Satamalee, Panitan
CS 235, Data Mining

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).

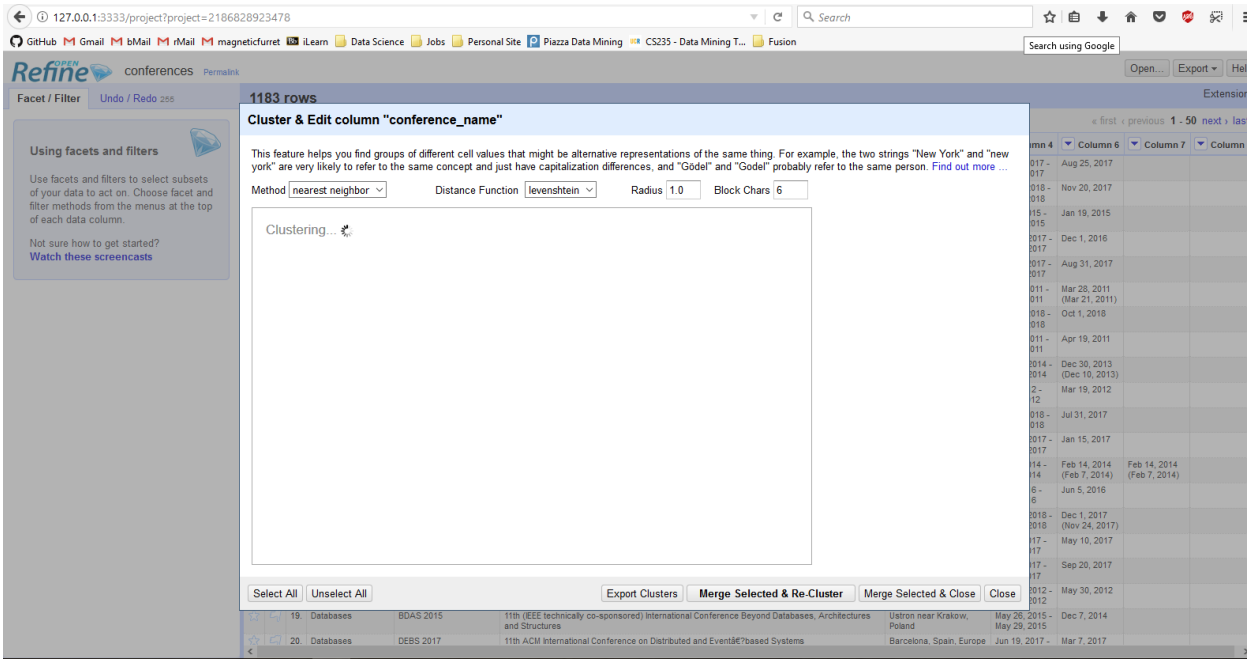

Fig 1: Here I clustered all the conference acronyms, conference names, and conference locations.

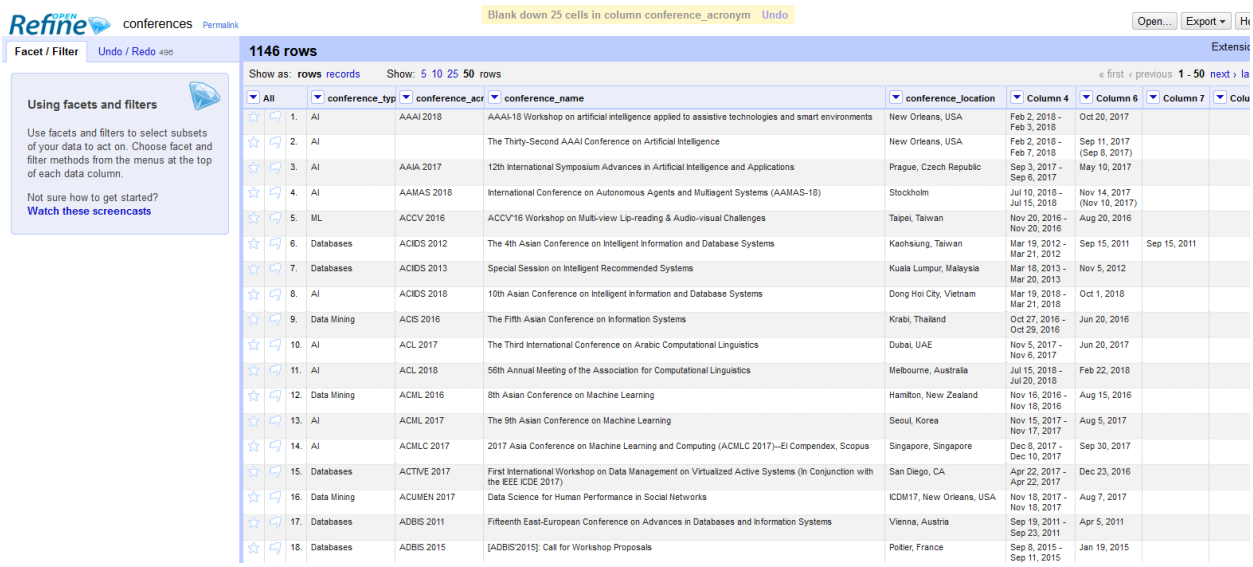

Fig 2: By sorting all the rows alphabetically and then using the “Blank down” option, I can find any duplicate conference_acronyms. Then using the “Facet” feature I can find all the blank acronyms and delete them.

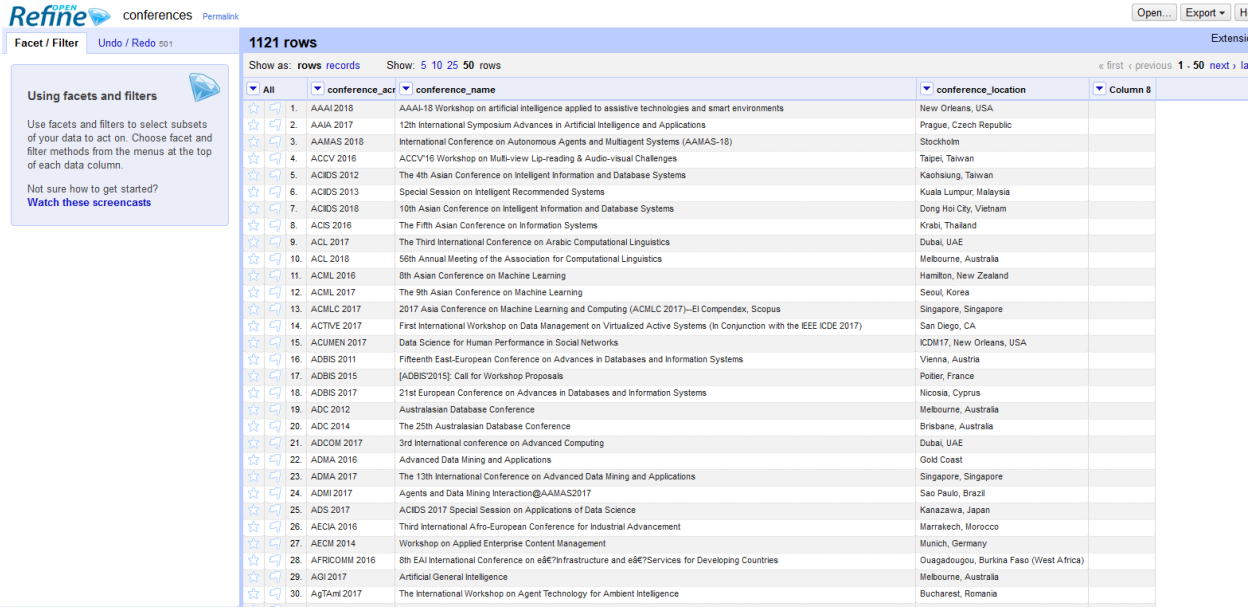

Fig 3: I cleaned the data through clustering and changing the conference acronyms and locations to be more uniform compared to the raw data. This picture is the final cleaned data.