Topics page:

The topics page provides information on a specific topic such as the most cited papers, most cited authors, and the most influential institution. There are two filtering options provided to select the topic, and then to narrow down the results to a specific period of interest.

At the start of the page is a table showing the most cited papers for the topic. This table contains the title of the paper, year of publication, the citation count, the authors of the paper, and lastly, a link to the url of the paper.

Following it, there is a scatter plot showing the citations of papers over the years. Each dot on this plot is a paper. This allows one to instantly gauge how interest in topic has evolved. Years with higher number of dots show that a greater number of papers have been published, and higher values show how much value the years are adding to the topic.

The pie chart on the page shows how much each sector has contributed to this topic (in terms of number of papers). The three sectors for this dataset are “industry”, “academia”, and “collaborative”. This gives a quick overview of how much the topic may be used in the industry. For example, the topic “Artificial Intelligence” has more papers by the academia than by industry. In contrast, topics such as “Computer Hardware” is more dominated by industry. Interestingly, the year filter here could help us visualize trends. For example, limiting “Computer Hardware” to the early years of 2006-2007, show that 80% of the papers come from academia, representing the time when it was still a major topic of interest in computer science.

A world choropleth map on the page shows the contribution for each topic by each country. The country is based on the publishing institution. This contribution is gauged by the number of citations for the papers published. Countries with a higher number of citations generally represent countries with a greater interest in a topic.

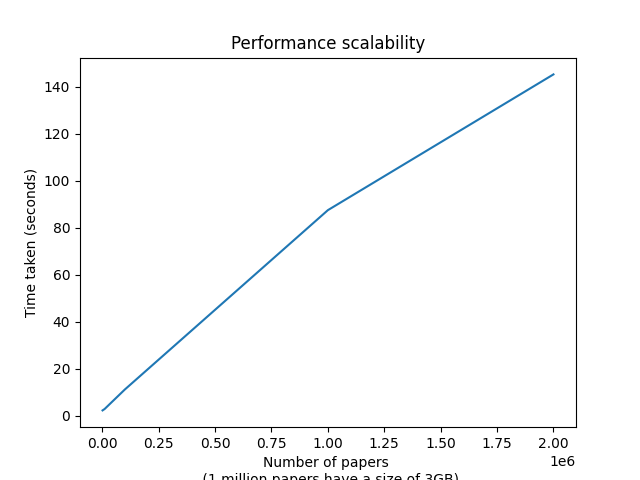
At the bottom of the page, visual cards display top authors and top institutions in the selected topic.

Data Preprocessing:

Data Preprocessing is where most of the knowledge of this course is applied. Since our original dataset is normally very large, and we only have a few rows of interest for any given topic in selected years, filtering the rows is the natural first step. This ensures that all subsequent operations such as aggregation and expansion of lists and dictionaries are performed only on the relatively small dataset. This data is then cached to be reused by all the plots. The plots use similar data sets with slight modifications, for example the choropleth map groups data by country and the pie chart groups data by type. For the choropleth, we also make use of gapminder’s country data to retrieve ISO alpha codes (insert citation <https://www.gapminder.org/data/> ).

Stress Testing / Scalability:

Scalability is the primary concern for most big data applications. While we demonstrate this dashboard on a sample of 50000 papers, extending it is not a major concern. Our tests show that the application scales linearly (refer to figure below). This is due to a good combination of memory utilization and code organization. The first step in all the graphs is to filter the rows and columns, and we do it once at the start and cache the results. This prevents huge costs in aggregation that normally come with increased data sizes. For example, consider the averages displayed on the authors page. These averages do not change with any option given in the dashboard, and hence are calculated only once at the start and then stored. This significantly improves performance as complicated calculations are performed once, and then a handful of key value pairs are stored. This allows the limited memory of the computer to be spent on storing useful data, instead of reading the entire data set every time.



The figure shows a linear increase in time taken to execute a set of plotting functions as we increase the number of papers. In the selected dataset, 1 million papers occupy approximately 1GB of space.