# **CS5346 Mini-Project 2 Conference Information Visualizations**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Dong Yizhou** | **Li Yuanda** |
| **Matric Number** | **A0078006J** | **A0078501J** |

# **Introduction**

According to research, number of science papers published had passed 50 million in 2009 with roughly 2.5 million new scientific papers published each year (Sarah Boon, 2016). Digging out meaningful information such as research topic trends, author ranks and article ranks can be a great challenge with such huge data source.

In this mini project, we use the first 200,000 lines from database of open research corpus with over 20 million published research papers in Computer Science, Neuroscience, and Biomedical as the data source to perform visualization and information digging. For this project we use both R and D3 to achieve the visualization. There are 2 main steps of our tasks:

1. Raw data preprocessing in preparation for the visualization
   1. Data format converting, to do the visualization in R, we managed to use the package rjson to convert the json object into dataframe before we proceed the visualization, while in D3 we proceed the visualization with json directly.
   2. Data cleaning, we noticed that there are quite a number of authors with no ids when we do the visualization for top authors. We dropped the rows with no author ids.
   3. Data aggregation, to calculate the count of authors and articles, we used the aggregation function in R.
2. Visualization with R and Javascript D3
   1. We use the library ggplot and plotly in R to perform the visualization. For Javascript we used D3.
   2. We want to do a cross comparison of the 2 programming language R and Javascript to understand the pros and cons of each language.

# **Visualizations**

|  |  |
| --- | --- |
| **Objective** | **Visualization** |
| **1** | **Barchart** |
| **2** | **Barchart, Bubble Chart** |
| **3** | **Line Chart, Donut Chart** |

**Objective 1 - Visualize the top 10 authors for venue arXiv**

**Visualization with R**

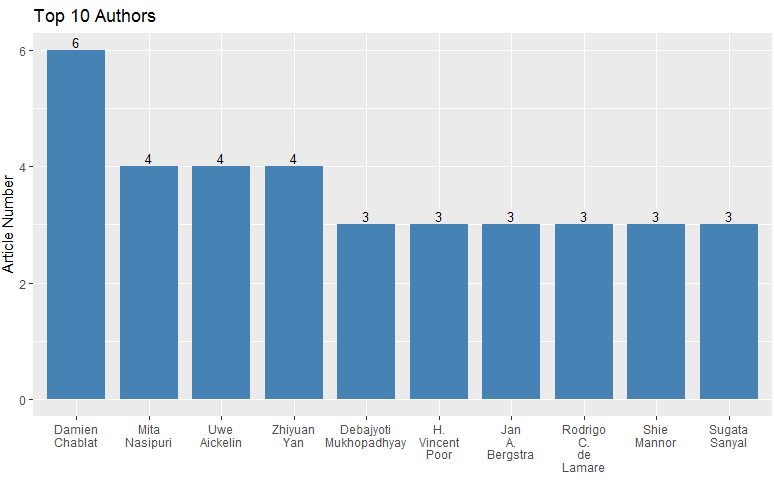
**Data Preparation:**

Library “rjson” was used to read in the json file and we retrieved the articles within venue arVix. Then we get the list of authors within the venue and dropped the authors with no ids. Since 1 author id will only be in 1 article once, so the total ids counts represent how many articles a particular author published. We get the id counts and match the id with the authors. After that, we use ggplot to plot the barchat with x axis as author names, y axis as article counts.



*R code for object 1*

Barchart was chosen as it can represent the number of articles and the rank of authors directly. Note that there are many authors published 3 articles our R code select the 6 authors with 3 articles in the barchart based on their ID rank. We can find author Damlen Chablat published the most number of articles for venue arVix.



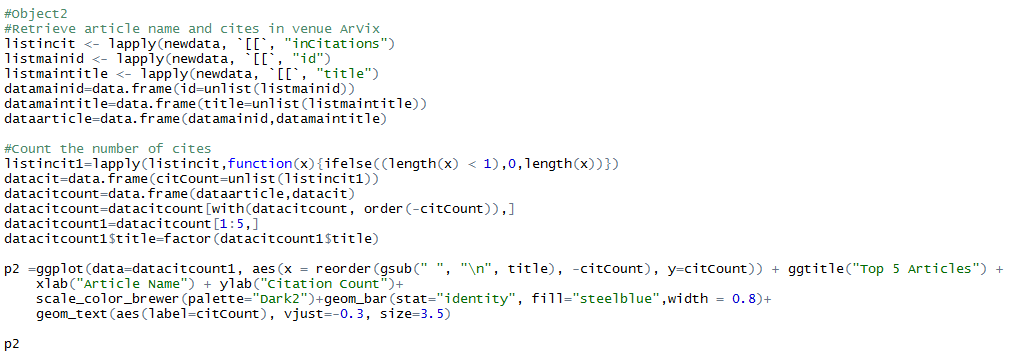
*R graph for object 1*

**Objective 2 - Visualize the top 5 papers for venue arXiv**

**Visualization with R**

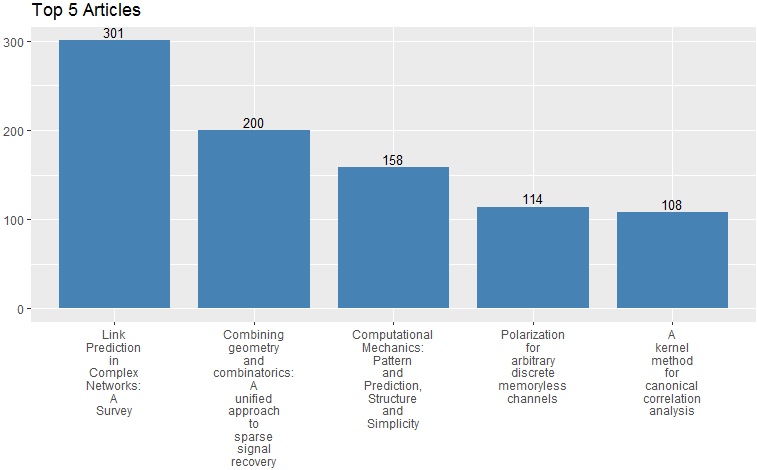
**Data Preparation:**

For object 2 R code, we retrieve the list of articled in venue arXiv and get the number of incitations for each article.



*R code for object 2*

Same as object 1 barchart was chosen as it can represent the number of citations and the rank of citations directly. We can tell the top article directly from the graph below and the top article in arVix is “Link Prediction in Complex Networks: A Survey”.



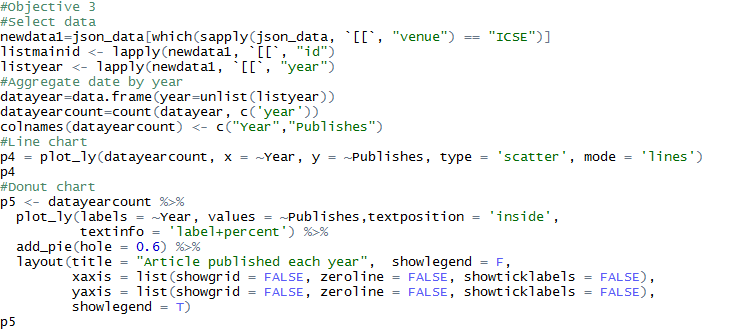
*R graph for object 2*

**Objective 3 - Visualize the trend of the amount of publications**

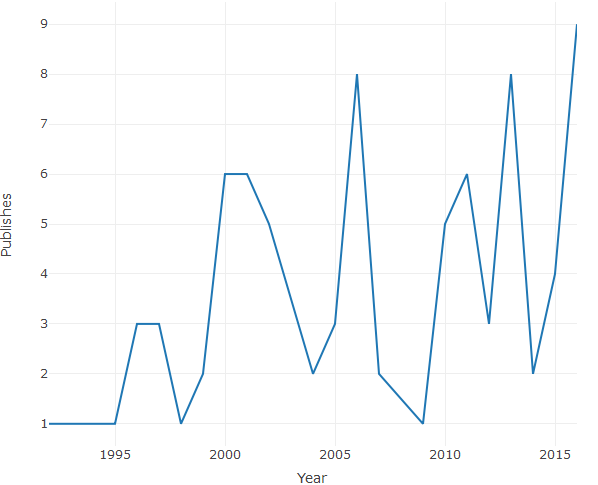
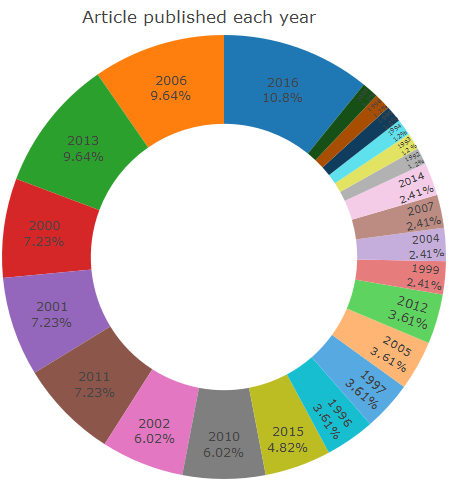
**Visualization with R**

**Data Preparation:**

We select all the articles in venue ICSE first. Then aggregate the data according to the years.



For objective 3 we created 2 graphs a Line Chart to visualize the general trend of article published each year and a Donut Chart to visualize the percentage contribution for the articles published each year. From the Line Chart we can see the article posted each year varies significantly and seems not following a time order. From the donut chart the percentage contribution is not in year sequence either. So there’s no significant trend for the article published.

*R graph for object 3*