CS235 Assignment 1, Phase 1

Alvin Thai, 861103107

October 20th, 2017

My approach to Crawling WikiCFP

After taking a look at the source html for one of the pages of WikiCFP, I figured that the necessary steps would be:

- Parse through the html to obtain the contents of the table, exclusively.
- Determine which of the resulting text to output.

I included the jsoup html parsing library to assist with the first step. With some experimentation, I managed to be able to use the tools provided in jsoup to extract the data only from the table of conferences/journals. From here I decided the best thing to do to output the proper information would be to store all the info (i.e. conference acronyms, conference names, and conference locations) in a vector.

There was an error, however - On page 4 of the first set of WikiCFP topics, there was a column that said "Expired CFPs" that was being still being read. As a fix, I added a conditional expression that would only save text to the vector if it did not read "Expired CFPs". This worked for all other crawled WikiCFP pages.

Once that was in the vector, I was able to determine a pattern in the strings and used modulo arithmetic to determine what to output. More specifically, after parsing, the data was stored in a specific order:

1. Conference Acronym
2. Conference Name
3. <newline>
4. Date range
5. Conference Location
6. Date

...and so on, so forth. Thus, I was able to determine that I only had to print out the first, second, and fifth elements in any set of 6. *(See code for implementation)*

As an aside, it should be mentioned that this is by no means the most efficient way to implement this crawler, though it was the first algorithm I came up with that made sense to implement.

Cleaning the Data

To clean the data, I applied a text facet that would allow me to easily view and sort through the rows of locations.

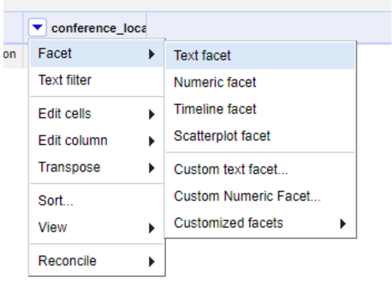
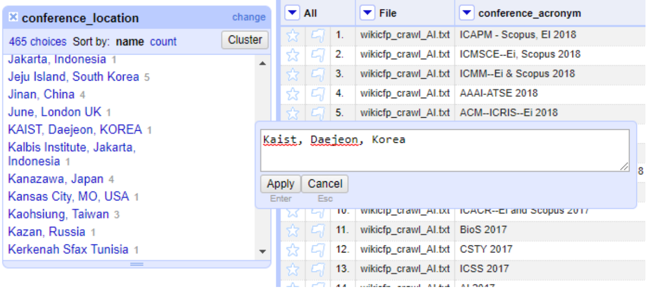


Figure 1: Adding a text facet to location



465 choices

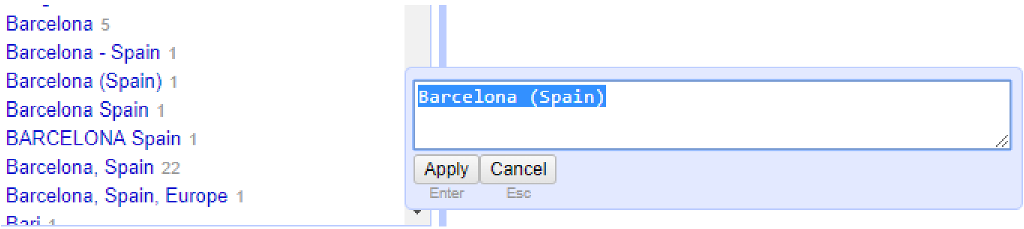
Jakarta, Indonesia 1
Jeju Island, South Korea 5
Jinan, China 4
June, London UK 1
KAIST, Daejeon, KOREA 1
Kalbis Institute, Jakarta, Indonesia 1
Kanazawa, Japan 4
Kansas City, MO, USA 1
Kaohsiung, Taiwan 3
Kazan, Russia 1
Kerkenah Sfax Tunisia 1

	File	conference_acronym
1.	wikicfp_crawl_AI.txt	ICAPM - Scopus, EI 2018
2.	wikicfp_crawl_AI.txt	ICMSCE--Ei, Scopus 2018
3.	wikicfp_crawl_AI.txt	ICMM--Ei & Scopus 2018
4.	wikicfp_crawl_AI.txt	AAAI-ATSE 2018
5.	wikicfp_crawl_AI.txt	ACM--ICRIS--Ei 2018
10.	wikicfp_crawl_AI.txt	ICACR--EI and Scopus 2017
11.	wikicfp_crawl_AI.txt	BioS 2017
12.	wikicfp_crawl_AI.txt	CSTY 2017
13.	wikicfp_crawl_AI.txt	ICSS 2017

Kaist, Daejeon, Korea

Figure 2: Text Facet List

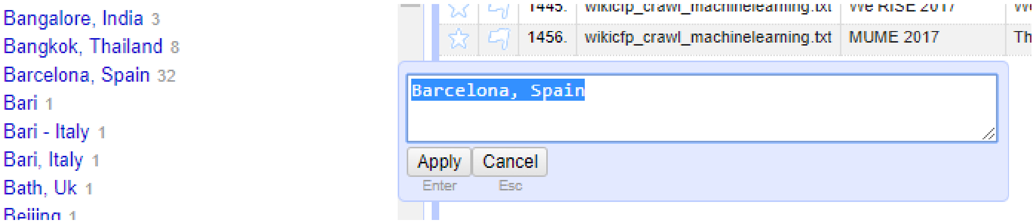
I decided to approach cleaning the data in a rather conservative way. I only made changes to locations that were either misspelled or represented in multiple ways. For example, the city of Barcelona, Spain was represented 7 different ways. To consolidate these, I opted for the most common way it was written.



Barcelona 5
Barcelona - Spain 1
Barcelona (Spain) 1
Barcelona Spain 1
BARCELONA Spain 1
Barcelona, Spain 22
Barcelona, Spain, Europe 1
Bari 1

Barcelona (Spain)

Figure 3: Barcelona, Spain before cleaning



Bangalore, India 3
Bangkok, Thailand 8
Barcelona, Spain 32
Bari 1
Bari - Italy 1
Bari, Italy 1
Bath, Uk 1
Beijing 1

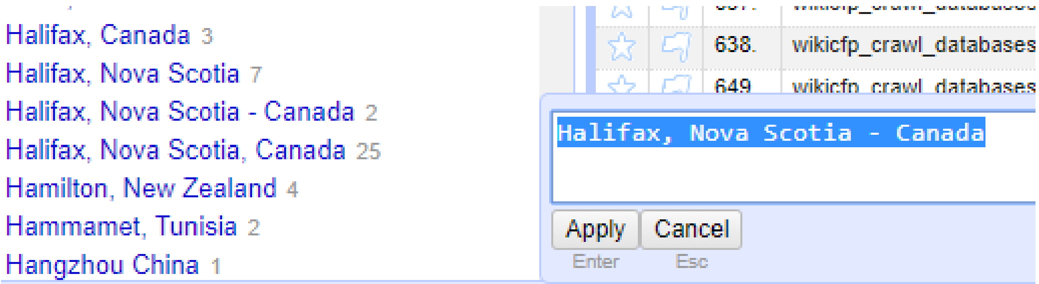
1456. wikicfp_crawl_machinelearning.txt MUME 2017 Th

Barcelona, Spain

Figure 4: Barcelona, Spain after cleaning

Similarly, cities like Halifax in Canada, Aberdeen in Scotland, as well as many more in the dataset were represented in more than one way. Also, there were cases where the location was only the name of a city. That required looking up the convention and confirming which country it was held in. When it came to choosing how to rename the location, I made sure it had proper capitalization and at least had the format of (*City, Country*), (*City, State*) or (*City, State, Country*).

That all being said, it should be noted that the specs for this lab were to only gather conference data, and thus I had to take care of Journal data during cleanup. It just so happened that when inspecting items with a location of "N/A", they all turned out to be journal publications. Thus, I chose to remove them entirely from the dataset.

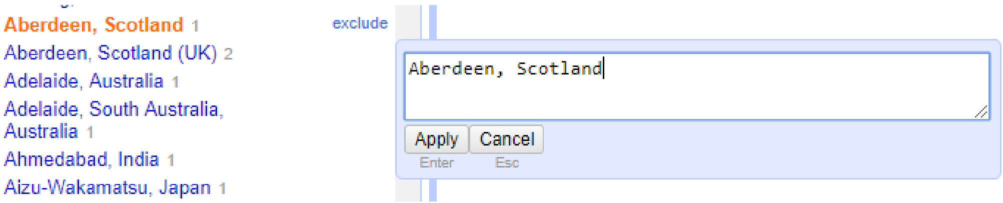


Halifax, Canada 3
Halifax, Nova Scotia 7
Halifax, Nova Scotia - Canada 2
Halifax, Nova Scotia, Canada 25
Hamilton, New Zealand 4
Hammamet, Tunisia 2
Hangzhou China 1

638. wikicfp_crawl_databases
649 wikicfp_crawl_databases

Halifax, Nova Scotia - Canada

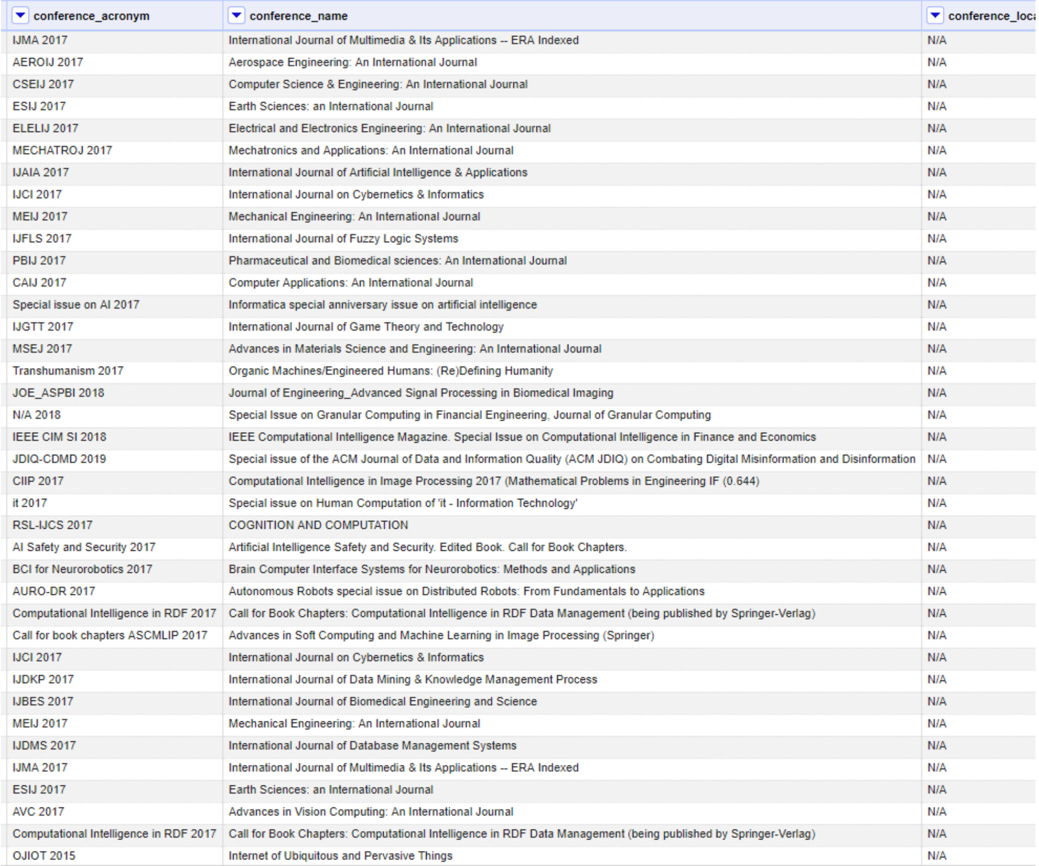
Figure 5: Halifax before cleaning



Aberdeen, Scotland 1
Aberdeen, Scotland (UK) 2
Adelaide, Australia 1
Adelaide, South Australia, Australia 1
Ahmedabad, India 1
Aizu-Wakamatsu, Japan 1

Aberdeen, Scotland

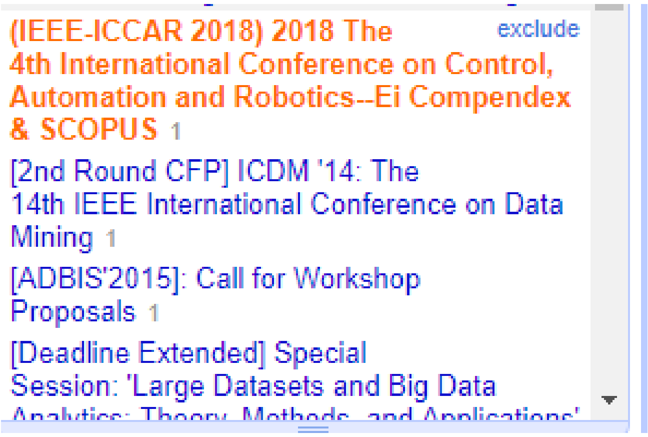
Figure 6: Aberdeen before cleaning



conference_acronym	conference_name	conference_loca
IJMA 2017	International Journal of Multimedia & Its Applications -- ERA Indexed	N/A
AEROIJ 2017	Aerospace Engineering: An International Journal	N/A
CSEIJ 2017	Computer Science & Engineering: An International Journal	N/A
ESIJ 2017	Earth Sciences: an International Journal	N/A
ELELIJ 2017	Electrical and Electronics Engineering: An International Journal	N/A
MECHATROJ 2017	Mechatronics and Applications: An International Journal	N/A
IJAIA 2017	International Journal of Artificial Intelligence & Applications	N/A
IJCI 2017	International Journal on Cybernetics & Informatics	N/A
MEIJ 2017	Mechanical Engineering: An International Journal	N/A
IJFLS 2017	International Journal of Fuzzy Logic Systems	N/A
PBIJ 2017	Pharmaceutical and Biomedical sciences: An International Journal	N/A
CAIJ 2017	Computer Applications: An International Journal	N/A
Special issue on AI 2017	Informatica special anniversary issue on artificial intelligence	N/A
IJGTT 2017	International Journal of Game Theory and Technology	N/A
MSEJ 2017	Advances in Materials Science and Engineering: An International Journal	N/A
Transhumanism 2017	Organic Machines/Engineered Humans: (Re)Defining Humanity	N/A
JOE_ASPBI 2018	Journal of Engineering_Advanced Signal Processing in Biomedical Imaging	N/A
N/A 2018	Special Issue on Granular Computing in Financial Engineering, Journal of Granular Computing	N/A
IEEE CIM SI 2018	IEEE Computational Intelligence Magazine. Special Issue on Computational Intelligence in Finance and Economics	N/A
JDIQ-CDMD 2019	Special issue of the ACM Journal of Data and Information Quality (ACM JDIQ) on Combating Digital Misinformation and Disinformation	N/A
CIIP 2017	Computational Intelligence in Image Processing 2017 (Mathematical Problems in Engineering IF (0.644)	N/A
it 2017	Special issue on Human Computation of 'it - Information Technology'	N/A
RSL-IJCS 2017	COGNITION AND COMPUTATION	N/A
AI Safety and Security 2017	Artificial Intelligence Safety and Security. Edited Book. Call for Book Chapters.	N/A
BCI for Neurorobotics 2017	Brain Computer Interface Systems for Neurorobotics: Methods and Applications	N/A
AURO-DR 2017	Autonomous Robots special issue on Distributed Robots: From Fundamentals to Applications	N/A
Computational Intelligence in RDF 2017	Call for Book Chapters: Computational Intelligence in RDF Data Management (being published by Springer-Verlag)	N/A
Call for book chapters ASCMLIP 2017	Advances in Soft Computing and Machine Learning in Image Processing (Springer)	N/A
IJCI 2017	International Journal on Cybernetics & Informatics	N/A
IJDKP 2017	International Journal of Data Mining & Knowledge Management Process	N/A
IJBES 2017	International Journal of Biomedical Engineering and Science	N/A
MEIJ 2017	Mechanical Engineering: An International Journal	N/A
IJDMS 2017	International Journal of Database Management Systems	N/A
IJMA 2017	International Journal of Multimedia & Its Applications -- ERA Indexed	N/A
ESIJ 2017	Earth Sciences: an International Journal	N/A
AVC 2017	Advances in Vision Computing: An International Journal	N/A
Computational Intelligence in RDF 2017	Call for Book Chapters: Computational Intelligence in RDF Data Management (being published by Springer-Verlag)	N/A
OJIOT 2015	Internet of Ubiquitous and Pervasive Things	N/A

Figure 7: List of journals with N/A location

I also went through and cleaned up the conference names, removing any irrelevant text in brackets. Irrelevant text, in this case, being a repeat of its acronym, or some sort of text that bears no relevance to the name.



(IEEE-ICCAR 2018) 2018 The 4th International Conference on Control, Automation and Robotics--Ei Compendex & SCOPUS 1

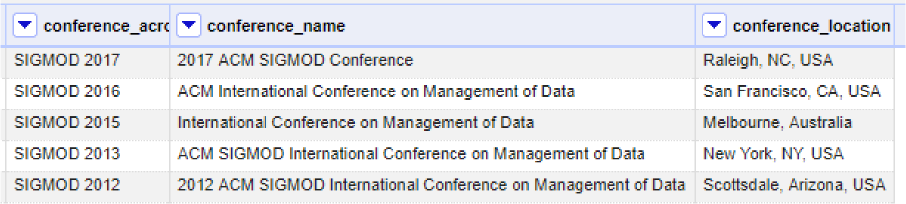
[2nd Round CFP] ICDM '14: The 14th IEEE International Conference on Data Mining 1

[ADBIS'2015]: Call for Workshop Proposals 1

[Deadline Extended] Special Session: 'Large Datasets and Big Data Analytics: Theory, Methods, and Applications'

Figure 8: A few conference names with some redundant information

The reason why I had chosen to do rather conservative cleaning was because I felt that some data should still be retained. For example, I did not go through and thoroughly clean up the conference acronyms because I felt that the acronyms themselves contain rather important information. They act as distinct identifiers for each conference in the dataset - giving us the name of the conference as well as the year that it occurred in.



conference_acro	conference_name	conference_location
SIGMOD 2017	2017 ACM SIGMOD Conference	Raleigh, NC, USA
SIGMOD 2016	ACM International Conference on Management of Data	San Francisco, CA, USA
SIGMOD 2015	International Conference on Management of Data	Melbourne, Australia
SIGMOD 2013	ACM SIGMOD International Conference on Management of Data	New York, NY, USA
SIGMOD 2012	2012 ACM SIGMOD International Conference on Management of Data	Scottsdale, Arizona, USA

Figure 9: Sigmod Conferences

For example, SIGMOD in this dataset has acronyms *SIGMOD 2017*, *SIGMOD 2016*, *SIGMOD 2015*, *SIGMOD 2013*, *SIGMOD 2012*. They could easily all be reduced to just *SIGMOD*, but each consecutive SIGMOD conference was held in a different location. The same applies for other conferences like *DSAA*, *DaWaK*, *QDB*, etc. While having years in the acronym can potentially dirty the data, it does serve as a way to better identify the specific conference it is.