CS235: DATA MINING TECHNIQUES

FALL 2016

ASSIGNMENT REPORT

SUBMITTED BY:
Abhishek Kumar Srivastava
#ID – 861307778

INSTRUCTOR:
Vagelis Papalexakis
University of California Riverside

Part 1: Data Crawling

I have written the crawler in Python. In the crawler for each section i.e Data Mining, Databases, Machine Learning & Artificial Intelligence I have called a general function ***data_from_url(target_url)*** passing appropriate URL for each section.

In this function, page number are added incrementally and using request API the web-page is requested. Then using beautiful soup I have extracted the URL for each conference and then requested again the web-page for that Conference from there I extracted Conference Name and Conference Location. Appropriate time limiter was added of 10 seconds between each page requests.

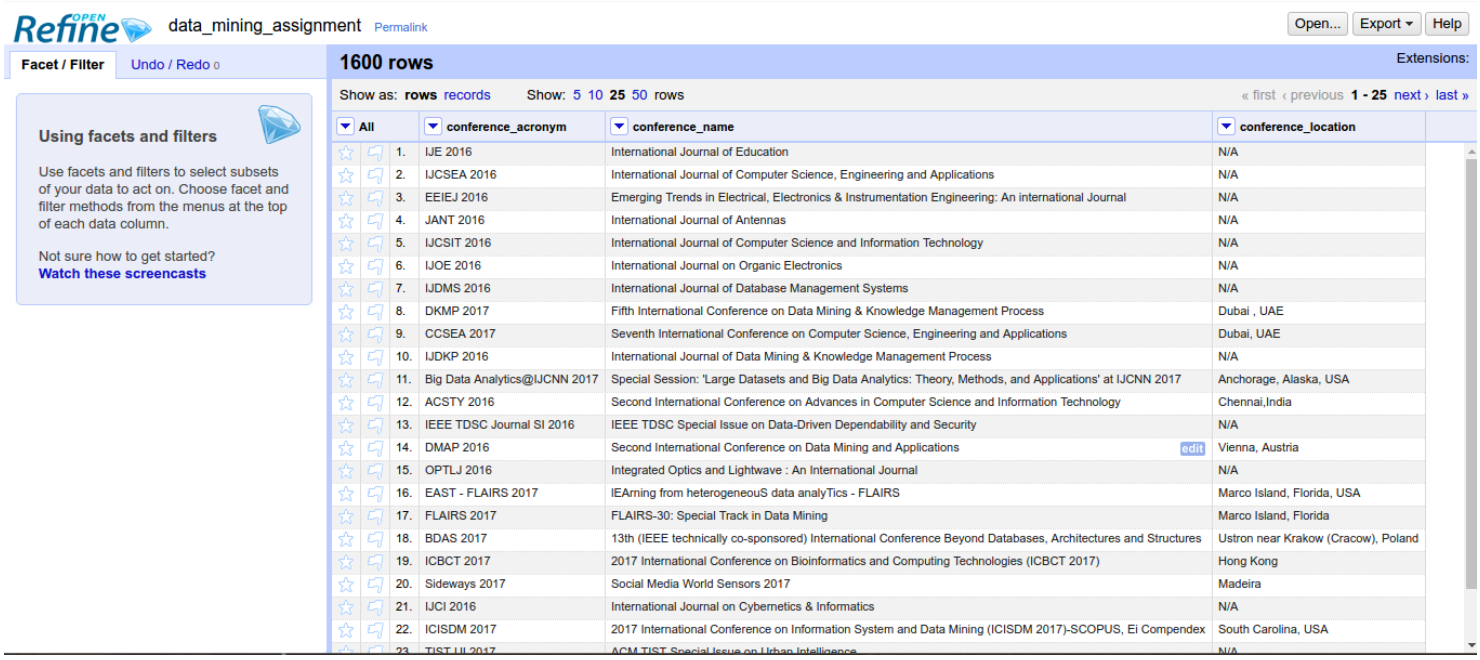
Finally after getting all the required data i.e Conference_Acronym,Conference_Name and Conference_Location for all 20 pages and for all 4 categories I added those values to the file in the tab separated manner in appending manner using CSV Writer.

Please check **crawler.py** for the crawler code and **data_mining_original.tsv** for the tab separate values crawled using the crawler.

Part 2: Data Cleaning

I have used **OpenRefine** for the data cleaning task as recommended by the Instructor.
Following steps were taken in data cleaning with the screen shots for few of them:

1. Importing Data

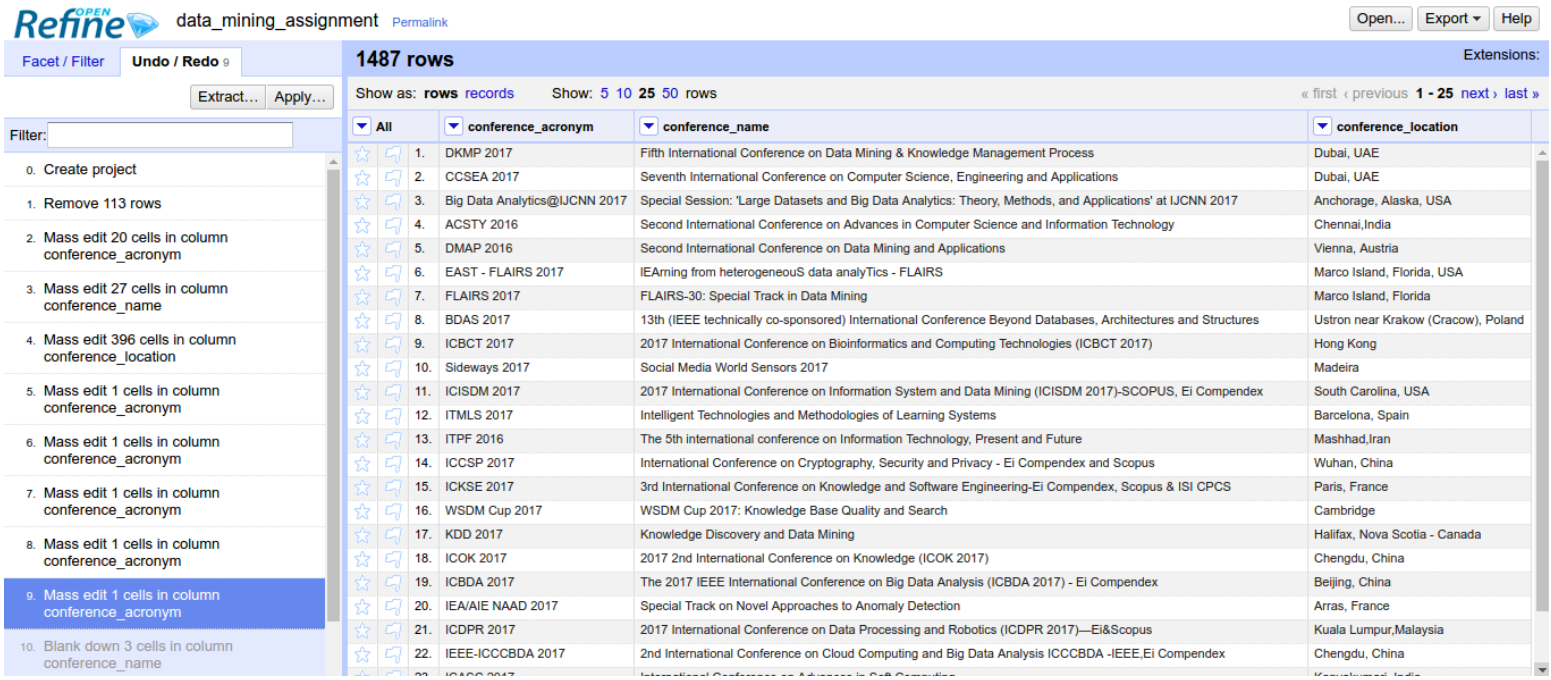


All	conference_acronym	conference_name	conference_location
1.	IJE 2016	International Journal of Education	N/A
2.	IJCSEA 2016	International Journal of Computer Science, Engineering and Applications	N/A
3.	EEIEJ 2016	Emerging Trends in Electrical, Electronics & Instrumentation Engineering: An international Journal	N/A
4.	JANT 2016	International Journal of Antennas	N/A
5.	IJCSIT 2016	International Journal of Computer Science and Information Technology	N/A
6.	IJOE 2016	International Journal on Organic Electronics	N/A
7.	IJDMS 2016	International Journal of Database Management Systems	N/A
8.	DKMP 2017	Fifth International Conference on Data Mining & Knowledge Management Process	Dubai , UAE
9.	CCSEA 2017	Seventh International Conference on Computer Science, Engineering and Applications	Dubai, UAE
10.	IJDKP 2016	International Journal of Data Mining & Knowledge Management Process	N/A
11.	Big Data Analytics@IJCNN 2017	Special Session: 'Large Datasets and Big Data Analytics: Theory, Methods, and Applications' at IJCNN 2017	Anchorage, Alaska, USA
12.	ACSTY 2016	Second International Conference on Advances in Computer Science and Information Technology	Chennai,India
13.	IEEE TDSC Journal SI 2016	IEEE TDSC Special Issue on Data-Driven Dependability and Security	N/A
14.	DMAP 2016	Second International Conference on Data Mining and Applications edit	Vienna, Austria
15.	OPTLJ 2016	Integrated Optics and Lightwave : An International Journal	N/A
16.	EAST - FLAIRS 2017	IEArning from heterogeneouS data analyTics - FLAIRS	Marco Island, Florida, USA
17.	FLAIRS 2017	FLAIRS-30: Special Track in Data Mining	Marco Island, Florida
18.	BDAS 2017	13th (IEEE technically co-sponsored) International Conference Beyond Databases, Architectures and Structures	Ustron near Krakow (Cracow), Poland
19.	ICBCT 2017	2017 International Conference on Bioinformatics and Computing Technologies (ICBCT 2017)	Hong Kong
20.	Sideways 2017	Social Media World Sensors 2017	Madeira
21.	IJCI 2016	International Journal on Cybernetics & Informatics	N/A
22.	ICISDM 2017	2017 International Conference on Information System and Data Mining (ICISDM 2017)-SCOPUS, Ei Compendex	South Carolina, USA
23.	TIST UI 2017	ACM TIST Special Issue on Urban Intelligence	N/A

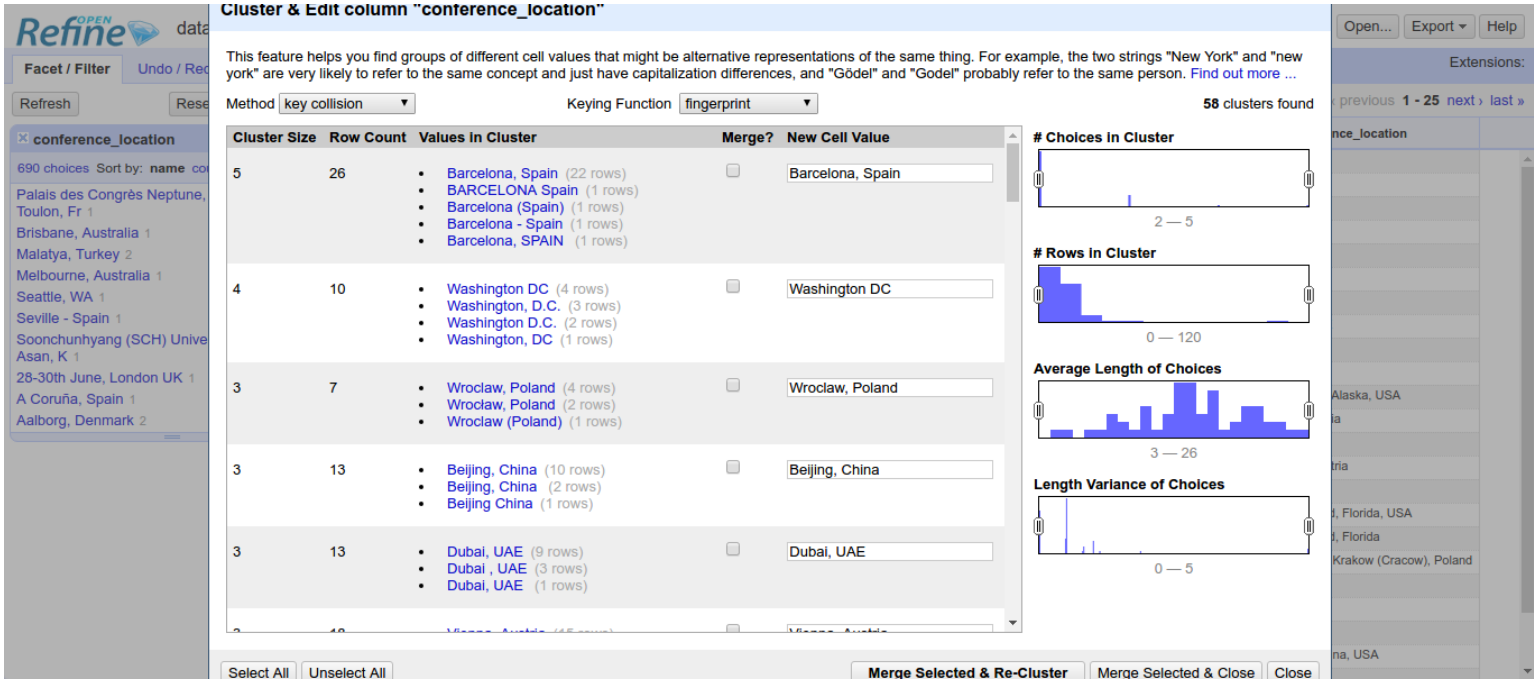
After importing the tsv file directly in the OpenRefine we can see that there are 1600 rows i.e 400 rows for each category (20 pages with 20 conference list each) which validated the correctness of the crawler that it has correctly crawled all the required pages needed to be crawled.

2. Removing whitespaces and conference_location as N/A

First for all 3 Columns trimmed leading and trailing whitespaces and since Conference Location is important requirement for the task to be done further I have removed the data rows in which location is not mentioned. After doing so 1487 rows were left useful.



3. Data Clustering



There were many mismatches for the same name so they were clustered using the clustering feature of OpenRefine. Clustering applied to all Conference name and rows. One example as a snapshot.

4. Anomalous City and Conferences:

Even after clustering there were many mismatched data set were there. They were fixed manually.

Multiple kind of Anomalies are attached:

4 matching rows (1231 total)

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

All		conference_acrc	conference_name	conference_loca
805.		IUI 2017	22nd ACM International Conference on Intelligent User Interfaces	Limassol
798.		IUI 2017	22nd ACM International Conference on Intelligent User Interfaces - Tutorials	Limassol
818.		IUI 2017	22nd ACM International Conference on Intelligent User Interfaces - Workshops	Limassol
786.		IUI 2017	Intelligent User Interfaces	Limassol

1 matching rows (1187 total)

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

All		conference_acrc	conference_name	conference_loca
95.		PAIS'2016 2016	The 2nd International Conference on Pattern Analysis and Intelligent Systems	Khenchela

1 matching rows (1187 total)

Show as: **rows** records Show: 5 10 **25** 50 rows Sort « first ‹ pr

All		conference_acrc	conference_name	conference_loca
343.			Ei&Scopus--2017 International Conference on Data Mining, Communications and Information Technology(DMCIT 2017)	Phuket

1 matching rows (1187 total)

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

All		conference_acrc	conference_name	conference_loca
723.		EI--AIACT 2017	2017 International Conference on Artificial Intelligence, Automation and Control Technologies--Ei & SCOPUS	Wuhan

5. Deletion of Repeated Values

There were many data rows which were repeated multiple times because they may be crawled because same conference may belong to multiple categories like Machine Learning and Artificial Intelligence.

4 matching rows (1487 total)

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

All		conference_acrc	conference_name	conference_location
255.		KDMILE 2015	3rd Symposium on Knowledge Discovery Mining and Learning	Itaipava - Petropolis (Brazil)
257.		KDMILE 2015	3rd Symposium on Knowledge Discovery Mining and Learning	Itaipava - Petropolis (Brazil)
1414.		KDMILE 2015	3rd Symposium on Knowledge Discovery Mining and Learning	Itaipava - Petropolis (Brazil)
1415.		KDMILE 2015	3rd Symposium on Knowledge Discovery Mining and Learning	Itaipava - Petropolis (Brazil)

There were some repetitions that were not so obvious

6 matching rows (1487 total)

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

All		conference_acrc	conference_name	conference_locat
233.		ICDM 2015	ICDM 2015 Student Travel Grants	Atlantic City, NJ, USA
293.		ICDM 2015	International Conference on Data Mining	Atlantic City, NJ, USA
406.		ICDM 2015	ICDM 2015 Student Travel Grants	Atlantic City, NJ, USA
409.		ICDM 2015	International Conference on Data Mining	Atlantic City, NJ, USA
1387.		ICDM 2015	ICDM 2015 Student Travel Grants	Atlantic City, NJ, USA
1442.		ICDM 2015	International Conference on Data Mining	Atlantic City, NJ, USA

3 matching rows (1487 total)

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

All		conference_acrc	conference_name	conference_location
975.		ECAI 2016	8th International Conference on Electronics, Computers and Artificial Intelligence	Ploiesti, Romania
1024.		ECAI 2016	European Conference on Artificial Intelligence	The Hague, the Netherlands
1323.		ECAI 2016	European Conference on Artificial Intelligence	The Hague, the Netherlands

These case I handled using markdown feature. I marked down similar rows as blank leaving only 1 of the possible data entries among the repeated ones and finally selecting the all blank rows and deleting them.

6. Splitting city from location and Year from acronym

I split City from the conference_location column using “,” delimiter and retained only first column and deleted the rest of the column which could have been States and Countries column. After that I replaced space in location by “_” to make them a single word.
I also split Year from the conference_acronym to get Conference acronym and Conference Year. Which were helpful for many different purposes. I used “<space> “ as a delimiter and kept only last column and joined rest of them together.
I also replaced “,” as “<space>” in the name stop its interference with the csv values which I have used as input file for evaluation of MapReduce functions.
I again trimmed leading and trailing whitespace in the data set.

7. Removing Random Unwanted Words from Conference Name

There were many mistyped words which I removed from the Conference_Name column few examples of the “Deadline Extended” ,“Extension”, “Ei Compendex”, “Scopus”

I removed these kind of words as many as possible by searching them and replacing them using OpenRefine replace methods to make more sense from the conference names.

After cleaning as much as possible in my opinion this was the my final data set of 1162 rows. Cleaned up data in a comma separate value is in the **data_mining_refined.csv** file.

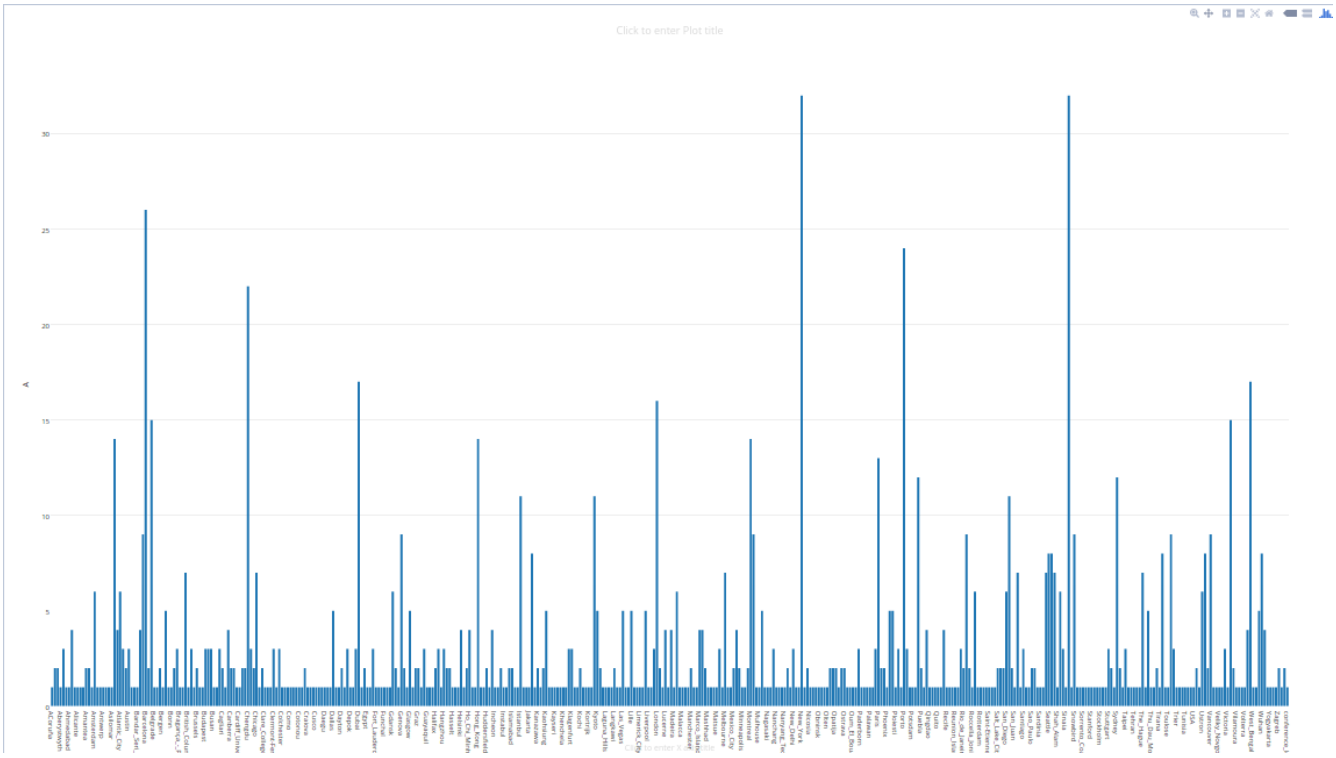
1162 rows Extensions

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

	All	conference_acrc	conference_yea	conference_name	conference_loca
1.	DKMP	2017	Fifth International Conference on Data Mining & Knowledge Management Process	Dubai
2.	CCSEA	2017	Seventh International Conference on Computer Science Engineering and Applications	Dubai
3.	IJCNN	2017	International Joint Conference on Neural Networks	Anchorage
4.	ACSTY	2016	Second International Conference on Advances in Computer Science and Information Technology	Chennai
5.	DMAP	2016	Second International Conference on Data Mining and Applications	Vienna
6.	FLAIRS	2017	FLAIRS-30: Special Track in Data Mining	Marco_Island
7.	BDAS	2017	13th (IEEE technically co-sponsored) International Conference Beyond Databases Architectures and Structures	Ustron
8.	ICBCT	2017	2017 International Conference on Bioinformatics and Computing Technologies (ICBCT 2017)	Hong_Kong
9.	Sideways	2017	Social Media World Sensors 2017	Madeira
10.	ICISDM	2017	2017 International Conference on Information System and Data Mining	South_Carolina
11.	ITMLS	2017	Intelligent Technologies and Methodologies of Learning Systems	Barcelona
12.	ITPF	2016	The 5th international conference on Information Technology Present and Future	Mashhad
13.	ICCSP	2017	International Conference on Cryptography Security and Privacy	Wuhan
14.	ICKSE	2017	3rd International Conference on Knowledge and Software Engineering edit	Paris
15.	WSDM	2017	International Conference on Web Search and Data Mining	Cambridge
16.	KDD	2017	Knowledge Discovery and Data Mining	Halifax
17.	ICOK	2017	2017 2nd International Conference on Knowledge (ICOK 2017)	Chengdu
18.	ICBDA	2017	2017 IEEE International Conference on Big Data Analysis	Beijing
19.	IEA/AIE	2017	Special Track on Novel Approaches to Anomaly Detection	Arras
20.	ICDPR	2017	2017 International Conference on Data Processing and Robotics	Kuala_Lumpur
21.	IEEE-ICCCBDA	2017	2nd International Conference on Cloud Computing and Big Data Analysis	Chengdu
22.	ICASC	2017	International Conference on Advances in Soft Computing	Kanyakumari
23.	DMCIT	2017	2017 International Conference on Data Mining Communications and Information Technology	Phuket

Part 3: Hadoop

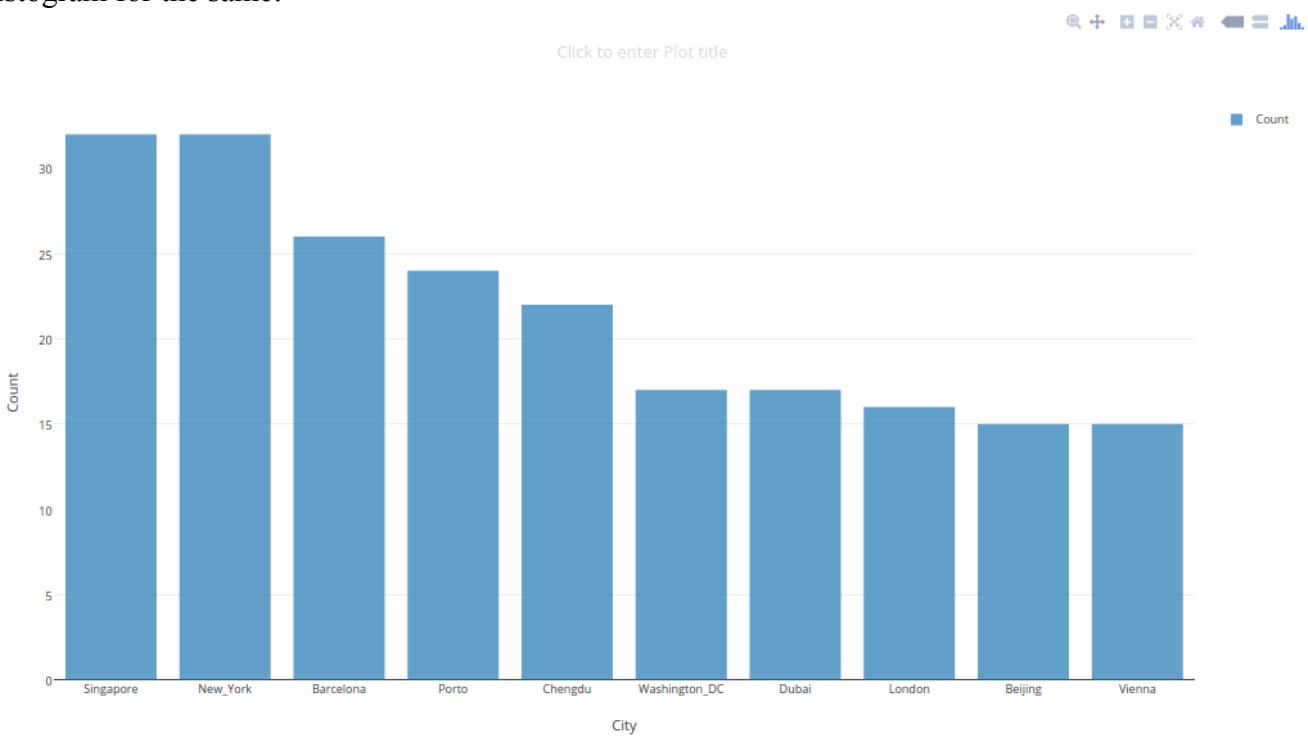
part a) To calculate the number of conferences per city I first split the row using "," as splitting point then from separate values I used conference_city and count 1 as the value.
Reducer added up all the values for a particular city resulting in total number of conference per city.
Result generated is in the **conf_per_city.txt** file and the code is in **conf_per_city.java**
Plot for the same is :



The Top 10 locations from my cleaned up data set is :

New_York:32, **Singapore**:32, **Barcelona**:26, **Porto**:24, **Chengdu**:22, **Dubai**:17, **Washington_DC**:17, **London**:16, **Beijing**:15, **Vienna**:15

Histogram for the same:



part b) To calculate the list of conferences per city I did the same thing as before by spliting the row using "," as splitting point then from separate values I used conference_city as the key and conference_acronym as the value.

In Reducer I concatenated all the value I get for a particular key which resulted in list of conference per city. But it can happen that if a particular conference happened in a particular city in different year it will result into repeated values so I also added the check in the Reducer that if a particular conference is already added then I wont add it again which will result only in unique conference occurred in the city. Although there were not many cases there this were frequent only for those cities where conferences occurs frequently.

Conference per city with repeated value is present in **list_conf_per_city.txt** and without repetition one is present in list_**conf_per_city_without_repetition.txt** and the code is in **list_conf_per_city.java**

Output example:

Acoruña : JISBD , **Aalborg** : CSE SPLINE, **Aberdeen** :EANN SoMePeAS, **Aberystwyth** : SAB, **Adelaide** : DPBA APCCM DICTA, **Agra** : HTC, **Ahmedabad** : COMAD, **Aizu-Wakamatsu** :DNIS MLSLP, **Aksaray** : SIN, **Alicante** : Big Data

part c) To calculate the list of cities per conference I did the same thing as above only change was I used conference_acronym as the key and conference_city as the value.
In Reducer I concatenated all the value I get for a particular key which will result in list of cities per conference. Same approach was taken to reduce the redundancy of cities per conference.

Cities per conference with repeated value is present in **list_city_per_conf.txt** and without repetition one is present in **list_city_per_conf_without_repetition.txt** and the code is in **list_city_per_conf.java**

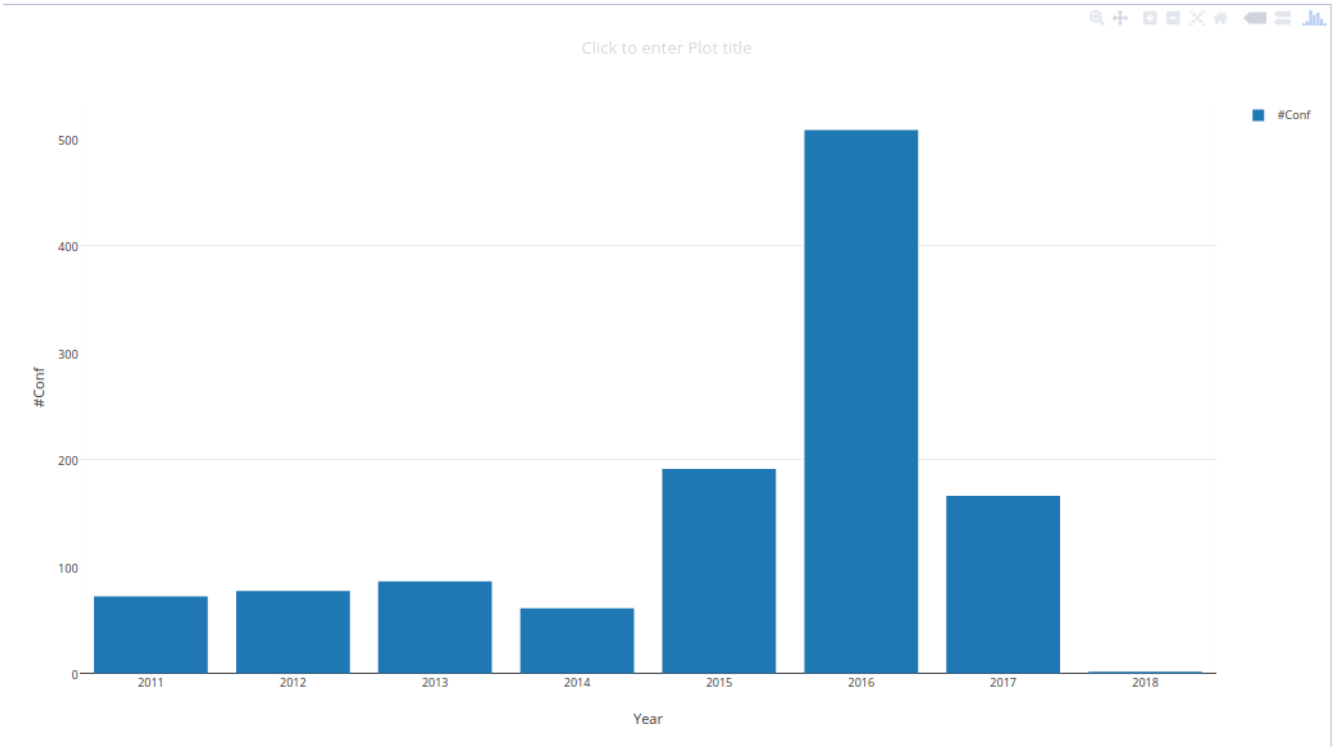
Output example:

4DML : USA, **AAIA** : Gdansk, **AAMAS** : Sao_Paulo Singapore, **ACALCI** : Canberra, **ACCV :** Taipai, **ACIIDS :** Kaohsiung Kuala_Lumpur, **ACIRS** : Tokyo Wuhan, **ACIS** : Krabi, **ACML** : Hong_Kong Hamilton, **ACMME** : Kuala_Lumpur

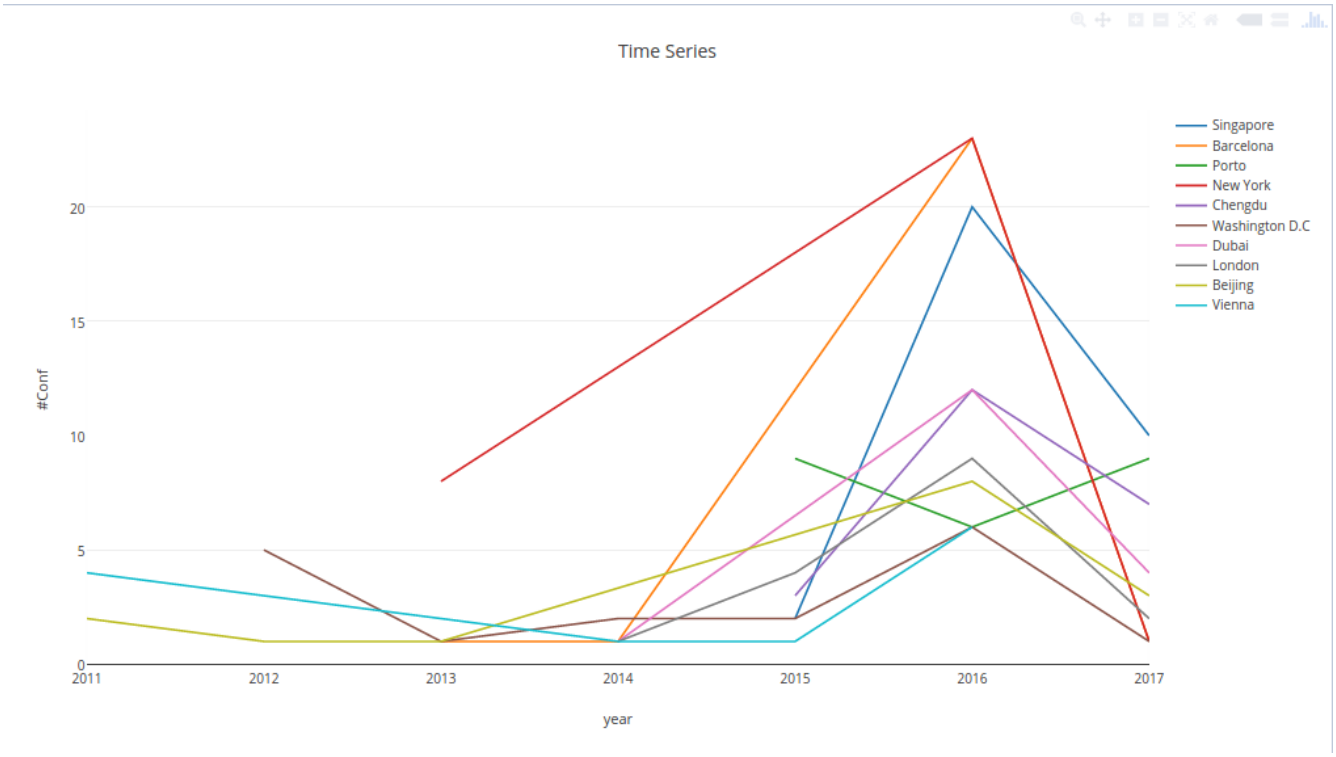
part d) To calculate the number of year wise conference for a particular I used conference_acronym + ":" + conference_year as the key and count '1' as the value.
Reason for using such key is it wont mix all the count for a particular city and wont mix count for a particular year. Thats why it will give me count of conference occurred in a particular city for a particular year.

In Reducer I just summed up all the value I get for a the key which will result in time series of conference per city.

This is the graph of number of conference occurred per year I calculate it just by using conference_year as a key and count 1 as the value and plotted the graph.



For Plotting the time series of a particular city I did it by again feeding the collected data into the OpenRefine and splitting conference acronym and for a particular city I extracted the year and count for top 10 cities and downloaded csv for each city.
Using online graph plotting tool I uploaded all csv's and plotted a line curve for each of the cities as year vs number of conferences and used different colored line to represent a different city.



Cities per conference with repeated value is present in **year_wise_conf.txt** and the code is in **year_wise_conf.java**

Output example:

ACoruña:2011	1
Aalborg:2013	1
Aalborg:2016	1
Aberdeen:2016	1
Aberdeen:2017	1
Aberystwyth:2016	1
Adelaide:2013	1
Adelaide:2015	2
Agra:2016	1
Ahmedabad:2013	1

References

[1] For All Python Related Stuff : http://docs.python-guide.org/en/latest/
[2] For Python Crawler : https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] For Fixing Compilation issues & Other Stuffs : https://stackoverflow.com
[4] For Graph Plotting : https://plot.ly/