Assignment 3 – CIRViz

Submission Date: April 1, Monday *by 10pm*
Demo (During the week of April 1, Schedule to be decided near to the time)

Type: Pair work **Weightage: 15%**

I. Introduction

Each minute, thousands of scientific documents are added to our knowledge base, however, digesting a big source of text or even making out meaningful insights from the data is challenging. General trends on the research topic and tracking their future implications are of great interest among people. Such trends also drive industry innovations.

. The purpose of this Conference Information Retrieval (CIRViz) assignment is to:

a) Extract the first n lines (n >= 200,000) from the (full) dataset at http://labs.semanticscholar.org/corpus/ and use them for the visualization tasks listed in section III. Note that the dataset is ~19GB in size. The download shall take some time. Also the lines you download will require a few GB of RAM for processing. Please contact us at cs5346.tutor@gmail.com in case you face any problem with download or extraction of data.

b) Visualize(given tasks) using a Viz tool. You could choose any visualization tool or framework to do this assignment (some examples include d3.js in JavaScript and ggplot in R or Tableau). If you have any doubt or query about which tool to use, contact tutor or lecturer at cs5346.tutor@gmail.com by 10 March.

II. Important Information

1. Read this document carefully.

2. If you have any query about this Assignment, send mail to cs5346.tutor@gmail.com. Work ahead of time and do not leave things to last minute.

3. A note about dataset:
The dataset at http://labs.semanticscholar.org/corpus/ provides data about over 7 million published research papers in Computer Science and Neuroscience. You will find two links – Full and Sample. For the assignment, extract the first n lines (n >= 200,000) of the FULL dataset. The link also gives short description of data attributes and an example.

4. Demonstrate visualizations, corresponding to the tasks set in Section III, **to your tutor in the week of April 1** (a schedule will be decided near to the time).

5. Submit a file labelled **{matric number-1}_{matric-number-2}_A3_CIRViz.zip** ,*omitting the brackets,* **containing**

(i) code (e.g. HTML & JS files with d3.min.js, tableau workbook). The code should be self contained and able to run without external dependencies,
(ii) a report (see template at the end of this document) . Your report could typically be 4-8 pages, and a single pdf. Exceeding the page guideline of 4-8 pages does not invite any penalty. **Label** the report document: **{Matric-number-1}_{Matric-number-2}_ CIRVizReport omitting the brackets** e.g. A0045396X_A0046342Y_CIRVizReport.pdf
(iii) a readme file containing any information about your code files or on how to run your code, or URL of your uploaded workbook in tableau public.

Submission folder : IVLE→Files→Student Submissions → A3-CIRViz

III. Task

Use **3 or more** different types of visualizations to achieve the tasks given below. Each task should be covered by at least 1 visualization.

1. Visualize the top 10 authors for venue arXiv based on the number of publications he/she has made across all available years for arXiv.

2. Visualize the top 5 papers for venue arXiv based on the number of citations across all available years for arXiv. (how many times this paper has been cited, so consider those with the largest inCitations from arXiv)

3. Visualize the trend of the amount of publications across all available years for venue ICSE.

4. Exploratory visualizations :
   (a) Design tasks (or find insights) to explore **author-co-author networks**. For example :
   - explore co-author information for any 'one' particular author based on the number of publications they have together for all available years. (Pick any author and consider all the co-authors in the publications by that author), and/or
   - explore co-author information for any 'one' particular author based on their Publication Affinity for all available years. (Pick any author and consider all the co-authors in the publications by that author. Publication Affinity between author X(author) and Y(co-author) = (number of papers published by X and Y together) / (total number of papers published by X) and/or
   - explore Topic affinity i.e. number of matching keywords (For each author-co-author pair, Topic Affinity between author X and Y = count of the subset of keywords in the publications by X and Y)

   (b) Design tasks (or find insights) to explore **Contribution index** of authors. For example
   - it could be based on number of years he/she has been active in publications eg year of latest publication - year of first publication.
   - or it could also mean if an author has published at multiple venues.
   - or it could be defined in terms of total number of publications and the duration in which these papers are published.
   - or it can be defined in terms of citations their papers have received.

Note:

a) You can use any types of visualization as long as you can achieve the above tasks. We value creativity.

b) You may need to do extra processing on data before visualization. You can use any tool to do extra processing. You don't have to report the extra processing script.

c) Before you start, take a look at the sample dataset provided in the link and get a sense of how the actual data looks like. Basically the two have the same format; the actual data is only bigger in size.

d) To help you get a better understanding on what you need to do, we provide a few visualizations, in Appendix I.

Report template

CS5346 Assignment 3 – CIRViz

Student Name		
Matriculation Number		

1. Introduction

(1-2 paragraphs including objective of assignment in your own words; individual contribution of each member in doing this assignment)

2. Visualizations - Purpose & Method

(i) State which visualization(s) did you select for each of the Task given in Section III. In order to facilitate grading, you can use a table showing which tasks are covered by which visualization. *An example is given below:*

Task	*Visualization*
1	*Heatmap*
2	*PieChart, Heatmap*
3	*Waterfall*

(ii) Provide an image of each of the visualizations you created. Label it to mention the task it addresses.

(iii) For **any two** of the visualizations *{atleast one of these should be from Task 4}*

Provide visual encoding . List Step wise method you followed in creating the visualization. Be precise and succinct. Write with a perspective such that your peers could easily use your method to create a similar visualization. Include any extra processing you did.

3. (optional) Any other comments or information you may have
