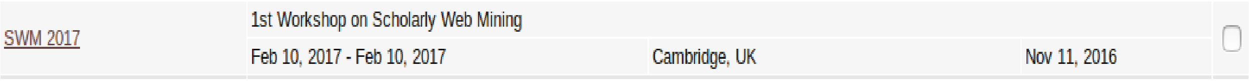
Introduction

In the world of Data mining, the major task is to crawl, clean and retrieve meaningful information from the data. In order to retrieve meaningful information from the **WikiCFP** website, I have crawled it using **Java** and parse the data using **Jsoup: Java HTML Parser**. Then for cleaning **OpenRefine** is used. I have used **Hadoop** to gain an insight from the data. The processes followed throughout the accomplishment of each task is discussed here.

1. Data Crawling

In order to crawl the WikiCFP website for the conferences of data mining, machine learning, database and AI, I have used the provided Java code. In order to fetch the conferences of these 4 categories, I have used a string array named **searchCategory** and fetched 20 pages of each category. After fetching the content of each page, a function named **parseTableData()** is called. In this function, the content of one page is parsed. For data parsing, **Jsoup: Java HTML Parser** is used. While parsing the data of one page, it is observed that there are multiple tables on every page. Among them, table number 2 is the one in which the conference data is displayed. Therefore, table 2 is selected from the parsed data.

After selecting the correct table, each row is considered to fetch the conference information. In the table, each conference information is represented using 2 consecutive rows. Therefore, in order to fetch the information of one conference, two row is selected at a time. From Figure 1, the real scenario can be found, where each conference consists of two rows. The first row contains conference name and the down row contains time, location and deadline of the conference.



SWM 2017	1st Workshop on Scholarly Web Mining			▢
	Feb 10, 2017 - Feb 10, 2017	Cambridge, UK	Nov 11, 2016	

Figure 1: One conference entry

From each conference entry, the required information is collected as follows,

1) Conference Acronym: The conference acronym is found at table column number 0 of first row. So this column is selected for this information. There are some conference entries in which the conference acronym is **Expired CFPs**. These entries are ignored for further processing i.e. the conference name and location is not collected for these entries.

2) Conference Name: After getting the correct conference acronym, the conference name is fetched by selecting column number 1 of first row.

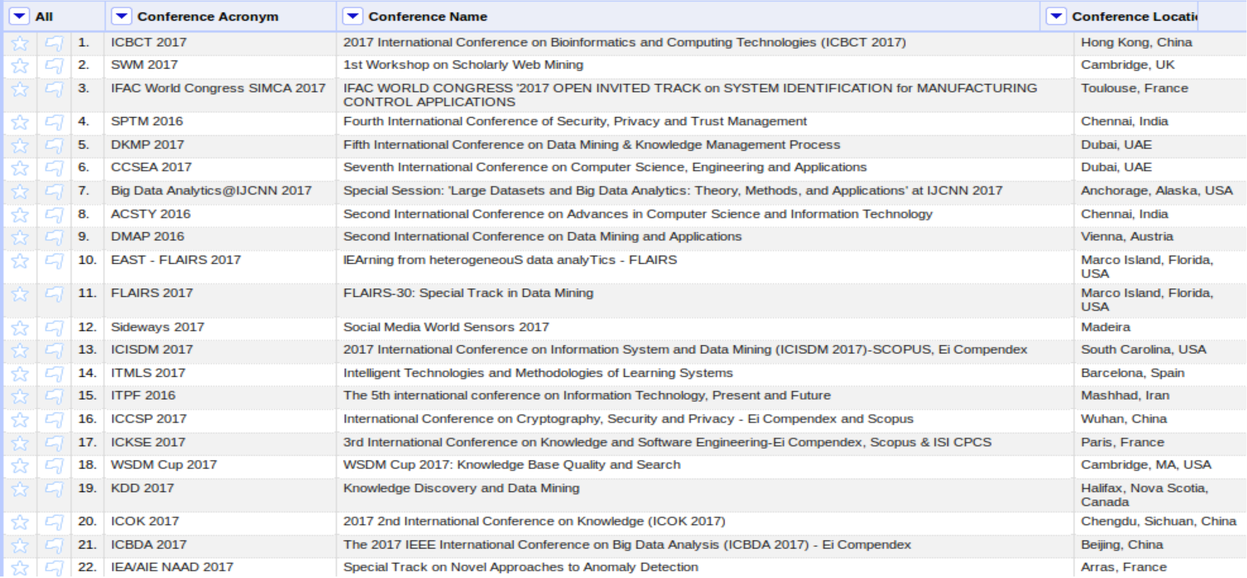
3) Conference Location: In order to fetch the conference location, the second row is selected. In second row, the conference location is found in column 1. Therefore, the conference location is selected from there. Though it is not required to select the time and deadline of the conference, these information are also selected to check if the required information is collected properly.

The three information is appended using tab between them in a string. Then all the conference entries of one page is returned to the calling routine where the returned string is written in the file named **wikicfp_crawl.txt**. The conference information of all categories i.e. Data mining, Artificial Intelligence, Databases and Machine Learning are stored in the same file.

In order to ensure the proper implementation of this portion, I have used netinstructions.com[1] to get an idea about how a java crawler works and Stackoverflow[2] to get an idea about how the table data are stored.

2. Data Cleaning

In the data cleaning portion, **OpenRefine** is used. Since the data were stored in tab separated format, when the file containing all conference data is uploaded in OpenRefine, all the data are automatically organized in a table format by OpenRefine as Figure 2.

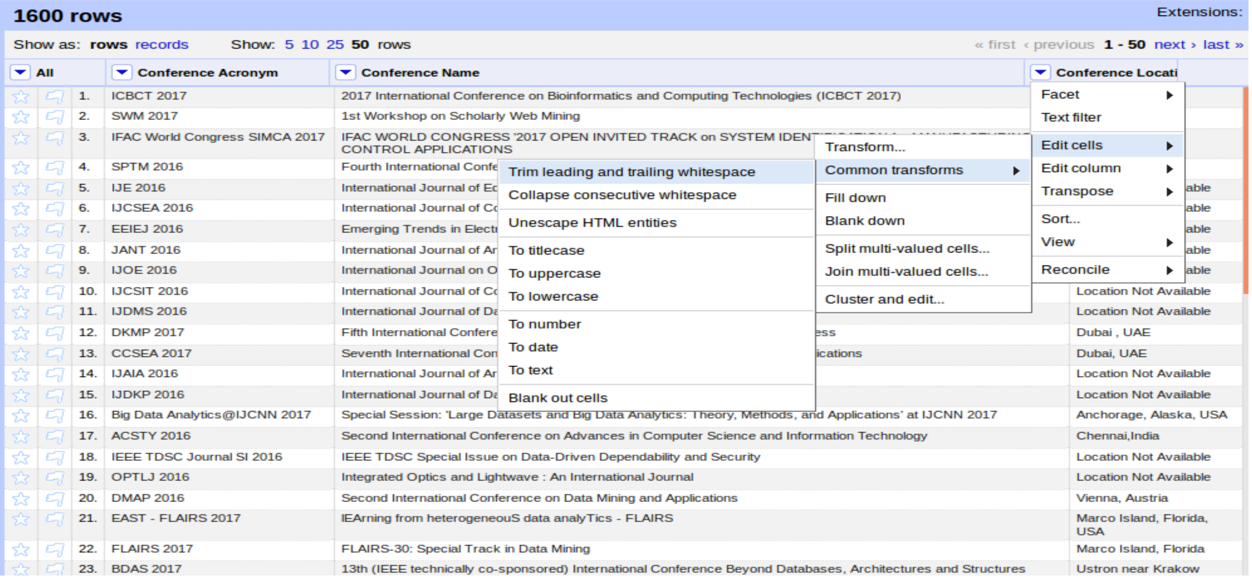


All	Conference Acronym	Conference Name	Conference Locati
1.	ICBCT 2017	2017 International Conference on Bioinformatics and Computing Technologies (ICBCT 2017)	Hong Kong, China
2.	SWM 2017	1st Workshop on Scholarly Web Mining	Cambridge, UK
3.	IFAC World Congress SIMCA 2017	IFAC WORLD CONGRESS '2017 OPEN INVITED TRACK on SYSTEM IDENTIFICATION for MANUFACTURING CONTROL APPLICATIONS	Toulouse, France
4.	SPTM 2016	Fourth International Conference of Security, Privacy and Trust Management	Chennai, India
5.	DKMP 2017	Fifth International Conference on Data Mining & Knowledge Management Process	Dubai, UAE
6.	CCSEA 2017	Seventh International Conference on Computer Science, Engineering and Applications	Dubai, UAE
7.	Big Data Analytics@IJCNN 2017	Special Session: 'Large Datasets and Big Data Analytics: Theory, Methods, and Applications' at IJCNN 2017	Anchorage, Alaska, USA
8.	ACSTY 2016	Second International Conference on Advances in Computer Science and Information Technology	Chennai, India
9.	DMAP 2016	Second International Conference on Data Mining and Applications	Vienna, Austria
10.	EAST - FLAIRS 2017	lEArning from heterogeneouS data analyTics - FLAIRS	Marco Island, Florida, USA
11.	FLAIRS 2017	FLAIRS-30: Special Track in Data Mining	Marco Island, Florida, USA
12.	Sideways 2017	Social Media World Sensors 2017	Madeira
13.	ICISDM 2017	2017 International Conference on Information System and Data Mining (ICISDM 2017)-SCOPUS, Ei Compendex	South Carolina, USA
14.	ITMLS 2017	Intelligent Technologies and Methodologies of Learning Systems	Barcelona, Spain
15.	ITPF 2016	The 5th international conference on Information Technology, Present and Future	Mashhad, Iran
16.	ICCSP 2017	International Conference on Cryptography, Security and Privacy - Ei Compendex and Scopus	Wuhan, China
17.	ICKSE 2017	3rd International Conference on Knowledge and Software Engineering-Ei Compendex, Scopus & ISI CPCS	Paris, France
18.	WSDM Cup 2017	WSDM Cup 2017: Knowledge Base Quality and Search	Cambridge, MA, USA
19.	KDD 2017	Knowledge Discovery and Data Mining	Halifax, Nova Scotia, Canada
20.	ICOK 2017	2017 2nd International Conference on Knowledge (ICOK 2017)	Chengdu, Sichuan, China
21.	ICBDA 2017	The 2017 IEEE International Conference on Big Data Analysis (ICBDA 2017) - Ei Compendex	Beijing, China
22.	IEA/AIE NAAD 2017	Special Track on Novel Approaches to Anomaly Detection	Arras, France

Figure 2: Data upload in OpenRefine

In order to clean the following steps are followed.

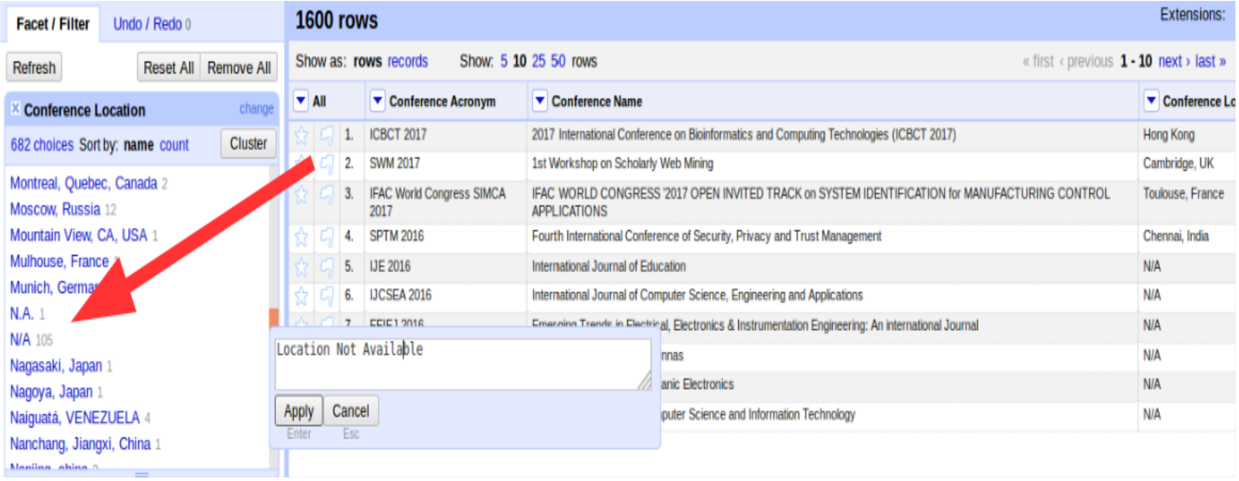
- **Trim white space**: At first, from all the columns the leading and trailing white spaces are removed. For this **Edit cells→Common transforms→Trim leading and trailing spaces**, steps are followed as shown in Figure 3.
- **Cleaning using Conference Location**: In order to clean the data, at first the the **Conference Location** column is selected. There were a number of inconsistencies in this column. In order to maintain consistency and make all the conference location to follow a specific format without loosing important information, **City name**, **Country** format is followed in general for conference location. While cleaning the data, the following scenarios are considered,



All	Conference Acronym	Conference Name	Conference Locati
1.	ICBCT 2017	2017 International Conference on Bioinformatics and Computing Technologies (ICBCT 2017)	
2.	SWM 2017	1st Workshop on Scholarly Web Mining	
3.	IFAC World Congress SIMCA 2017	IFAC WORLD CONGRESS '2017 OPEN INVITED TRACK on SYSTEM IDEN CONTROL APPLICATIONS	
4.	SPTM 2016	Fourth International Confe	
5.	IJE 2016	International Journal of Ec	able
6.	IJCSEA 2016	International Journal of Cc	able
7.	EEIEJ 2016	Emerging Trends in Electr	able
8.	JANT 2016	International Journal of Ar	able
9.	IJOE 2016	International Journal on O	able
10.	IJCSIT 2016	International Journal of Cc	Location Not Available
11.	IJDMS 2016	International Journal of Da	Location Not Available
12.	DKMP 2017	Fifth International Confere ess	Dubai , UAE
13.	CCSEA 2017	Seventh International Con ications	Dubai, UAE
14.	IJAIA 2016	International Journal of Ar	Location Not Available
15.	IJDKP 2016	International Journal of Da	Location Not Available
16.	Big Data Analytics@IJCNN 2017	Special Session: 'Large Datasets and Big Data Analytics: Theory, Methods, and Applications' at IJCNN 2017	Anchorage, Alaska, USA
17.	ACSTY 2016	Second International Conference on Advances in Computer Science and Information Technology	Chennai,India
18.	IEEE TDSC Journal SI 2016	IEEE TDSC Special Issue on Data-Driven Dependability and Security	Location Not Available
19.	OPTLJ 2016	Integrated Optics and Lightwave : An International Journal	Location Not Available
20.	DMAP 2016	Second International Conference on Data Mining and Applications	Vienna, Austria
21.	EAST - FLAIRS 2017	lEArning from heterogeneouS data analyTics - FLAIRS	Marco Island, Florida, USA
22.	FLAIRS 2017	FLAIRS-30: Special Track in Data Mining	Marco Island, Florida
23.	BDAS 2017	13th (IEEE technically co-sponsored) International Conference Beyond Databases, Architectures and Structures	Ustron near Krakow

Figure 3: Trimming white spaces from columns

1. **N/A**: There are some conference entries in which the location is **N/A** or **N.A.** according to Figure 4. At first, these locations are considered as **Location Not Available**.

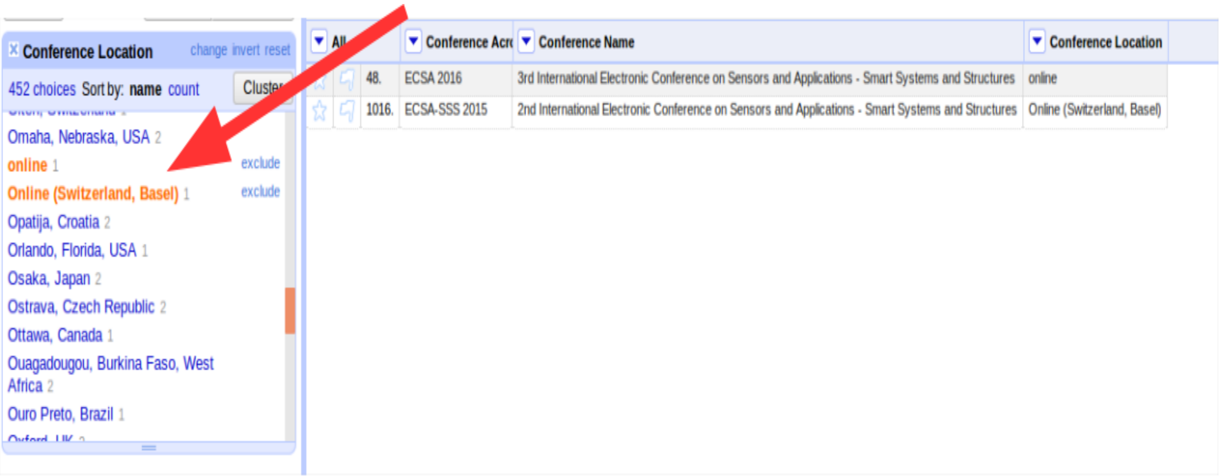


All	Conference Acronym	Conference Name	Conference Lo
1.	ICBCT 2017	2017 International Conference on Bioinformatics and Computing Technologies (ICBCT 2017)	Hong Kong
2.	SWM 2017	1st Workshop on Scholarly Web Mining	Cambridge, UK
3.	IFAC World Congress SIMCA 2017	IFAC WORLD CONGRESS '2017 OPEN INVITED TRACK on SYSTEM IDENTIFICATION for MANUFACTURING CONTROL APPLICATIONS	Toulouse, France
4.	SPTM 2016	Fourth International Conference of Security, Privacy and Trust Management	Chennai, India
5.	IJE 2016	International Journal of Education	N/A
6.	IJCSEA 2016	International Journal of Computer Science, Engineering and Applications	N/A
7.	EEIEJ 2016	Emerging Trends in Electrical, Electronics & Instrumentation Engineering: An international Journal	N/A
		nnas	N/A
		anic Electronics	N/A
		puter Science and Information Technology	N/A

Figure 4: Removal of Location N/A to Location Not Available

2. **Online**: There are also some conference entries with location **Online** or **online** as Figure 5. Initially, we have also converted them to **Location Not Available**.
3. **Journal Name**: There are some conference entries which are actually call for journal article. In this case, the journal name is included as location. For example, in Figure 6, the location is shown as **Studies in Computational Intelligence**. These entries are also replaced by **Location Not Available**.

After observing the above 3 situations, there were total 110 entries with **Location Not**



Conference Location change invert reset

452 choices Sort by: name count Cluster

Omaha, Nebraska, USA 2

online 1 exclude

Online (Switzerland, Basel) 1 exclude

Opatija, Croatia 2

Orlando, Florida, USA 1

Osaka, Japan 2

Ostrava, Czech Republic 2

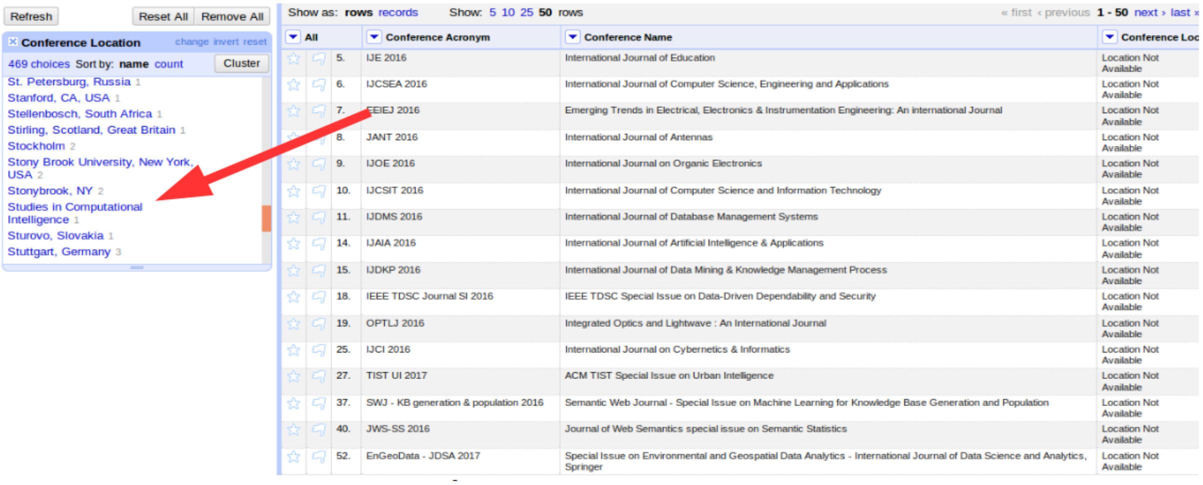
Ottawa, Canada 1

Ouagadougou, Burkina Faso, West Africa 2

Ouro Preto, Brazil 1

All	Conference Acr	Conference Name	Conference Location
48.	ECSA 2016	3rd International Electronic Conference on Sensors and Applications - Smart Systems and Structures	online
1016.	ECSA-SSS 2015	2nd International Electronic Conference on Sensors and Applications - Smart Systems and Structures	Online (Switzerland, Basel)

Figure 5: Conferences with Location Online



Refresh Reset All Remove All

Conference Location change invert reset

469 choices Sort by: name count Cluster

St. Petersburg, Russia 1

Stanford, CA, USA 1

Stellenbosch, South Africa 1

Stirling, Scotland, Great Britain 1

Stockholm 2

Stony Brook University, New York, USA 2

Stonybrook, NY 2

Studies in Computational Intelligence 1

Sturovo, Slovakia 1

Stuttgart, Germany 3

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

All	Conference Acronym	Conference Name	Conference Loc
5.	IJE 2016	International Journal of Education	Location Not Available
6.	IJCSEA 2016	International Journal of Computer Science, Engineering and Applications	Location Not Available
7.	EEIEJ 2016	Emerging Trends in Electrical, Electronics & Instrumentation Engineering: An international Journal	Location Not Available
8.	JANT 2016	International Journal of Antennas	Location Not Available
9.	IJOE 2016	International Journal on Organic Electronics	Location Not Available
10.	IJCSIT 2016	International Journal of Computer Science and Information Technology	Location Not Available
11.	IJDMS 2016	International Journal of Database Management Systems	Location Not Available
14.	IJAIA 2016	International Journal of Artificial Intelligence & Applications	Location Not Available
15.	IJDKP 2016	International Journal of Data Mining & Knowledge Management Process	Location Not Available
18.	IEEE TDSC Journal SI 2016	IEEE TDSC Special Issue on Data-Driven Dependability and Security	Location Not Available
19.	OPTLJ 2016	Integrated Optics and Lightwave : An International Journal	Location Not Available
25.	IJCI 2016	International Journal on Cybernetics & Informatics	Location Not Available
27.	TIST UI 2017	ACM TIST Special Issue on Urban Intelligence	Location Not Available
37.	SWJ - KB generation & population 2016	Semantic Web Journal - Special Issue on Machine Learning for Knowledge Base Generation and Population	Location Not Available
40.	JWS-SS 2016	Journal of Web Semantics special issue on Semantic Statistics	Location Not Available
52.	EnGeoData - JDSA 2017	Special Issue on Environmental and Geospatial Data Analytics - International Journal of Data Science and Analytics, Springer	Location Not Available

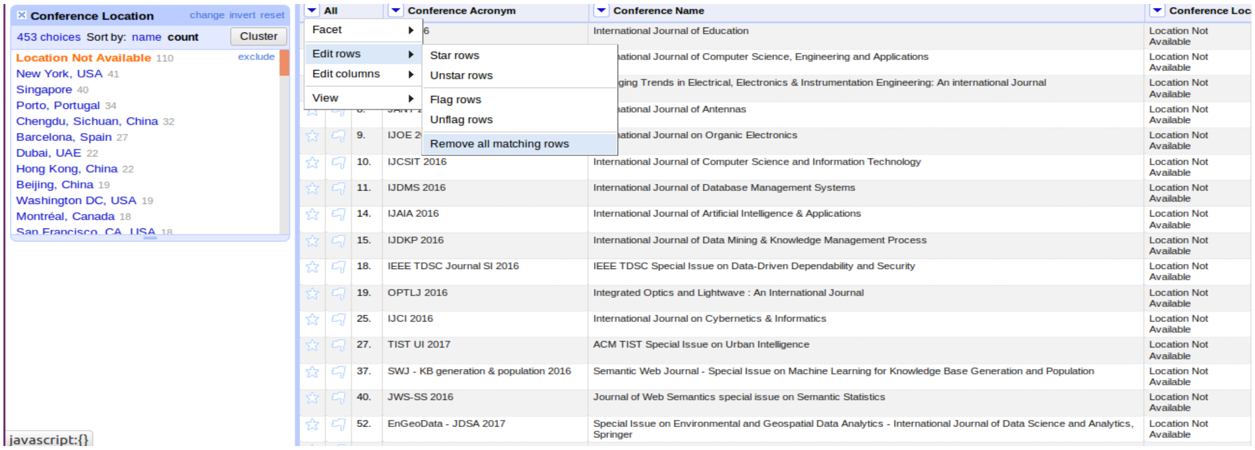
Figure 6: Conferences with Journal name as location

Available. Since the location information is important, these entries are removed. The removal operation is shown in Figure 7.

After this operation, the clusters among the conference location data is determined by the Cluster option of OpenRefine. With proper observation, the entries were merged. The screen shots of the cluster options are given on Figure 8.

Then, there are some conference entries for which the location format was still inconsistent. For example, for some conference the location is **San Fransisco, CA, USA** and for some conference the location is **San Fransisco, California, USA**. Therefore, such entries are found out and converted to a single format. For example, for this case I have converted it to **San Fransisco, CA, USA**. In order to find such entries, from the **Conference Location column →Facet →Text Facet**. From the facet, the repetitions are found according to Figure 9.

After cleaning the data, using conference location column, I found that, there are some confer-



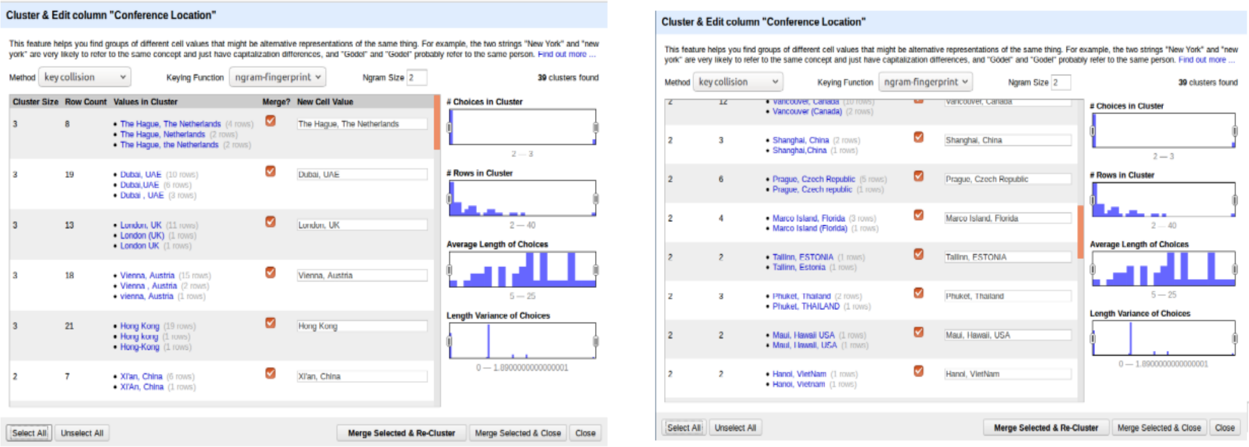
Conference Location change invert reset

453 choices Sort by: name count Cluster

Location Not Available	110
New York, USA	41
Singapore	40
Porto, Portugal	34
Chengdu, Sichuan, China	32
Barcelona, Spain	27
Dubai, UAE	22
Hong Kong, China	22
Beijing, China	19
Washington DC, USA	19
Montréal, Canada	18
San Francisco, CA, USA	18

All		Conference Acronym	Conference Name	Conference Loc
		6	International Journal of Education	Location Not Available
			ational Journal of Computer Science, Engineering and Applications	Location Not Available
			ging Trends in Electrical, Electronics & Instrumentation Engineering: An international Journal	Location Not Available
			ational Journal of Antennas	Location Not Available
	9.	IJOE 2	ational Journal on Organic Electronics	Location Not Available
	10.	IJCSIT 2016	International Journal of Computer Science and Information Technology	Location Not Available
	11.	IJDMS 2016	International Journal of Database Management Systems	Location Not Available
	14.	IJAIA 2016	International Journal of Artificial Intelligence & Applications	Location Not Available
	15.	IJDKP 2016	International Journal of Data Mining & Knowledge Management Process	Location Not Available
	18.	IEEE TDSC Journal SI 2016	IEEE TDSC Special Issue on Data-Driven Dependability and Security	Location Not Available
	19.	OPTLJ 2016	Integrated Optics and Lightwave : An International Journal	Location Not Available
	25.	IJCI 2016	International Journal on Cybernetics & Informatics	Location Not Available
	27.	TIST UI 2017	ACM TIST Special Issue on Urban Intelligence	Location Not Available
	37.	SWJ - KB generation & population 2016	Semantic Web Journal - Special Issue on Machine Learning for Knowledge Base Generation and Population	Location Not Available
	40.	JWS-SS 2016	Journal of Web Semantics special issue on Semantic Statistics	Location Not Available
	52.	EnGeoData - JDSA 2017	Special Issue on Environmental and Geospatial Data Analytics - International Journal of Data Science and Analytics, Springer	Location Not Available

Figure 7: Removal of entries with **Location Not Available**



Cluster & Edit column "Conference Location"

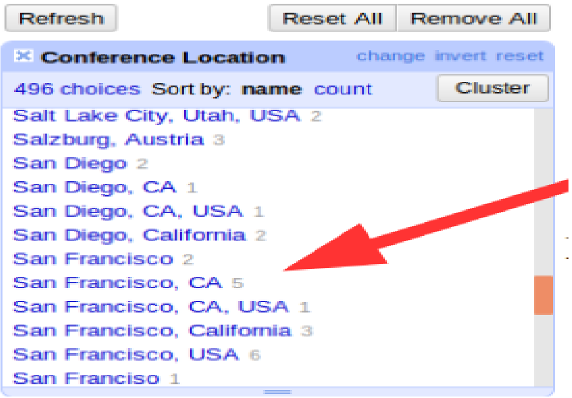
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method: key collision Keying Function: ngram-fingerprint Ngram Size: 2 39 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	8	The Hague, The Netherlands (4 rows) The Hague, Netherlands (2 rows) The Hague, the Netherlands (2 rows)	☑	The Hague, The Netherlands
3	19	Dubai, UAE (10 rows) Dubai,UAE (6 rows) Dubai , UAE (3 rows)	☑	Dubai, UAE
3	13	London, UK (11 rows) London (UK) (1 rows) London UK (1 rows)	☑	London, UK
3	18	Vienna, Austria (15 rows) Vienna , Austria (2 rows) vienna, Austria (1 rows)	☑	Vienna, Austria
3	21	Hong Kong (19 rows) Hong kong (1 rows) Hong-Kong (1 rows)	☑	Hong Kong
2	7	Xi'an, China (6 rows) Xi'An, China (1 rows)	☑	Xi'an, China

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	12	Vancouver, Canada (10 rows) Vancouver (Canada) (2 rows)	☑	Vancouver, Canada
2	3	Shanghai, China (2 rows) Shanghai,China (1 rows)	☑	Shanghai, China
2	6	Prague, Czech Republic (5 rows) Prague, Czech republic (1 rows)	☑	Prague, Czech Republic
2	4	Marco Island, Florida (3 rows) Marco Island (Florida) (1 rows)	☑	Marco Island, Florida
2	2	Tallinn, ESTONIA (1 rows) Tallinn, Estonia (1 rows)	☑	Tallinn, ESTONIA
2	3	Phuket, Thailand (2 rows) Phuket, THAILAND (1 rows)	☑	Phuket, Thailand
2	2	Maui, Hawaii USA (1 rows) Maui, Hawaii, USA (1 rows)	☑	Maui, Hawaii, USA
2	2	Hanoi, VietNam (1 rows) Hanoi, Vietnam (1 rows)	☑	Hanoi, VietNam

Figure 8: Cluster Options



Conference Location change invert reset

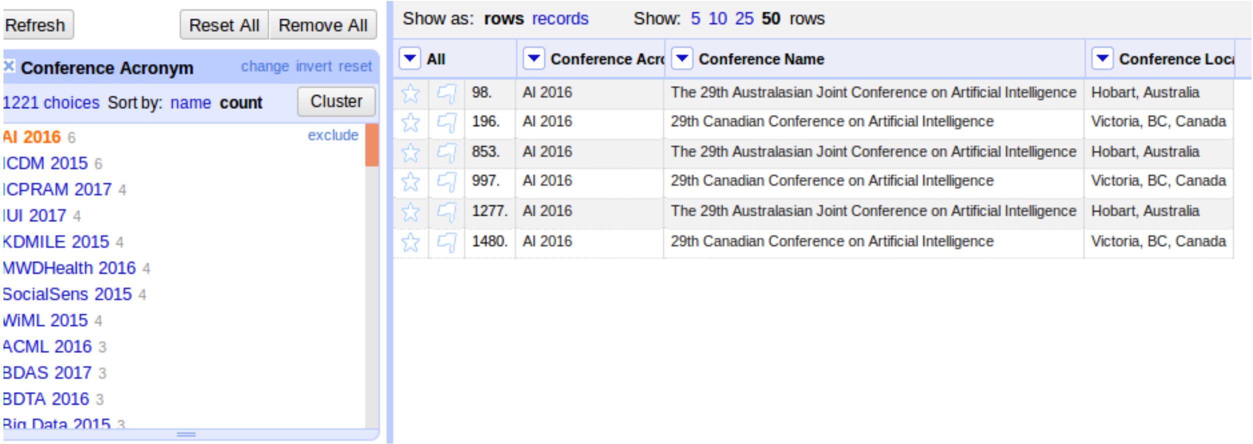
496 choices Sort by: name count Cluster

Salt Lake City, Utah, USA	2
Salzburg, Austria	3
San Diego	2
San Diego, CA	1
San Diego, CA, USA	1
San Diego, California	2
San Francisco	2
San Francisco, CA	5
San Francisco, CA, USA	1
San Francisco, California	3
San Francisco, USA	6
San Franciso	1

Figure 9: Inconsistency in the name of locations

ence entries that are repeated multiple times. The reason behind this is, there are some conference which are included as Data mining and Artificial Intelligence categories or more. Therefore, in order to get a clear view of the data, I removed the repeated entries.

In order to accomplish this, at first I tried to merge duplicate rows according to same Conference Acronym. But in that case, some other conference entries might be removed. For example, Figure 10 is showing that conferences with Acronym **AI 2016** has 6 entries but only 2 different conferences. In that case, I removed the repeated 4 conferences using star. The operation is, **All →Facet →Facet by star**. From **Starred Rows** select **true**. Then only the starred rows will appear. After that **All →Edit rows →Remove all matching rows**.

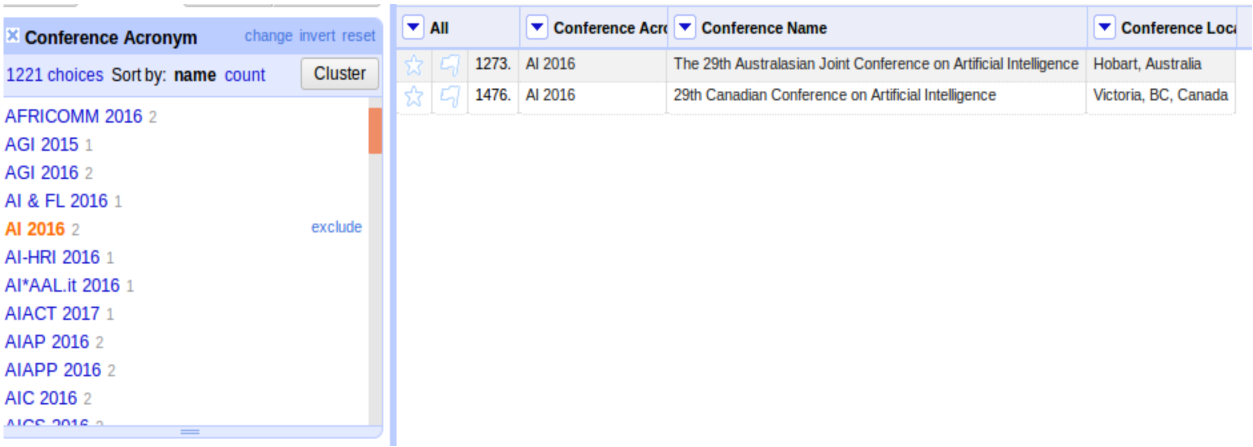


Conference Acronym	Count
AI 2016	6
ICDM 2015	6
ICPRAM 2017	4
IUI 2017	4
KDMILE 2015	4
MWDHealth 2016	4
SocialSens 2015	4
WiML 2015	4
ACML 2016	3
BDAS 2017	3
BDTA 2016	3
Big Data 2015	3

All	Conference Acro	Conference Name	Conference Loca
98.	AI 2016	The 29th Australasian Joint Conference on Artificial Intelligence	Hobart, Australia
196.	AI 2016	29th Canadian Conference on Artificial Intelligence	Victoria, BC, Canada
853.	AI 2016	The 29th Australasian Joint Conference on Artificial Intelligence	Hobart, Australia
997.	AI 2016	29th Canadian Conference on Artificial Intelligence	Victoria, BC, Canada
1277.	AI 2016	The 29th Australasian Joint Conference on Artificial Intelligence	Hobart, Australia
1480.	AI 2016	29th Canadian Conference on Artificial Intelligence	Victoria, BC, Canada

Figure 10: Problem of removing repeated entries using conference acronym

After removing the repeated entries it becomes like Figure 11.

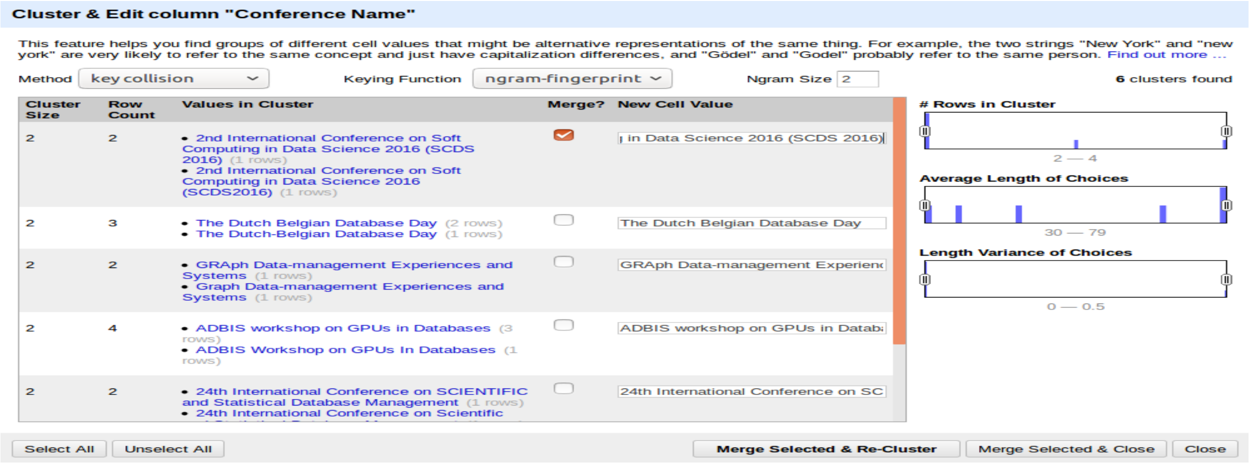


Conference Acronym	Count
AFRICOMM 2016	2
AGI 2015	1
AGI 2016	2
AI & FL 2016	1
AI 2016	2
AI-HRI 2016	1
AI*AAL.it 2016	1
AIACT 2017	1
AIAP 2016	2
AIAPP 2016	2
AIC 2016	2

All	Conference Acro	Conference Name	Conference Loca
1273.	AI 2016	The 29th Australasian Joint Conference on Artificial Intelligence	Hobart, Australia
1476.	AI 2016	29th Canadian Conference on Artificial Intelligence	Victoria, BC, Canada

Figure 11: Scenario after removing repeated entries

The problem of conference acronym also occurs with conference name. Therefore, after observing the clusters manually, I clustered the same conferences together as shown in Figure 12.



Cluster & Edit column "Conference Name"

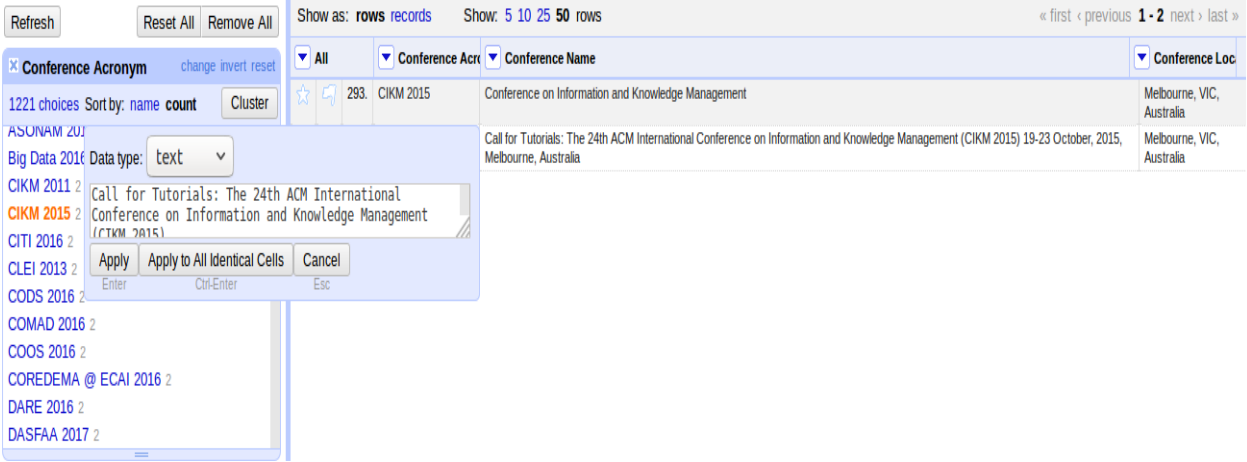
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method: key collision Keying Function: ngram-fingerprint Ngram Size: 2 **6** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	• 2nd International Conference on Soft Computing in Data Science 2016 (SCDS 2016) (1 rows) • 2nd International Conference on Soft Computing in Data Science 2016 (SCDS2016) (1 rows)	☑	in Data Science 2016 (SCDS 2016)
2	3	• The Dutch Belgian Database Day (2 rows) • The Dutch-Belgian Database Day (1 rows)	☐	The Dutch Belgian Database Day
2	2	• GRAph Data-management Experiences and Systems (1 rows) • Graph Data-management Experiences and Systems (1 rows)	☐	GRAph Data-management Experienc
2	4	• ADBIS workshop on GPUs in Databases (3 rows) • ADBIS Workshop on GPUs In Databases (1 rows)	☐	ADBIS workshop on GPUs in Databa
2	2	• 24th International Conference on SCIENTIFIC and Statistical Database Management (1 rows) • 24th International Conference on Scientific	☐	24th International Conference on SC

Figure 12: Clustering according to conference name

There are some conference names in which the time and location of the conference are appended, I removed them as shown in Figure 13.



1221 choices Sort by: name count

ASONAM 201
Big Data 2016
CIKM 2011 2
CIKM 2015 2
CITI 2016 2
CLEI 2013 2
CODS 2016 2
COMAD 2016 2
COOS 2016 2
COREDEMA @ ECAI 2016 2
DARE 2016 2
DASFAA 2017 2

Data type: text

```
Call for Tutorials: The 24th ACM International
Conference on Information and Knowledge Management
(CIKM 2015)
```

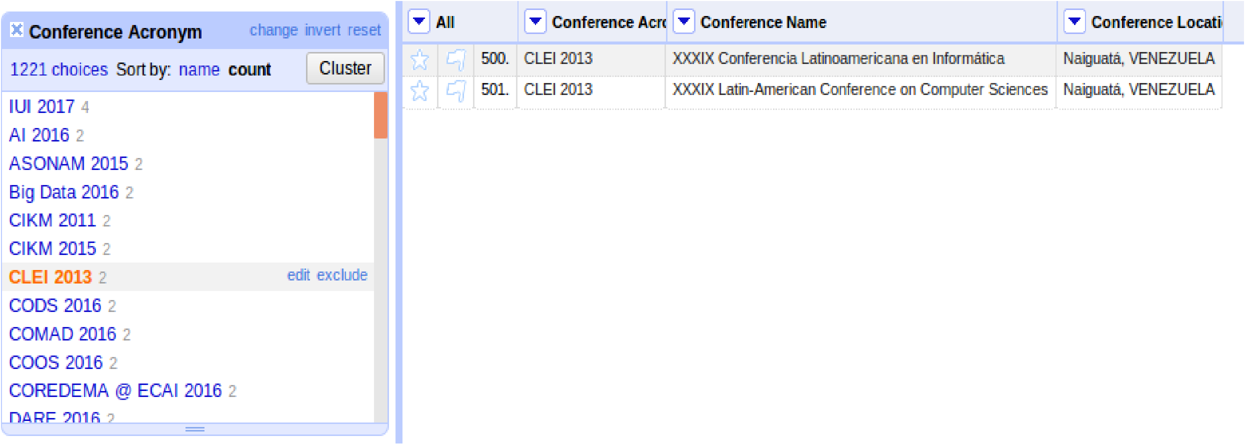
Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 2 next › last »

All	Conference Acro	Conference Name	Conference Loc
293.	CIKM 2015	Conference on Information and Knowledge Management	Melbourne, VIC, Australia
		Call for Tutorials: The 24th ACM International Conference on Information and Knowledge Management (CIKM 2015) 19-23 October, 2015, Melbourne, Australia	Melbourne, VIC, Australia

Figure 13: Correcting the conference name

Though, a thorough cleaning operation is performed on the Conference Location column, there are still some inconsistencies in the data. For example, same conference is entered once with English name. Then again with Spanish name as shown in Figure 14.

After removing repeated data and all inconsistencies now I have **1238** conference entries from initially collected **1600** conference entries. The number of location is reduced from **682** to **442**. In order to perform the data cleaning operation efficiently, I followed the videos of OpenRefine[3]. From this portion, the project is exported as **WikiCFP_CleanData.tsv** file.



Conference Acronym change invert reset

1221 choices Sort by: name count Cluster

IUI 2017 4
AI 2016 2
ASONAM 2015 2
Big Data 2016 2
CIKM 2011 2
CIKM 2015 2
CLEI 2013 2 edit exclude
CODS 2016 2
COMAD 2016 2
COOS 2016 2
COREDEMA @ ECAI 2016 2
DARE 2016 2

All	Conference Acro	Conference Name	Conference Locati
500.	CLEI 2013	XXXIX Conferencia Latinoamericana en Informática	Naiguatá, VENEZUELA
501.	CLEI 2013	XXXIX Latin-American Conference on Computer Sciences	Naiguatá, VENEZUELA

Figure 14: Repeated conference with the name written in different languages

3. Hadoop

In order to work with hadoop, at first I have followed all the steps of **Apache HADOOP**[4][5] tutorials. After running the **WordCount.java**, I have started working on the problems. For all the problems, in mapper portion, **StringTokenizer** is used to tokenize the string. Here as delimiter new line character is used, since in the **WikiCFP_CleanData.txt** file all the entires of conferences are in each new line. Then each new line is split using tab character as a delimiter. In each of the mapper function, I have checked, if current line's location is “Conference Location”or not. In order to prevent the column heading to be mistakenly considered by the program, I have deleted such entries.

- ***Compute and plot the number of conferences per city. Which are the top 10 locations?***
  In this problem, the **Conference Location** is considered as **key**. For each location, **1** is stored as **value**. Here key is a Text element and value is a IntWritable element. In the reducer part, the values are calculated which is actually the frequency of locations and written with the location name. After the proper execution of the code, the output file is generated. The file is copied from the distributed file system to the local system. The data of this file is used to plot the graph through excel. The graph for the number of conferences per city is shown in Figure 15.

  The top 10 locations in which the highest number of conferences are taking place are shown in Figure 16. Figure 17 is showing the result of the hadoop program from terminal.

- ***Output the list of conferences per city?***
  In this problem, the mapper has considered the **Conference Location** as **key** and the **Conference Name** as **value**. Therefore, both key and value are Text in this case. In the reducer, as **Iterable**, Text is considered which is the conference name. In each iteration, the conference names for a particular location are concatenated as a single string. Then the reducer writes the conference location and the conference names of that location (as a single string) in the output file. In order to solve this problem, I have used Text iterable and got

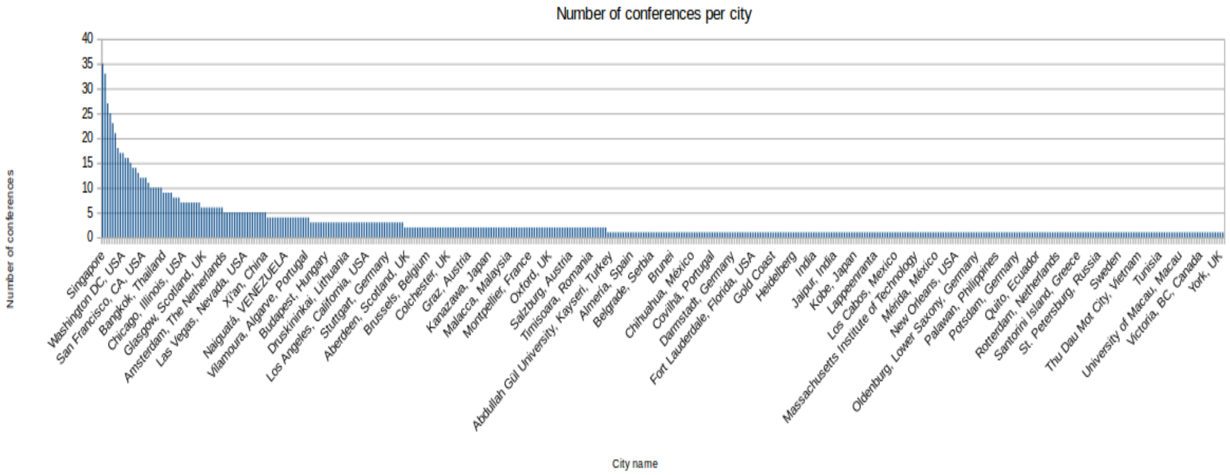
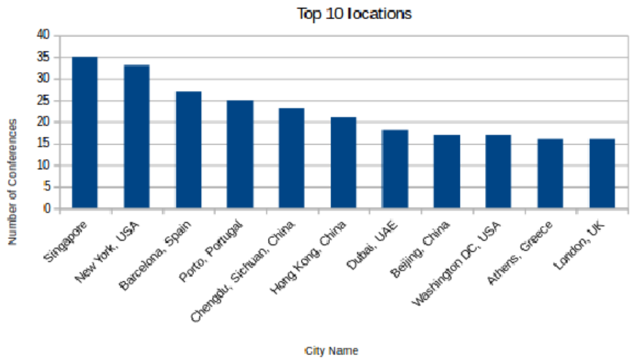
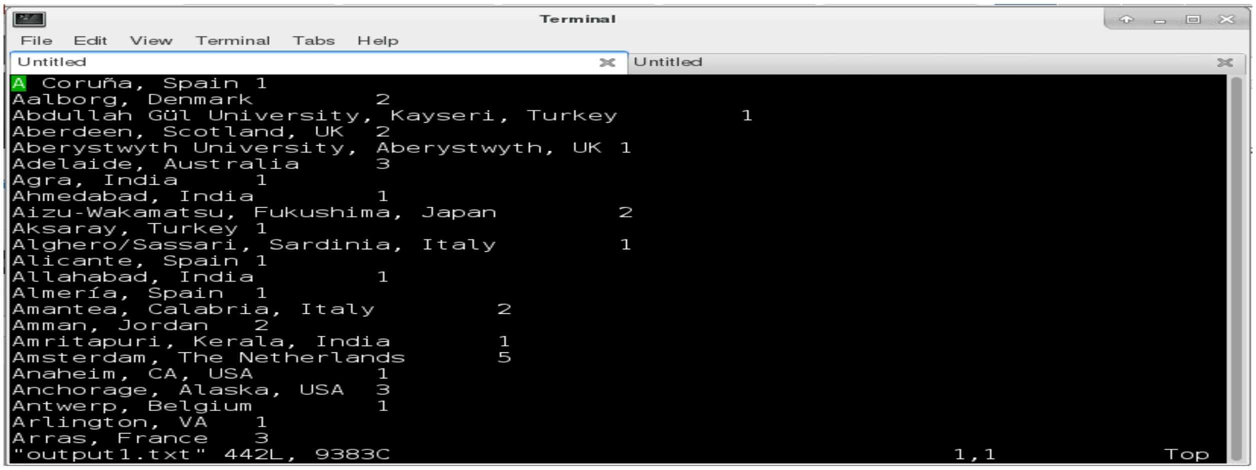


Figure 15: Plot showing number of conferences per city



City Name	Number of Conferences
Singapore	35
New York, USA	33
Barcelona, Spain	27
Porto, Portugal	25
Chengdu, Sichuan, China	23
Hong Kong, China	21
Dubai, UAE	18
Beijing, China	17
Washington DC, USA	17
Athens, Greece	16
London, UK	16

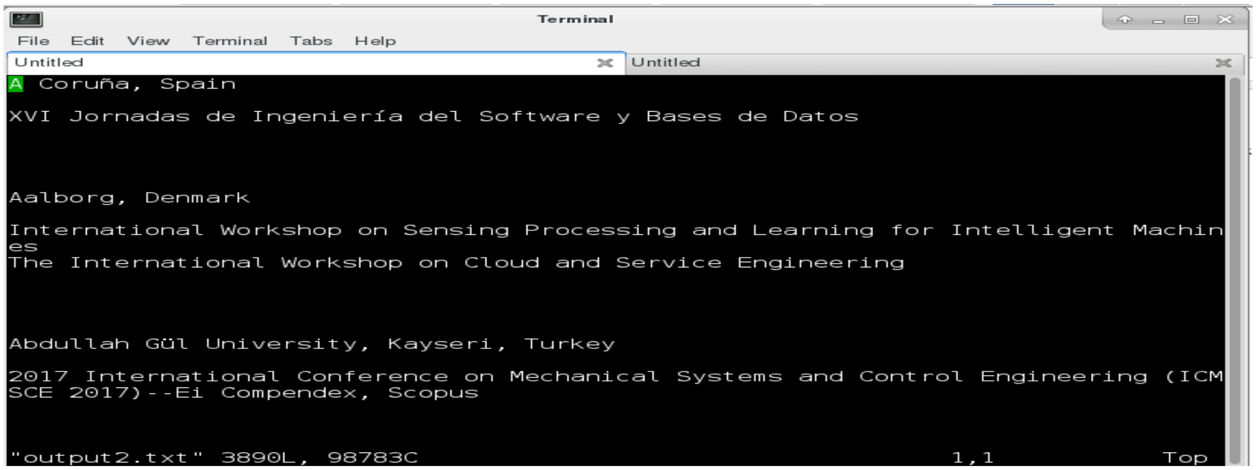
Figure 16: Plot showing top 10 cities with highest number of conferences



```
A Coruña, Spain 1
Aalborg, Denmark        2
Abdullah Gül University, Kayseri, Turkey        1
Aberdeen, Scotland, UK  2
Aberystwyth University, Aberystwyth, UK 1
Adelaide, Australia     3
Agra, India     1
Ahmedabad, India        1
Aizu-Wakamatsu, Fukushima, Japan        2
Aksaray, Turkey 1
Alghero/Sassari, Sardinia, Italy        1
Alicante, Spain 1
Allahabad, India        1
Almería, Spain  1
Amantea, Calabria, Italy        2
Amman, Jordan   2
Amritapuri, Kerala, India       1
Amsterdam, The Netherlands      5
Anaheim, CA, USA        1
Anchorage, Alaska, USA  3
Antwerp, Belgium        1
Arlington, VA   1
Arras, France   3
"output1.txt" 442L, 9383C                                     1,1           Top
```

Figure 17: Result of Problem 1

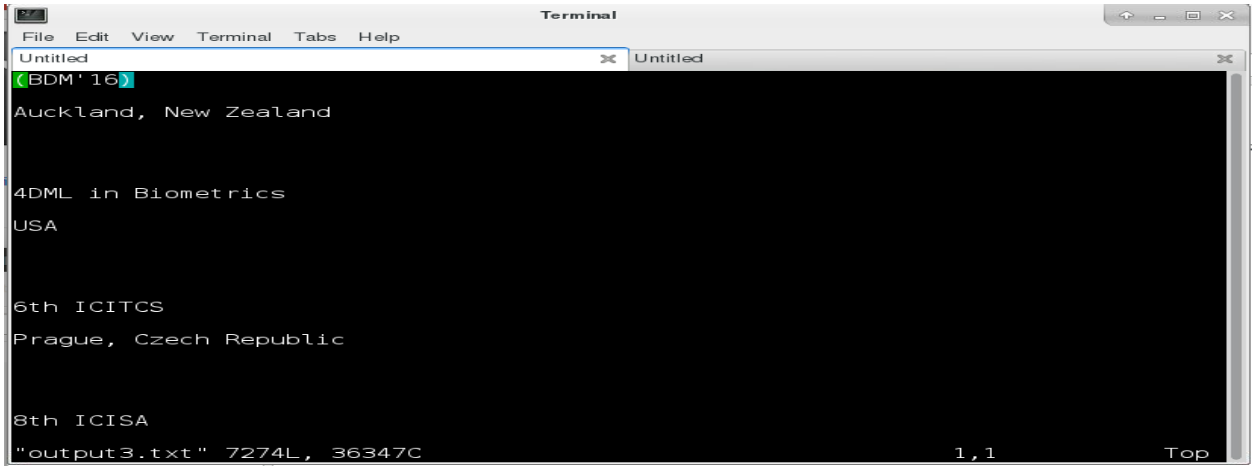
the initial idea about it from CONVERGE Blog[6]. Figure 18 is showing the result generated by the hadoop program in terminal.



```
A Coruña, Spain

XVI Jornadas de Ingeniería del Software y Bases de Datos



Aalborg, Denmark

International Workshop on Sensing Processing and Learning for Intelligent Machin
es
The International Workshop on Cloud and Service Engineering



Abdullah Gül University, Kayseri, Turkey

2017 International Conference on Mechanical Systems and Control Engineering (ICM
SCE 2017)--Ei Compendex, Scopus


"output2.txt" 3890L, 98783C                                   1,1           Top
```

Figure 18: Result of Problem 2

- ***For each conference regardless of the year (e.g., KDD), output the list of cities.*** In this problem, at first the conference acronym is separated from the year. Then only **Conference Acronym** is emitted as **key** and the **Conference Location** is emitted as **value** by the mapper. In the reducer part, the conference location is again considered as text iterable. For a particular conference/conference acronym, the locations are concatenated in one string which is then emitted with the conference acronym as output. Figure 19 is showing the result generated by the hadoop program in terminal.

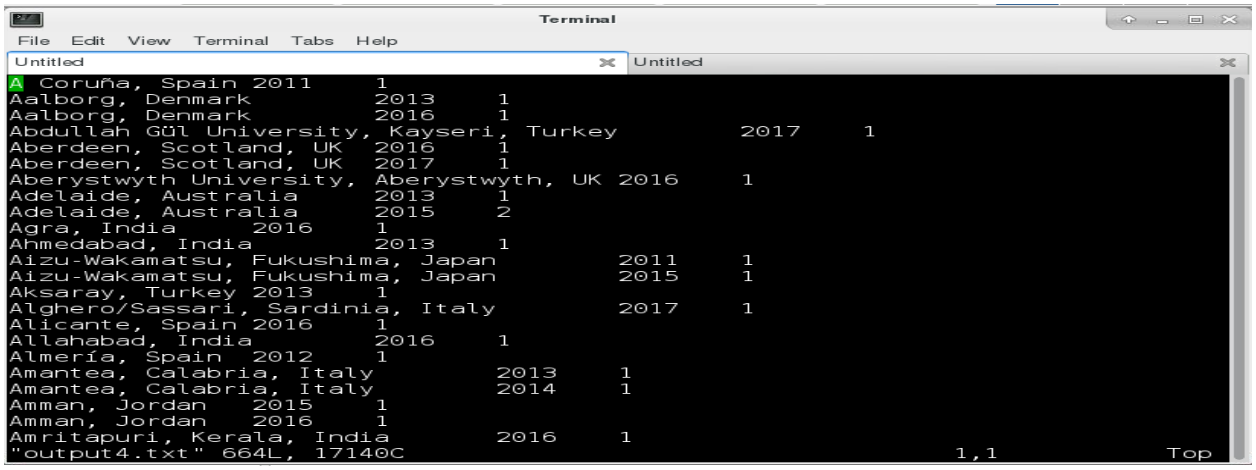


```
(BDM'16)

Auckland, New Zealand



4DML in Biometrics

USA



6th ICITCS

Prague, Czech Republic



8th ICISA
"output3.txt" 7274L, 36347C                                   1,1           Top
```

Figure 19: Result of Problem 3

- ***For each city compute and plot a time series of number of conferences per year.*** In order to solve this part of the problem, in the mapper part, the Year information is collected from the conference acronym. While separating the year information from the acronym, I considered the last 4 characters of it. However, I find that there are some discrepancies.

For example, for conference acronym like “ACTION15”, considering the last 4 digits will give “ON15”. In order to remove this problem, I have checked the first 2 letters. If they are not numbers, then I replaced them by “20”, assuming that the record must be showing information for conference that is going to take place by year 2000 – 2099. Then this Year information is concatenated with conference location information. This merged information of **Conference location and Year** is emitted as **key**. Then for each such key a value of **1** is emitted as **value** from the mapper. Then in the reducer part, for each conference location and year key, the frequency is calculated and emitted as output. Here as iterable the frequency is considered as **IntWritable**. Figure 20 is showing the result generated by the hadoop program in terminal.



```
A Coruña, Spain 2011    1
Aalborg, Denmark        2013    1
Aalborg, Denmark        2016    1
Abdullah Gül University, Kayseri, Turkey        2017    1
Aberdeen, Scotland, UK  2016    1
Aberdeen, Scotland, UK  2017    1
Aberystwyth University, Aberystwyth, UK 2016    1
Adelaide, Australia     2013    1
Adelaide, Australia     2015    2
Agra, India     2016    1
Ahmedabad, India        2013    1
Aizu-Wakamatsu, Fukushima, Japan        2011    1
Aizu-Wakamatsu, Fukushima, Japan        2015    1
Aksaray, Turkey 2013    1
Alghero/Sassari, Sardinia, Italy        2017    1
Alicante, Spain 2016    1
Allahabad, India        2016    1
Almería, Spain  2012    1
Amantea, Calabria, Italy        2013    1
Amantea, Calabria, Italy        2014    1
Amman, Jordan   2015    1
Amman, Jordan   2016    1
Amritapuri, Kerala, India       2016    1
"output4.txt" 664L, 17140C                                        1,1           Top
```

Figure 20: Result of Problem 4

For this problem, I have plotted the top 11 cities. The plots for the top 11 cities are shown in Figure 21 - 23.

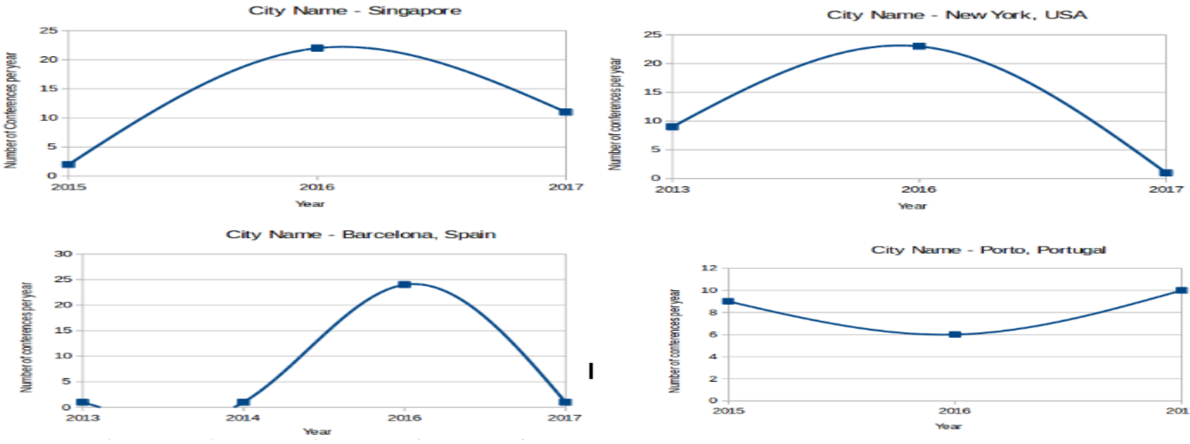


Figure 21: Plot for Singapore with rank 1, New York, USA with rank 2, Barcelona, Spain with rank 3 and Porto, Portugal with rank 4

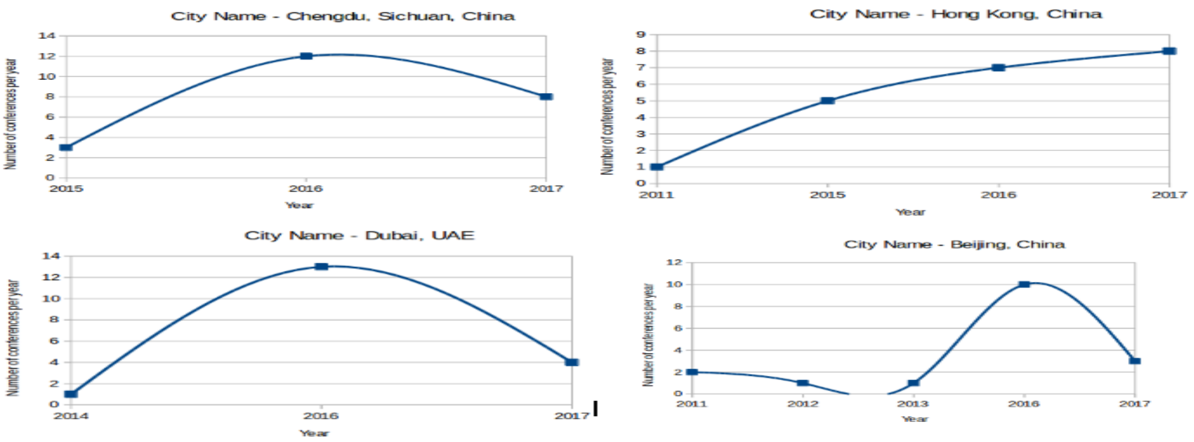


Figure 22: Plot for Chengdu, Sichuan, China with rank 5, Hong Kong, China with rank 6, Dubai, UAE with rank 7 and Beijing, China with rank 8

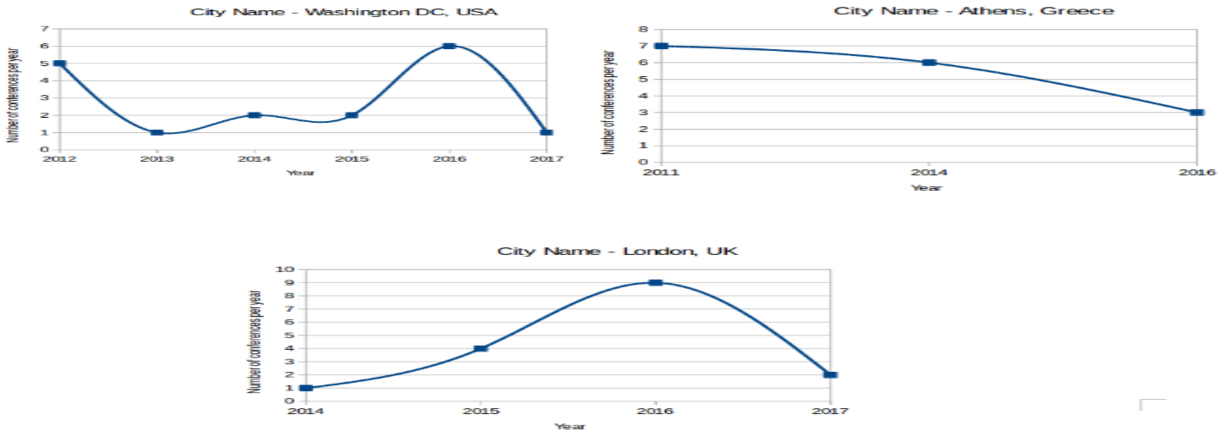


Figure 23: Plot for Washington DC, USA with rank 9, Athens, Greece with rank 10 and London, UK with rank 11

Commands

The solution code for each problem is compiled and run using the commands in Table 1.

Table 1: Commands for the Hadoop program

Task	**Command**
Copy .java file	/extra/netid/hadoop/hadoop-2.7.3/bin
Code compilation	hadoop com.sun.tools.javac.Main Code_File_Name.java
Jar creation	jar cf Code_File_Name.jar Code_File_Name*.class
Run application	hadoop jar Code_File_Name.jar Code_File_Name input output
Copy output file from HDFS	hadoop fs -copyToLocal output/part-r-00000 /extra/netid
Copy output file to local system	scp netid@machineid:/extra/netid/part-r-00000 /home/csgrads/netid

References

[1] StackOverflow, "How to parse HTML table using jsoup?" Last accessed: November 16, 2016. [Online]. Available: http://stackoverflow.com/questions/24772828/how-to-parse-html-table-using-jsoup

[2] Net Instructions, "How to make a simple web crawler in Java," Last accessed: November 16, 2016. [Online]. Available: http://www.netinstructions.com/how-to-make-a-simple-web-crawler-in-java/

[3] Google, "OpenRefine," Last accessed: November 16, 2016. [Online]. Available: http://openrefine.org/

[4] Apache Hadoop - The Apache Software Foundation, "Hadoop: Setting up a Single Node Cluster," Last accessed: November 16, 2016. [Online]. Available: https://hadoop.apache.org/docs/r2.7.0/hadoop-project-dist/hadoop-common/SingleCluster.html

[5] Apache Hadoop, "MapReduce Tutorial," Last accessed: November 16, 2016. [Online]. Available: https://hadoop.apache.org/docs/r2.7.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[6] CONVERGE Blog, "How to Write a MapReduce Program," Last accessed: November 16, 2016. [Online]. Available: https://www.mapr.com/blog/how-write-mapreduce-program