Mini-Project 2 – CIRViz

Submission Date: March 26, Monday *Report(by 10am)*
Demo (During the week of Mar 26, Schedule to be posted near to the time)

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

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 the mini-project Conference Information Retrieval (CIRViz) is to:

a) Extract the first 200,000 lines 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 200,000 lines 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). Note: however, you are **NOT** allowed to use visualization softwares for it (Microsoft Excel, Tableau, Power BI, etc). We want you to get hands on programming experiences and hence code your own plots.

II. Important Information

1. Read this document carefully.

2. If you have any query about this Assignment, send mail to cs5346.tutor@gmail.com.

3. Use first 200,000 lines of (full) dataset http://labs.semanticscholar.org/corpus/ for this assignment. Note that the dataset file is about 19GB in size. The extraction of data may take several mins. Also the 200,000 lines will require about a few GB of RAM for processing. Please contact us at cs5346.tutor@gmail.com in case you have any issues with extraction time or RAM size of your machine.

4. 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 200,000 lines of the FULL dataset. The link also gives short description of data attributes and an example.

5. Demonstrate visualizations, corresponding to the tasks set in Section III, **to your tutor in the week of March 26**(a schedule will be posted near to the time).

6. Submit a report (typically 2-5 pages, single pdf), **as per Report template given at the end of this document, by Monday, March 26, 10AM** in IVLE **folder MP-CIRViz-Report** in IVLE Files(workbin). Exceeding the page guideline of 2-5 pages does not invite any penalty.
Label the report document: **CIRViz_<Matric-number-1>_<Matric-number-2>**
e.g. CIRViz_A0045396X_A0046342Y.pdf

III. Task

Use **2 or more different types of visualizations to achieve the tasks given below**. **Each task should be covered by at least 1 visualization**.

There are in total **3 tasks:**

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**.

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 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 on next page.

--- Report Template ---

Mini-Project 2: CIRViz

Student Name		
Matriculation Number		

1. Introduction

(up to 1 paragraph 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 objectives given in Section III. In order to facilitate grading, you can use a table showing which objectives are covered by which visualization. *An example is given below:*

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

(ii) Provide an image of each of the visualizations you created

(iii) For **any one** of the visualizations:

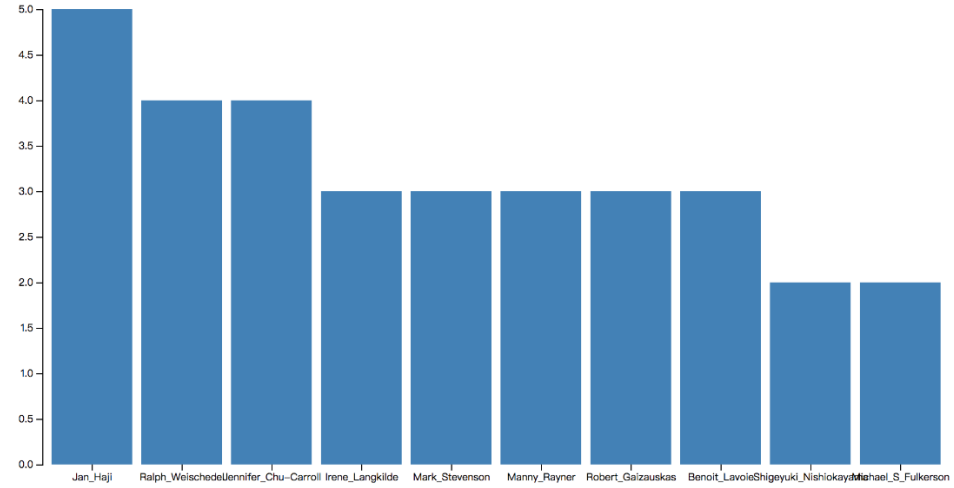
List Step wise method you followed in creating the visualization. Be precise and succinct. Write with a perspective such that your peers could easily use your method to create a similar visualization.

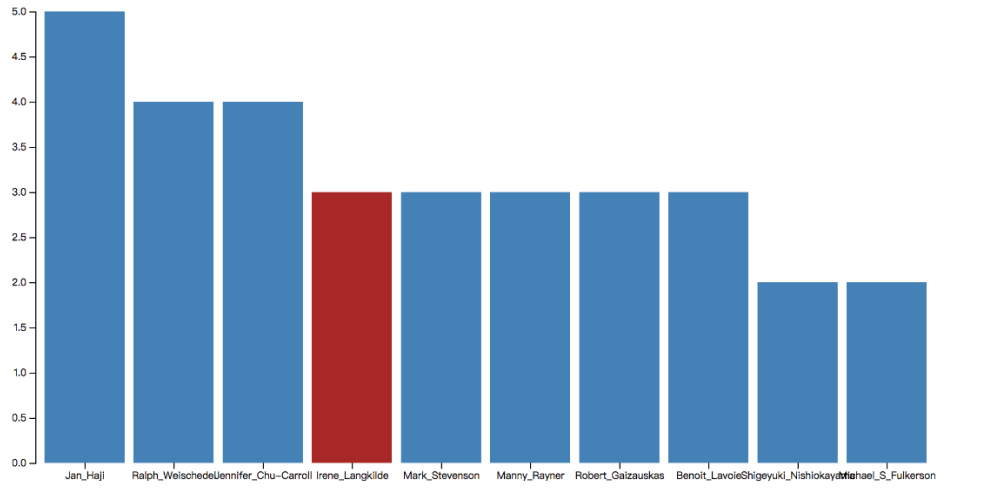
3. (optional) Any other comments or information you may have

Appendix I

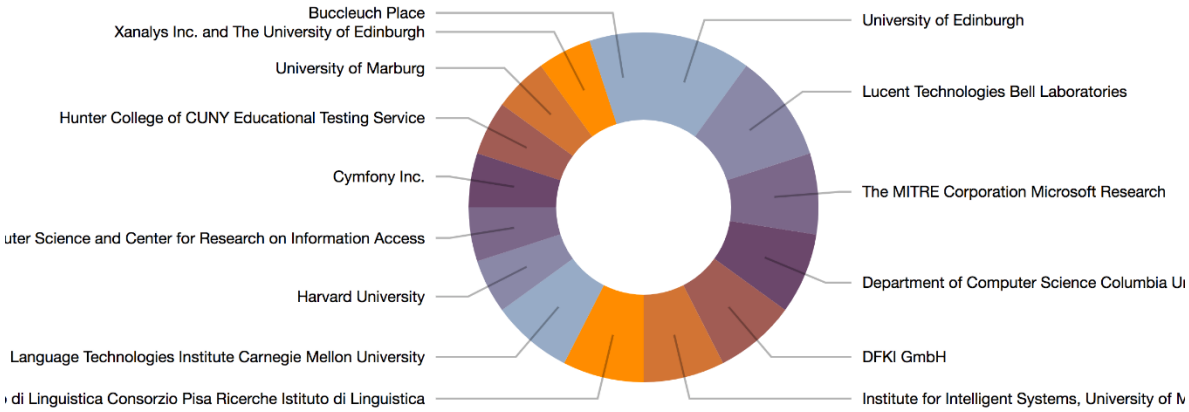
The graph/plots given below are based on the data from another dataset.

1. The top 10 authors:





2. The top affiliations/universities:



3. The trends of number of publications: