50 next » la

		conference_acr	conference_name	conference_location	Column 8
1.	AAAI 2018	AAAI-18 Workshop on artificial intelligence applied to assistive technologies and smart environments	New Orleans, USA		
2.	AAIA 2017	12th International Symposium Advances in Artificial Intelligence and Applications	Prague, Czech Republic		
3.	AAMAS 2018	International Conference on Autonomous Agents and Multiagent Systems (AAMAS-18)	Stockholm		
4.	ACCV 2016	ACCV'16 Workshop on Multi-view Lip-reading & Audio-visual Challenges	Taipei, Taiwan		
5.	ACIDS 2012	The 4th Asian Conference on Intelligent Information and Database Systems	Kaohsiung, Taiwan		
6.	ACIDS 2013	Special Session on Intelligent Recommended Systems	Kuala Lumpur, Malaysia		
7.	ACIDS 2018	10th Asian Conference on Intelligent Information and Database Systems	Dong Hoi City, Vietnam		
8.	ACIS 2016	The Fifth Asian Conference on Information Systems	Krabi, Thailand		
9.	ACL 2017	The Third International Conference on Arabic Computational Linguistics	Dubai, UAE		
10.	ACL 2018	56th Annual Meeting of the Association for Computational Linguistics	Melbourne, Australia		
11.	ACML 2016	8th Asian Conference on Machine Learning	Hamilton, New Zealand		
12.	ACML 2017	The 9th Asian Conference on Machine Learning	Seoul, Korea		
13.	ACMLC 2017	2017 Asia Conference on Machine Learning and Computing (ACMLC 2017)—EI Compendex, Scopus	Singapore, Singapore		
14.	ACTIVE 2017	First International Workshop on Data Management on Virtualized Active Systems (in Conjunction with the IEEE ICDE 2017)	San Diego, CA		
15.	ACUMEN 2017	Data Science for Human Performance in Social Networks	ICDM'17, New Orleans, USA		
16.	ADBS 2011	Fifteenth East-European Conference on Advances in Databases and Information Systems	Vienna, Austria		
17.	ADBS 2015	[ADBS'2015]: Call for Workshop Proposals	Polier, France		
18.	ADBS 2017	21st European Conference on Advances in Databases and Information Systems	Nicosia, Cyprus		
19.	ADC 2012	Australasian Database Conference	Melbourne, Australia		
20.	ADC 2014	The 25th Australasian Database Conference	Brisbane, Australia		
21.	ADCOM 2017	3rd International conference on Advanced Computing	Dubai, UAE		
22.	ADMA 2016	Advanced Data Mining and Applications	Gulf Coast		
23.	ADMA 2017	The 13th International Conference on Advanced Data Mining and Applications	Singapore, Singapore		
24.	ADMI 2017	Agents and Data Mining Interaction@AAMAS2017	Sao Paulo, Brazil		
25.	ADS 2017	ACIDS 2017 Special Session on Applications of Data Science	Kanazawa, Japan		
26.	AECIA 2016	Third International Afro-European Conference for Industrial Advancement	Marrakech, Morocco		
27.	AECM 2014	Workshop on Applied Enterprise Content Management	Munich, Germany		
28.	AFRICOMM 2016	8th EAI International Conference on e&E7nfrastucture and e&E7Services for Developing Countries	Ouagadougou, Burkina Faso (West Africa)		
29.	AGI 2017	Artificial General Intelligence	Melbourne, Australia		
30.	AgTAm 2017	The International Workshop on Agent Technology for Ambient Intelligence	Bucharest, Romania		

Fig 3: I cleaned the data through clustering and changing the conference acronyms and locations to be more uniform compared to the raw data. This picture is the final cleaned data.