Satamalee, Panitan
CS 235, Data Mining

Assignment Phase 2 Report

Hadoop

For the Hadoop portion, I altered the “WordCount” example that was given in the Apache tutorial. I created a separate Java/Hadoop program for each computation (labeled p1 for the first problem) which generated the results shown in the respective answer files (labeled p1_ans for the first problem).

In more detail, for the first problem, I included a delimiter for the StringTokenizer from word count to be the newline character, which is what separates the different data points. Then I split each value by the tab characters which gives me the conference acronym, conference name, and conference location. The first run I replace the key from WordCount to be the conference location, but, when I looked at the results, it showed that the preprocessing was not the best. I then split the conference location to only include the first word (or just the city most of the time) and that lead to better computations, although still not perfect.

For the second problem, instead of “one” being the value in the mapper from the first problem, I replaced it with the conference acronym, and I changed the reducer to keep attaching the acronyms together for the same conference location to create a long list of values at the end.

For the third problem, I swapped the key and value from the second problem. This means the key is the conference acronym, which was also parsed for the first word that doesn’t include the year, and the value is the city. The reducer was left the same since the value from the mapper is a Text variable that needed to be concatenated.

For the fourth problem, I reverted the reducer to the one from WordCount. I then set the value in the mapper back to one. The key in the mapper is the parsed city name concatenated with the year of the conference which was obtained from the conference acronym.

*Examples of outputs from Hadoop are shown below (Figures 11-14).

Top 10 Location According to Number of Conferences (Regardless of Year):

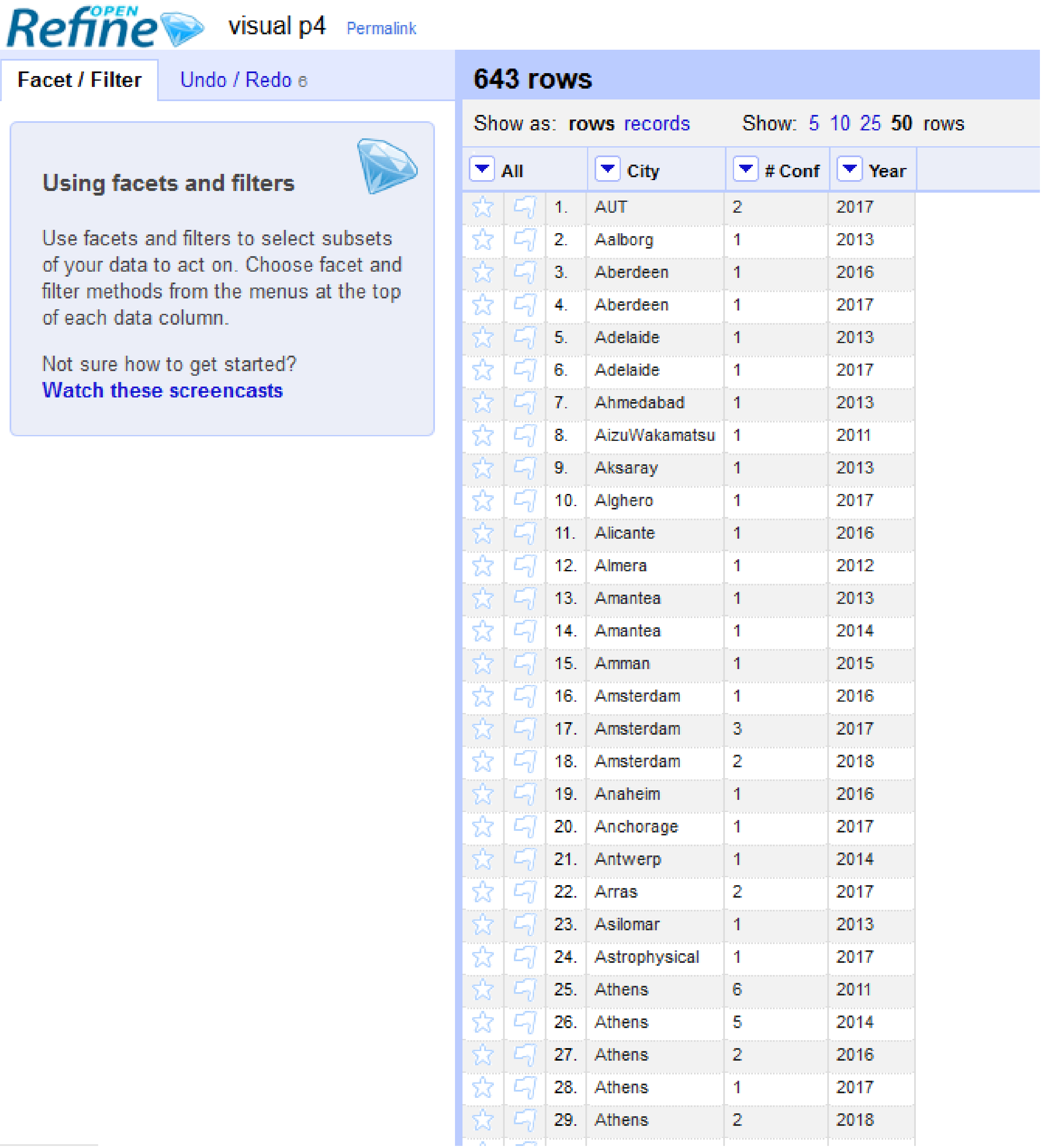
1. Barcelona (29)
2. New York (27)
3. Dubai (26)
4. Melbourne (23)
5. San Francisco (21)
6. Vienna (17)
7. Athens (16)
8. Prague (16)
9. Singapore (16)
10. Shanghai (15)

Visualization

For the visualization portion, I utilized the Fusion Table feature that is available on Google Drive. This takes in a table of cities and creates a heatmap using Google Maps. I was able to take a screen shot of each time period as I was able to filter through the years using the app's features. I had to process the Hadoop data using OpenRefine and Python before importing into Fusion Tables

In OpenRefine, I separated the years and conference acronyms into two separate columns (Fig. 1). This allows for easier filtering in Fusion Table. Then I used Python (Fig. 2) to duplicate rows where more than one conference happened during that year. This allows for Fusion Table to create the heat map since it takes into account how many rows of the same cities there are. The weight feature of the heatmap for Fusion Table seems very inaccurate when using with number of conferences since most cities don't show up if the number is too low.

From the visualization, we see that the amount of conferences continuously grow from 2011 to 2015 when it seems to be more congregated in Europe. Most of the conferences in this time are in Europe with a few starting in North America and Southeast Asia. After that the conferences grew even more around Europe and even more in North America and South East Asia. We also see more growth in several other regions such as India and Australia.



visual p4

643 rows

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

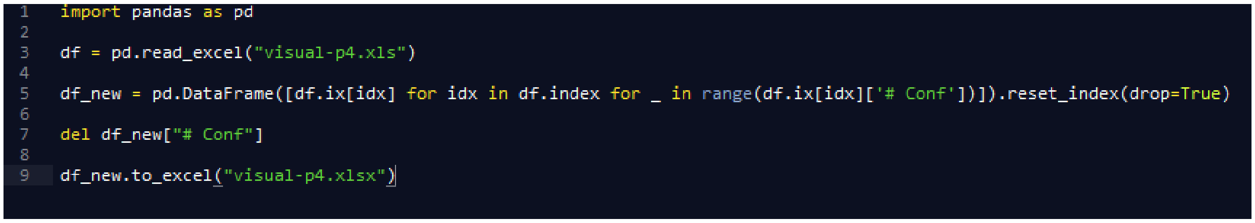
Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	City	# Conf	Year
1.	AUT	2	2017
2.	Aalborg	1	2013
3.	Aberdeen	1	2016
4.	Aberdeen	1	2017
5.	Adelaide	1	2013
6.	Adelaide	1	2017
7.	Ahmedabad	1	2013
8.	AizuWakamatsu	1	2011
9.	Aksaray	1	2013
10.	Alghero	1	2017
11.	Alicante	1	2016
12.	Almera	1	2012
13.	Amantea	1	2013
14.	Amantea	1	2014
15.	Amman	1	2015
16.	Amsterdam	1	2016
17.	Amsterdam	3	2017
18.	Amsterdam	2	2018
19.	Anaheim	1	2016
20.	Anchorage	1	2017
21.	Antwerp	1	2014
22.	Arras	2	2017
23.	Asilomar	1	2013
24.	Astrophysical	1	2017
25.	Athens	6	2011
26.	Athens	5	2014
27.	Athens	2	2016
28.	Athens	1	2017
29.	Athens	2	2018

Figure 1. OpenRefine process for visualization. Separating conference acronym to two columns.



```
import pandas as pd

df = pd.read_excel("visual-p4.xls")

df_new = pd.DataFrame([df.ix[idx] for idx in df.index for _ in range(df.ix[idx]['# Conf'])]).reset_index(drop=True)

del df_new["# Conf"]

df_new.to_excel("visual-p4.xlsx")
```

Figure 2. Python code for making duplicate rows according to conference number.

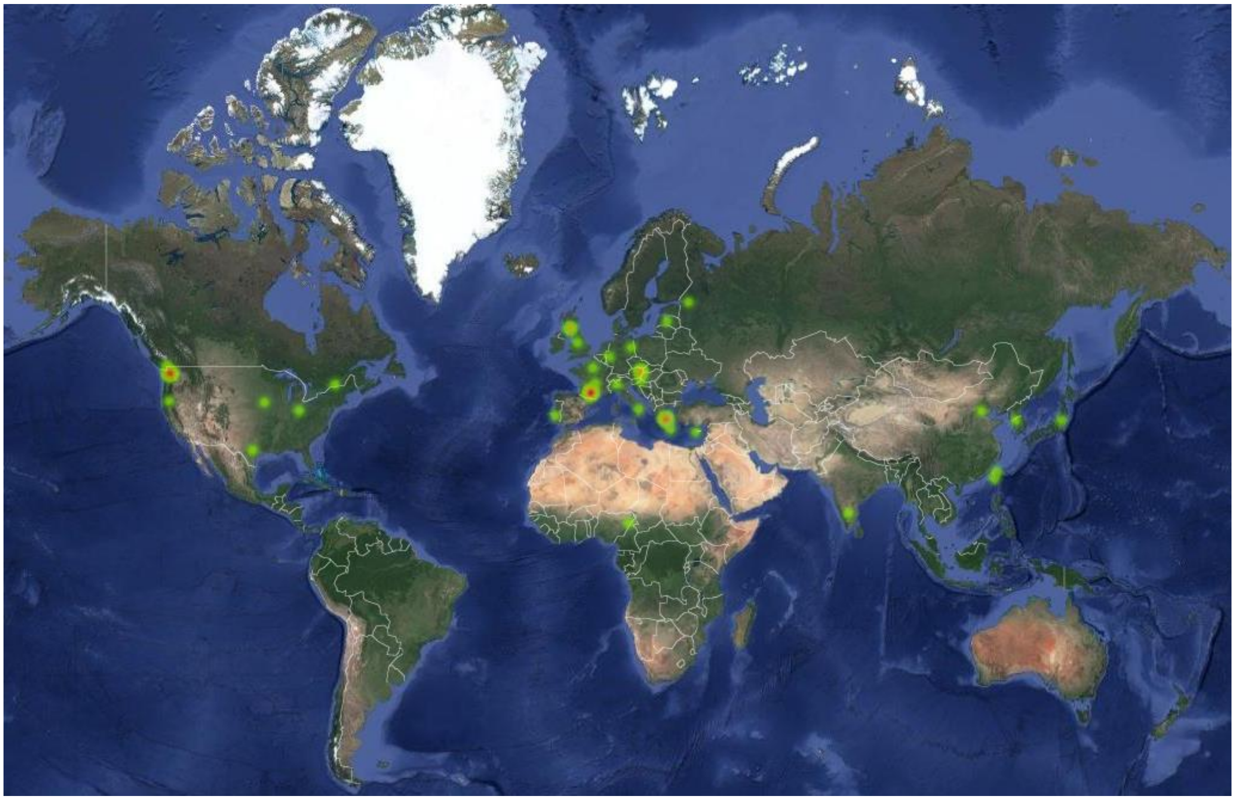


Figure 3. Heatmap for 2011 conferences.

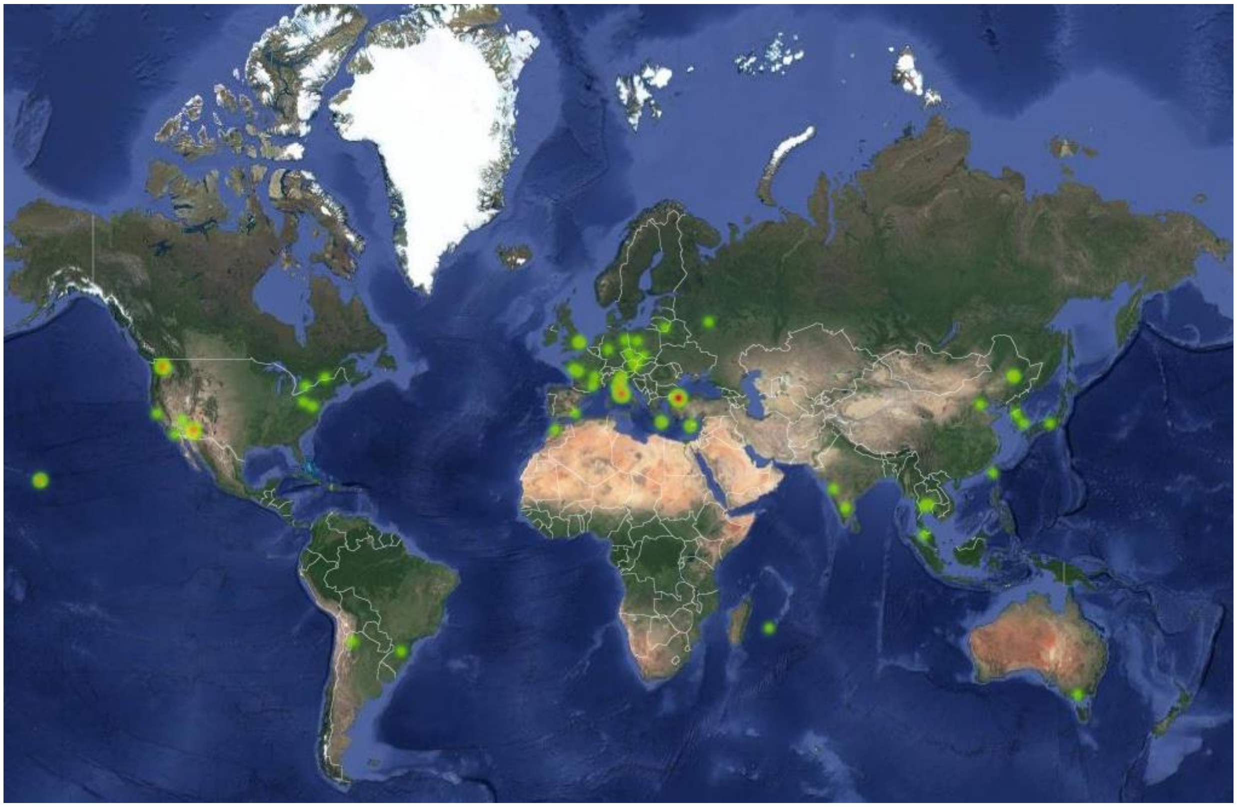


Figure 4. Heatmap for 2012 conferences.

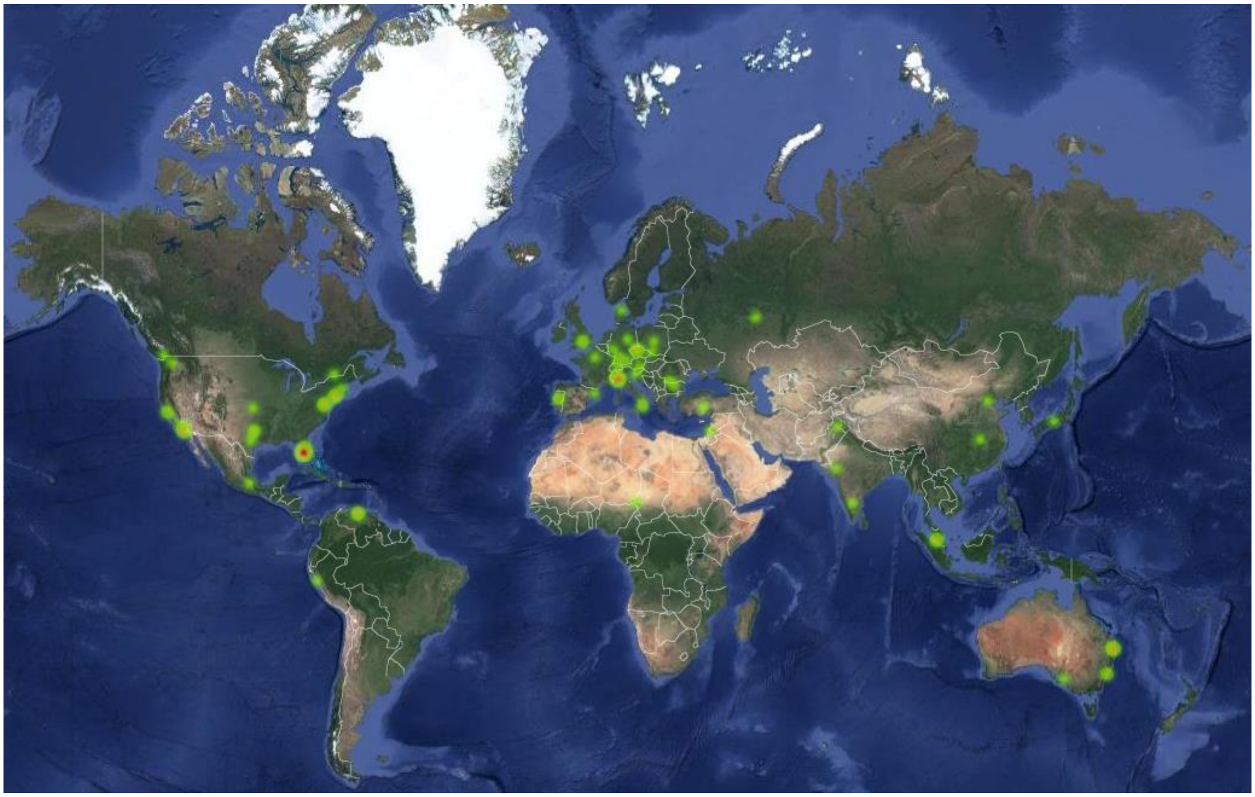


Figure 5. Heatmap for 2013 conferences

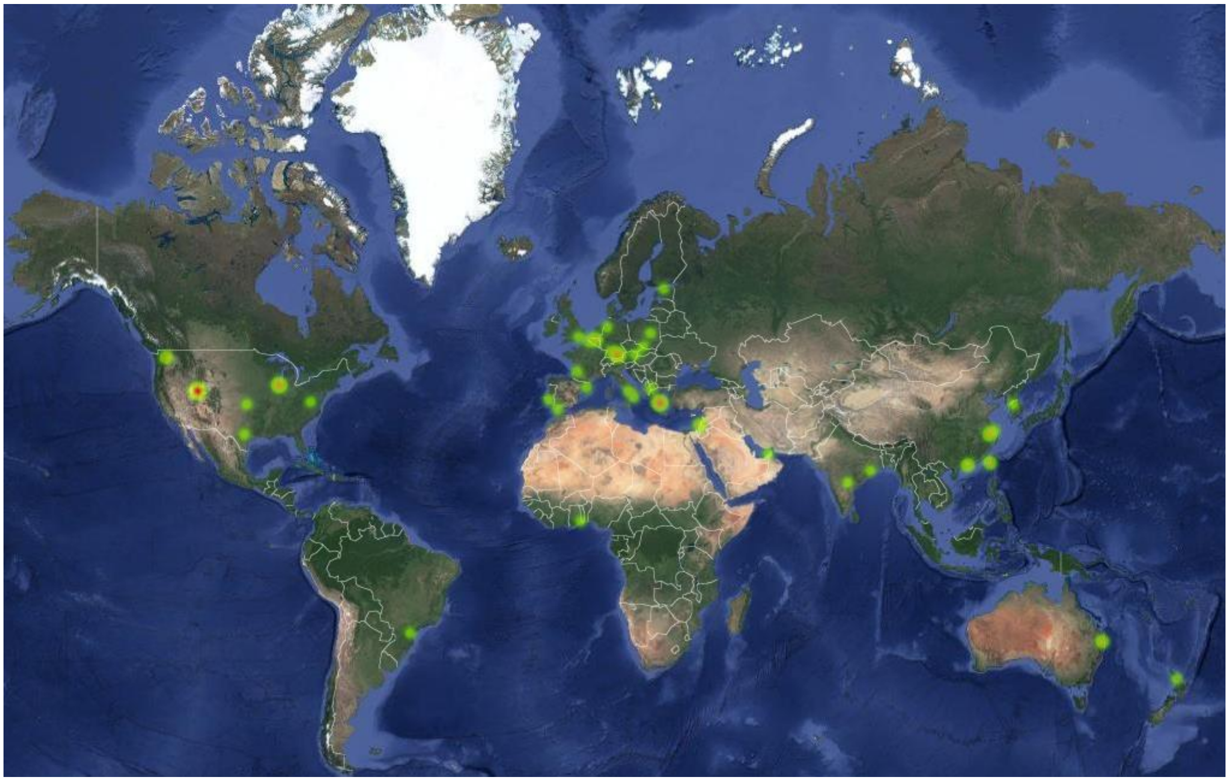


Figure 6. Heatmap for 2014 conferences.

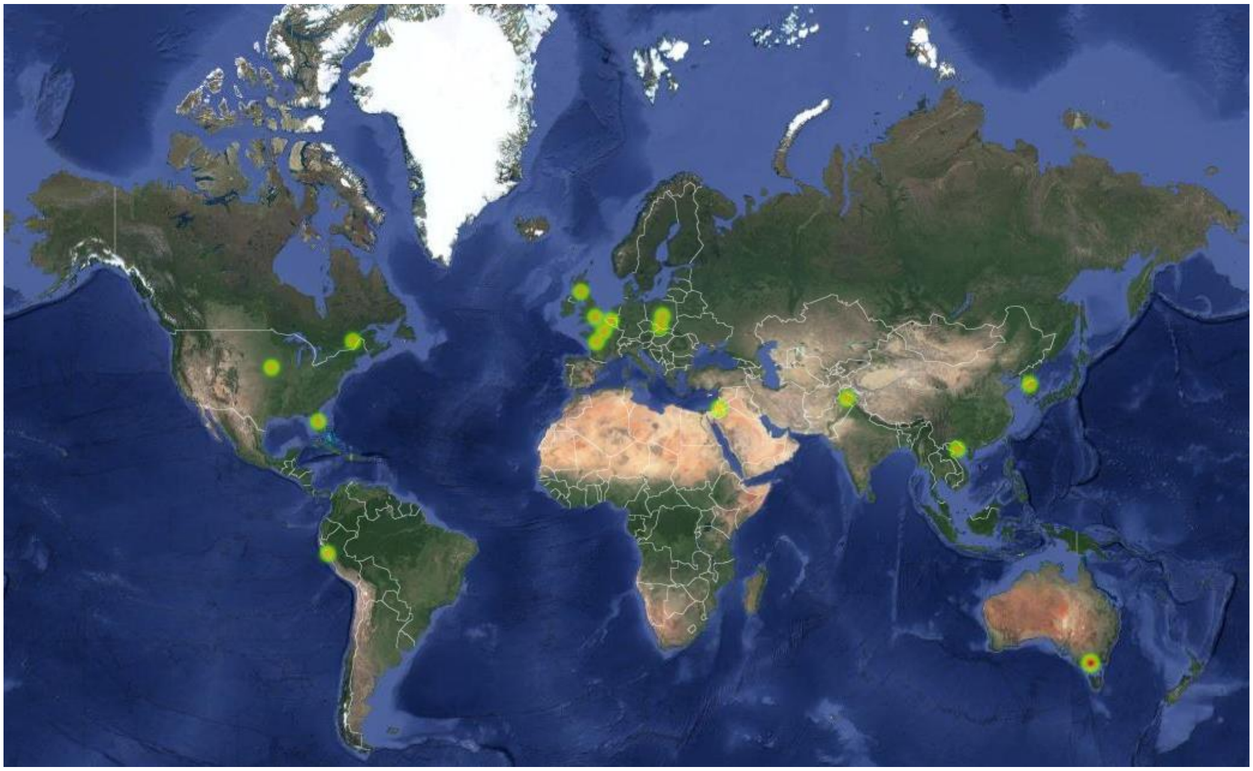


Figure 7. Heatmap for 2015 conferences

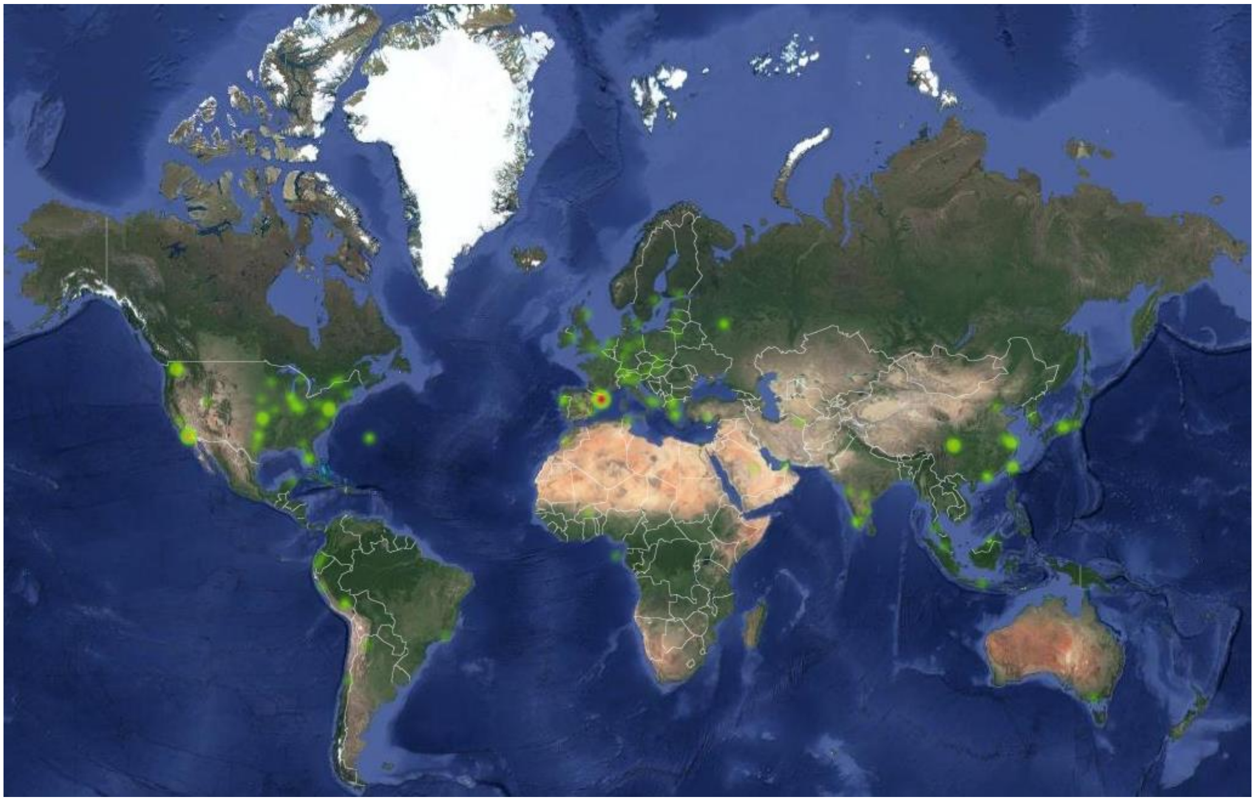


Figure 8. Heatmap for 2016 conferences.

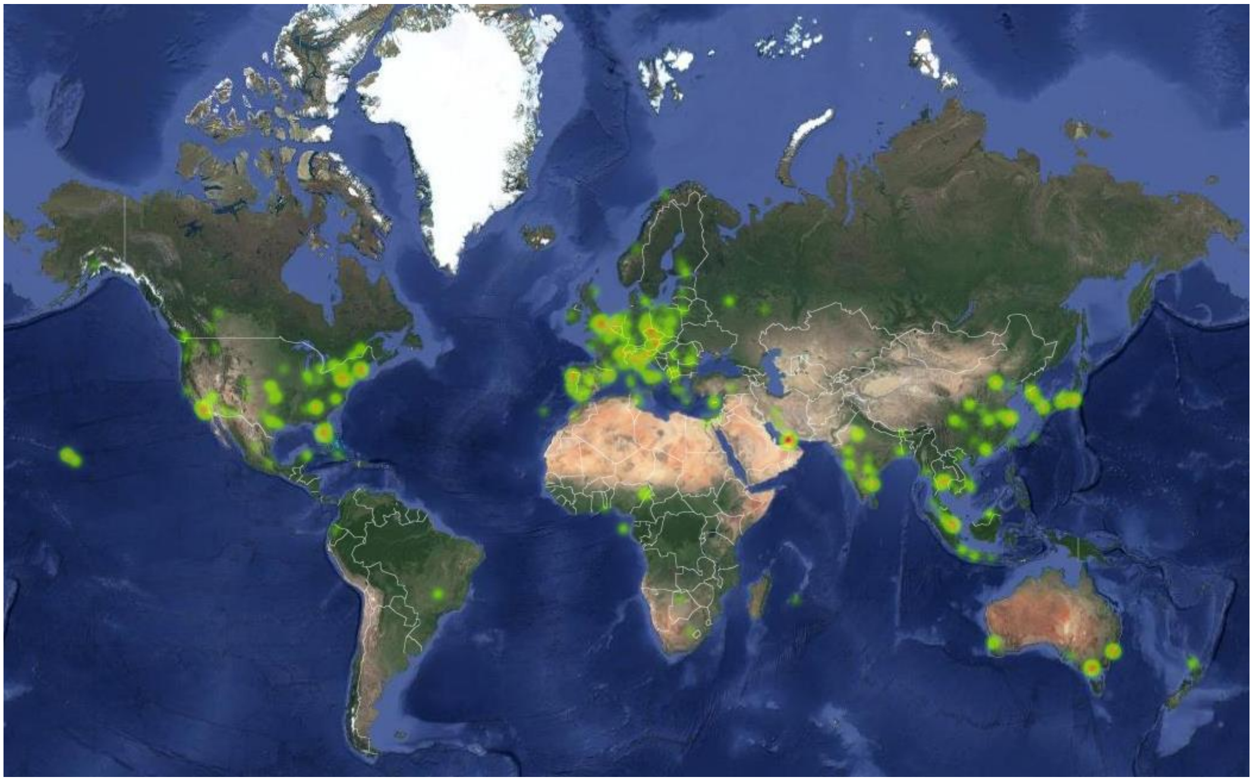


Figure 9. Heatmap for 2017 conferences.

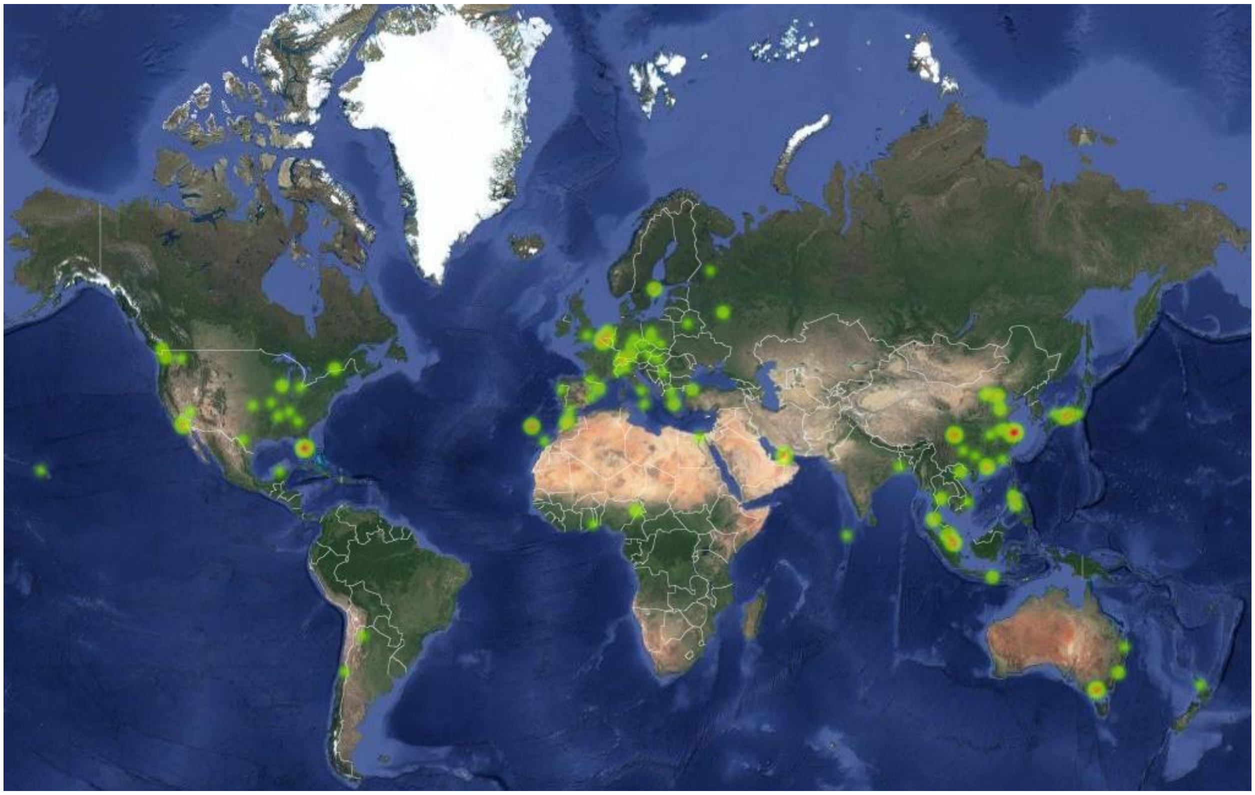
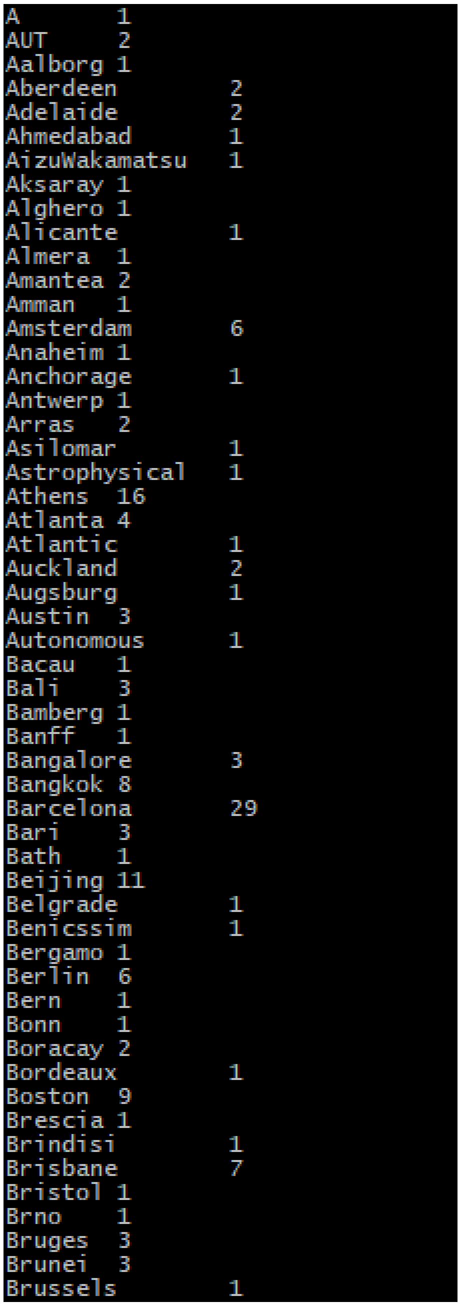
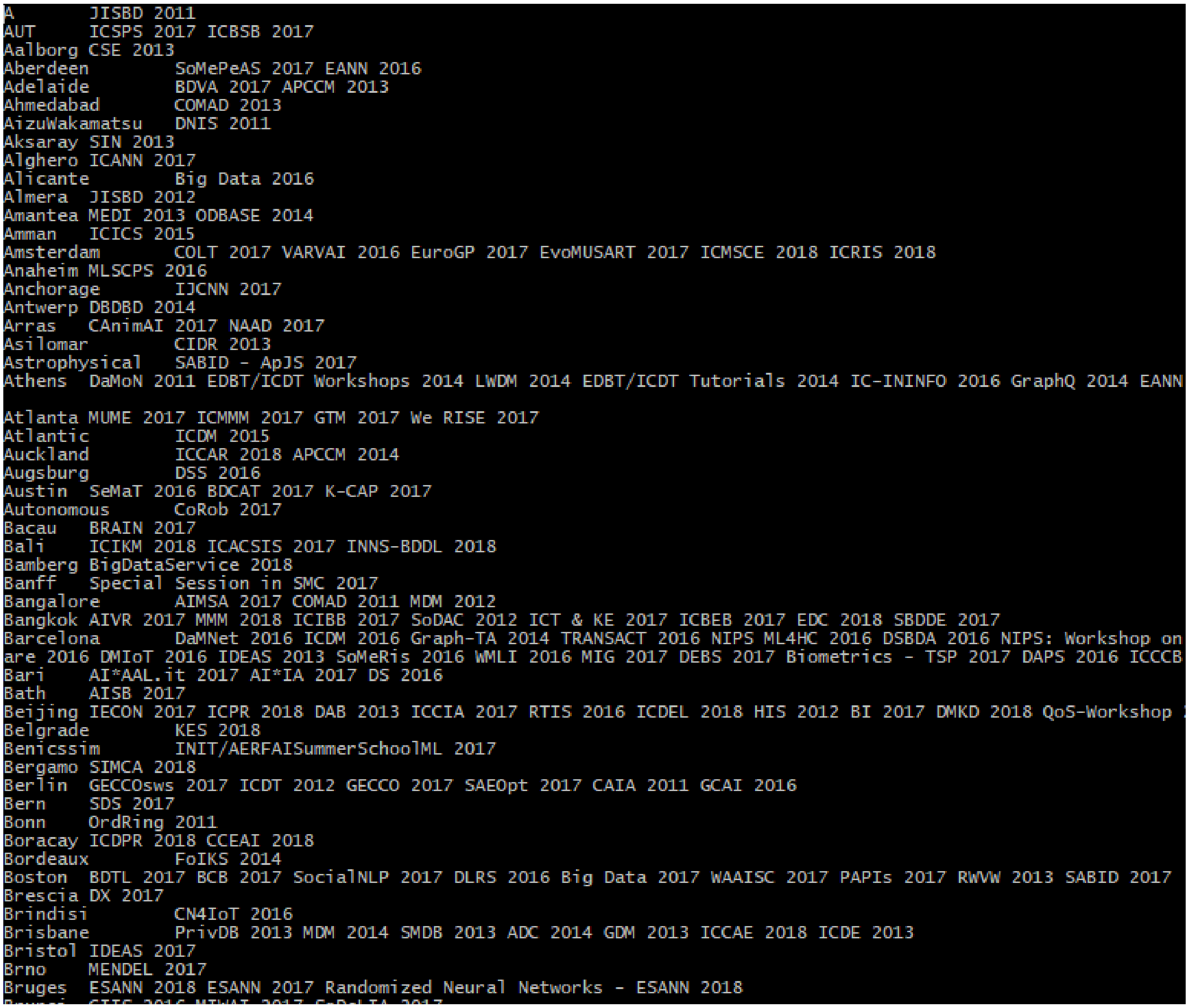


Figure 10. Heatmap for 2018 conferences.



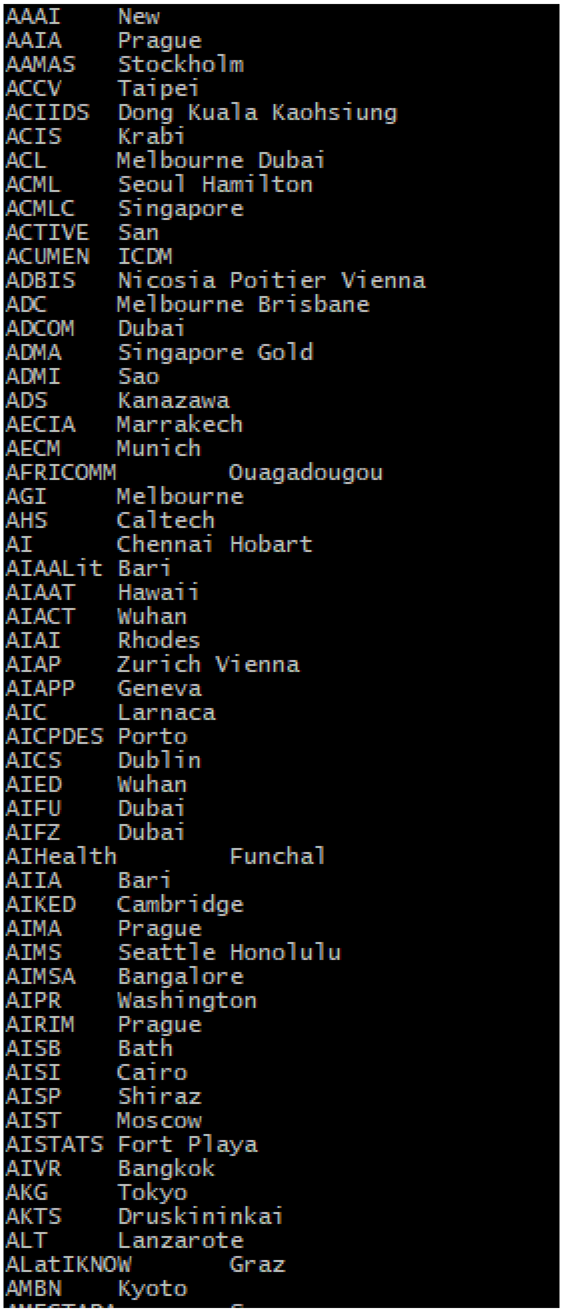
```
A       1
AUT     2
Aalborg 1
Aberdeen        2
Adelaide        2
Ahmedabad       1
AizuWakamatsu   1
Aksaray 1
Alghero 1
Alicante        1
Almera  1
Amantea 2
Amman   1
Amsterdam       6
Anaheim 1
Anchorage       1
Antwerp 1
Arras   2
Asilomar        1
Astrophysical   1
Athens  16
Atlanta 4
Atlantic        1
Auckland        2
Augsburg        1
Austin  3
Autonomous      1
Bacau   1
Bali    3
Bamberg 1
Banff   1
Bangalore       3
Bangkok 8
Barcelona       29
Bari    3
Bath    1
Beijing 11
Belgrade        1
Benicssim       1
Bergamo 1
Berlin  6
Bern    1
Bonn    1
Boracay 2
Bordeaux        1
Boston  9
Brescia 1
Brindisi        1
Brisbane        7
Bristol 1
Brno    1
Bruges  3
Brunei  3
Brussels        1
```

Figure 11. Example answer for problem 1.



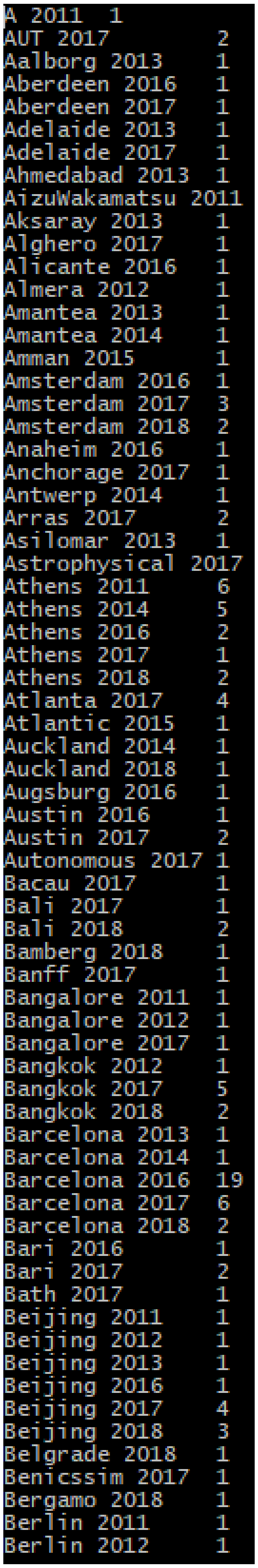
```
A       JISBD 2011
AUT     ICSPS 2017 ICBSB 2017
Aalborg CSE 2013
Aberdeen        SoMePeAS 2017 EANN 2016
Adelaide        BDVA 2017 APCCM 2013
Ahmedabad       COMAD 2013
AizuWakamatsu   DNIS 2011
Aksaray SIN 2013
Alghero ICANN 2017
Alicante        Big Data 2016
Almera  JISBD 2012
Amantea MEDI 2013 ODBASE 2014
Amman   ICICS 2015
Amsterdam       COLT 2017 VARVAI 2016 EuroGP 2017 EvoMUSART 2017 ICMSCE 2018 ICRIS 2018
Anaheim MLSCPS 2016
Anchorage       IJCNN 2017
Antwerp DBDBD 2014
Arras   CAnimAI 2017 NAAD 2017
Asilomar        CIDR 2013
Astrophysical   SABID - ApJS 2017
Athens  DaMoN 2011 EDBT/ICDT Workshops 2014 LWDM 2014 EDBT/ICDT Tutorials 2014 IC-ININFO 2016 GraphQ 2014 EANN

Atlanta MUME 2017 ICMMM 2017 GTM 2017 We RISE 2017
Atlantic        ICDM 2015
Auckland        ICCAR 2018 APCCM 2014
Augsburg        DSS 2016
Austin  SeMaT 2016 BDCAT 2017 K-CAP 2017
Autonomous      CoRob 2017
Bacau   BRAIN 2017
Bali    ICIKM 2018 ICACSIS 2017 INNS-BDDL 2018
Bamberg BigDataService 2018
Banff   Special Session in SMC 2017
Bangalore       AIMSA 2017 COMAD 2011 MDM 2012
Bangkok AIVR 2017 MMM 2018 ICIBB 2017 SoDAC 2012 ICT & KE 2017 ICBEB 2017 EDC 2018 SBDDE 2017
Barcelona       DaMNet 2016 ICDM 2016 Graph-TA 2014 TRANSACT 2016 NIPS ML4HC 2016 DSBDA 2016 NIPS: Workshop on
are 2016 DMIoT 2016 IDEAS 2013 SoMeRis 2016 WMLI 2016 MIG 2017 DEBS 2017 Biometrics - TSP 2017 DAPS 2016 ICCCB
Bari    AI*AAL.it 2017 AI*IA 2017 DS 2016
Bath    AISB 2017
Beijing IECON 2017 ICPR 2018 DAB 2013 ICCIA 2017 RTIS 2016 ICDEL 2018 HIS 2012 BI 2017 DMKD 2018 QoS-Workshop
Belgrade        KES 2018
Benicssim       INIT/AERFAISummerSchoolML 2017
Bergamo SIMCA 2018
Berlin  GECCOsws 2017 ICDT 2012 GECCO 2017 SAEOpt 2017 CAIA 2011 GCAI 2016
Bern    SDS 2017
Bonn    OrdRing 2011
Boracay ICDPR 2018 CCEAI 2018
Bordeaux        FoIKS 2014
Boston  BDTL 2017 BCB 2017 SocialNLP 2017 DLRS 2016 Big Data 2017 WAAISC 2017 PAPIs 2017 RWVW 2013 SABID 2017
Brescia DX 2017
Brindisi        CN4IoT 2016
Brisbane        PrivDB 2013 MDM 2014 SMDB 2013 ADC 2014 GDM 2013 ICCAE 2018 ICDE 2013
Bristol IDEAS 2017
Brno    MENDEL 2017
Bruges  ESANN 2018 ESANN 2017 Randomized Neural Networks - ESANN 2018
```

Figure 12. Example answer for problem 2.



```
AAAI    New
AAIA    Prague
AAMAS   Stockholm
ACCV    Taipei
ACIIDS  Dong Kuala Kaohsiung
ACIS    Krabi
ACL     Melbourne Dubai
ACML    Seoul Hamilton
ACMLC   Singapore
ACTIVE  San
ACUMEN  ICDM
ADBIS   Nicosia Poitier Vienna
ADC     Melbourne Brisbane
ADCOM   Dubai
ADMA    Singapore Gold
ADMI    Sao
ADS     Kanazawa
AECIA   Marrakech
AECM    Munich
AFRICOMM        Ouagadougou
AGI     Melbourne
AHS     Caltech
AI      Chennai Hobart
AIAALit Bari
AIAAT   Hawaii
AIACT   Wuhan
AIAI    Rhodes
AIAP    Zurich Vienna
AIAPP   Geneva
AIC     Larnaca
AICPDES Porto
AICS    Dublin
AIED    Wuhan
AIFU    Dubai
AIFZ    Dubai
AIHealth        Funchal
AIIA    Bari
AIKED   Cambridge
AIMA    Prague
AIMS    Seattle Honolulu
AIMSA   Bangalore
AIPR    Washington
AIRIM   Prague
AISB    Bath
AISI    Cairo
AISP    Shiraz
AIST    Moscow
AISTATS Fort Playa
AIVR    Bangkok
AKG     Tokyo
AKTS    Druskininkai
ALT     Lanzarote
ALatIKNOW       Graz
AMBN    Kyoto
```

Figure 13. Example answer for problem 3.



```
A 2011  1
AUT 2017        2
Aalborg 2013    1
Aberdeen 2016   1
Aberdeen 2017   1
Adelaide 2013   1
Adelaide 2017   1
Ahmedabad 2013  1
AizuWakamatsu 2011
Aksaray 2013    1
Alghero 2017    1
Alicante 2016   1
Almera 2012     1
Amantea 2013    1
Amantea 2014    1
Amman 2015      1
Amsterdam 2016  1
Amsterdam 2017  3
Amsterdam 2018  2
Anaheim 2016    1
Anchorage 2017  1
Antwerp 2014    1
Arras 2017      2
Asilomar 2013   1
Astrophysical 2017
Athens 2011     6
Athens 2014     5
Athens 2016     2
Athens 2017     1
Athens 2018     2
Atlanta 2017    4
Atlantic 2015   1
Auckland 2014   1
Auckland 2018   1
Augsburg 2016   1
Austin 2016     1
Austin 2017     2
Autonomous 2017 1
Bacau 2017      1
Bali 2017       1
Bali 2018       2
Bamberg 2018    1
Banff 2017      1
Bangalore 2011  1
Bangalore 2012  1
Bangalore 2017  1
Bangkok 2012    1
Bangkok 2017    5
Bangkok 2018    2
Barcelona 2013  1
Barcelona 2014  1
Barcelona 2016  19
Barcelona 2017  6
Barcelona 2018  2
Bari 2016       1
Bari 2017       2
Bath 2017       1
Beijing 2011    1
Beijing 2012    1
Beijing 2013    1
Beijing 2016    1
Beijing 2017    4
Beijing 2018    3
Belgrade 2018   1
Benicssim 2017  1
Bergamo 2018    1
Berlin 2011     1
Berlin 2012     1
```

Figure 14. Example answer for problem 4.