README

A Hadoop map/reduce program for parallel processing of the DBLP dataset. Various statistics about the number of co-authors is computed. The following statistics were computed;

1. Bin the number of co-authors (e.g., BIN_1 is 1 co-author, BIN_2 is 2-3 co-authors, BIN_3 is 4-6 co-authors, BIN_4 is 7-10 co-authors, BIN_5 is 11-15 co-authors and the rest in BIN_6).
2. Histogtram stratified by journals, inproceddings, and years of publications
3. Stratified breakdown of co-author count by publication venues in addition to the cumulative statistics across all venues.
4. The max, median, and the average number of authors for publication on which the name of the author appears.
5. Authorship score for each author calculated based on "*the score of the last co-author is credited 1/(4*N) *leaving it* 3*N/4 of the original score. The next co-author to the left is debited 1/(4*N) *and the process repeats until the first author is reached."*
6. List of top 100 authors in the descending order who publish with most co-authors and the list of 100 authors who publish with least co-authors.

To Run:

1. Run the command "**sbt clean compile assembly**"

2. Copy the generated EuniceDaphneHW2.jar file to HDP Sandbox VM

3. Load the dblp.xml as input file to HDP Sandbox VM

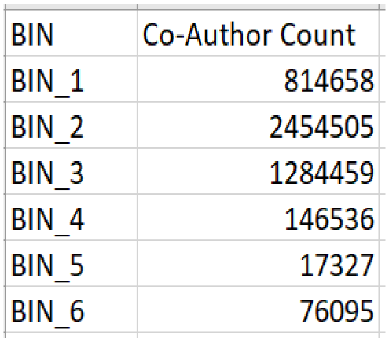
4. Run the following command,

   ```
   hadoop jar EuniceDaphneHW2.jar path/to/input/file path/to/output/file
   ```

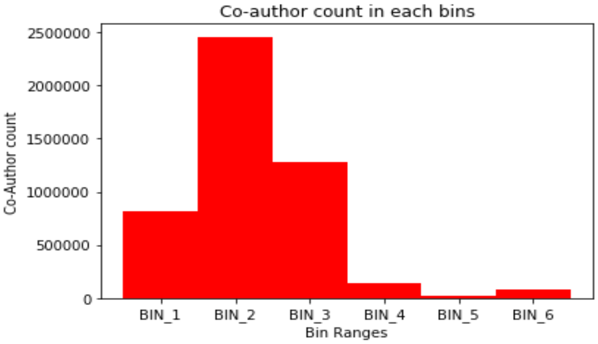
5. Copy the generated output file "part-00000" to local machine and run the python script (Cloud.ipynb) to convert each statistic to separate csv file, draw histogram and to obtain the top 100 authors in ascending and descending order.
   (*Note: Change the source file and destination folder name at the top of the "resources/Cloud.ipynb" file*)

6. The sample output for each statistic is as follows,

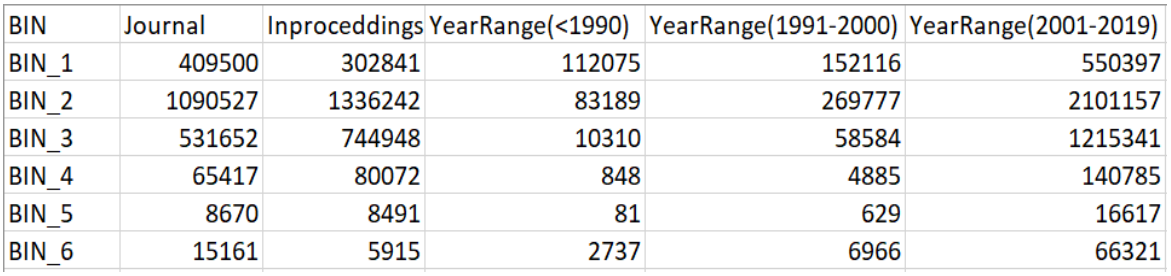
   (i) The histogram of co-author count in each bin is stored in *".../CoAuthor.csv"*. These are the sample csv table and histogram obtained.



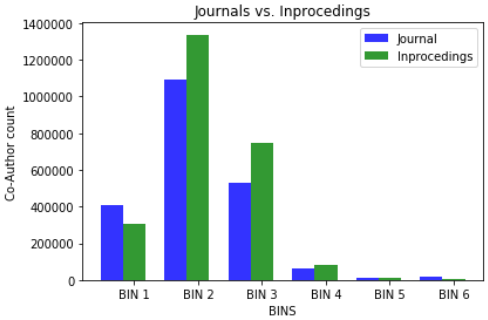
BIN	Co-Author Count
BIN_1	814658
BIN_2	2454505
BIN_3	1284459
BIN_4	146536
BIN_5	17327
BIN_6	76095

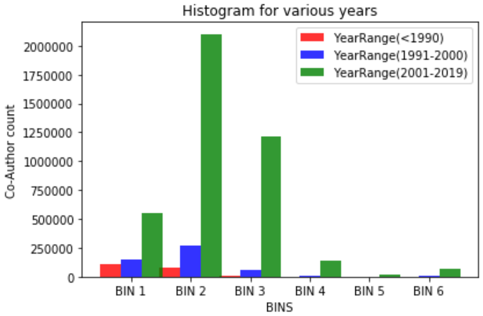


(ii) Histogtram stratified by journals, inproceddings, and years of publications. The table is stored in *".../journal_inproc_year.csv"*

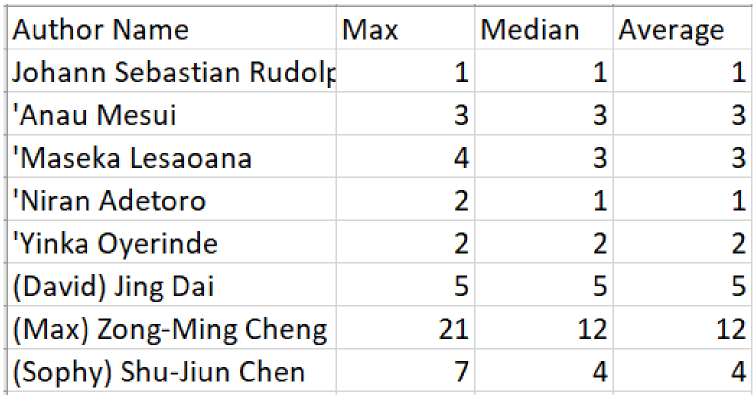


BIN	Journal	Inproceddings	YearRange(<1990)	YearRange(1991-2000)	YearRange(2001-2019)
BIN_1	409500	302841	112075	152116	550397
BIN_2	1090527	1336242	83189	269777	2101157
BIN_3	531652	744948	10310	58584	1215341
BIN_4	65417	80072	848	4885	140785
BIN_5	8670	8491	81	629	16617
BIN_6	15161	5915	2737	6966	66321





(iii) The max, median, and the average number of authors for publication on which the name of the author appears is computed in the ".../MMA.csv".



Author Name	Max	Median	Average
Johann Sebastian Rudolp	1	1	1
'Anau Mesui	3	3	3
'Maseka Lesaoana	4	3	3
'Niran Adetoro	2	1	1
'Yinka Oyerinde	2	2	2
(David) Jing Dai	5	5	5
(Max) Zong-Ming Cheng	21	12	12
(Sophy) Shu-Jiun Chen	7	4	4

(iv) List of Top 100 authors with least co-authors

```
['Ai Kaiho', '0.00379']
['Akiko Saka', '0.00379']
['Alan J. Knox', '0.00379']
['Albert S. B. Edge', '0.00379']
['Alessandro Bonetti', '0.00379']
['Alka Saxena', '0.00379']
['Anthony G. Beckhouse', '0.00379']
.
.
.
['Thomas J. Ha', '0.00379']
['Tomokatsu Ikawa', '0.00379']
['Tony J. Kenna', '0.00379']
['Toshio Kitamura', '0.00379']
['Tsugumi Kawashima', '0.00379']
```

(v) List of Top 100 authors with the most co-authors

```
['Edward J. Delp', '99.950424']
['Kun Wang', '99.922455']
['Madhu Sudan', '99.83746']
['Sadaaki Miyamoto', '99.816696']
['Jiandong Li 0001', '99.80573']
['Toshihide Ibaraki', '99.801094']
['Yukio Ohsawa', '99.77469']
['Xiao Li', '99.76136']
['Jianfeng Ma', '99.72973']
['Jun Ma', '99.72175']
['Noam Nisan', '99.71767']
.
.
.
.
['Godfried T. Toussaint', '97.13786']
['Kokichi Sugihara', '97.133736']
['Peng Xu', '97.11642']
['Xinghuo Yu', '97.048096']
['Ling Li', '97.02137']
```

The Youtube link on steps to deploy the map/reduce program on Amazon EMR can be found here.

Resources Folder:

Contains Cloud.ipynb for generating the csv files and the top 100 author counts. Also, a pdf of the Cloud.ipynb with results. Use Jupyter Notebook to run the .ipynb file.

Implementation:

Mapper:

- The mapper will output the key-value pairs uniquely for each statistic as a string is attached to each key to identify that statistic.

- The key to obtain co-author count in each bin will contain ***"Co-AuthorCount,+BIN_1"*** if the number of co-authors is 1 for that record.

- The key to obtain the author score will contain ***"AuthorScore,+score"***, where *score* is the calculated authorship score for that author.

- The key to obtain the stratified breakdown of co-author count in all venues will contain ***"All_venues,+bin"*** where bin can be BIN_1, BIN_2,….

- The key to obtain the stratified breakdown of co-author count in all venues will contain ***"MMA,+author"*** where author is the name of the author.

- The key to obtain the stratified histogram for journal, inproceedings and years will contain ***"Journal_Inproceedings_Year,+bin"*** where bin can be BIN_1, BIN_2,….

Reducer:

- The reducer will split the key and get the string corresponding to the statistics and perform the reducer function for that statistics.

- Based on the key the reducer will calculate sum, max, median and average operations.