Foundations of Data Science Course Project Report

Farida Qadipasha - 400101731

February 9, 2025

1 Section 0: Crawling

Crawling a dataset to be used as the test dataset was completed in a previous phase, and the resulting dataset is stored in the `test.csv` file.

2 Section 1: EDA

2.1 1.1 Basic

2.1.1 1.1.1 Bar Charts of Publications in Year Ranges (1937–1950, 1950–1970, 1970–1990)

Three bar charts were created to show the number of publications in each of these year ranges.

- **Observation:** In more recent years, the number of publications has grown substantially.

2.1.2 1.1.2 Bar Chart of the Number of References Over the Years

A bar chart illustrating the total number of references per year was generated.

- **Observation:** The references increase over time, with a high peak around 2016.

2.1.3 1.1.3 Bar Chart of the Number of Authors Over the Years

A bar chart showing the number of authors per year was plotted.

- **Observation:** This trend is similar to the reference count chart, indicating a general increase in authorship over time.

2.1.4 1.1.4 Correlation Between Number of Authors and Number of References

Both the Pearson correlation coefficient (for linear relationships) and the Spearman rank correlation coefficient (for monotonic relationships) were computed.

- **Pearson:** Indicates almost no linear relationship between the number of authors and the number of references.
- **Spearman:** Suggests a very weak but slightly stronger monotonic relationship, implying that papers with more authors may *slightly* reference more papers, though the effect is minimal.

2.1.5 1.1.5 Correlation Between Number of Authors and Number of Citations

Both Pearson and Spearman correlations were computed.

- **Pearson:** Shows that increasing or decreasing the number of authors does not predict the number of citations in a linear manner.
- **Spearman:** Suggests papers with more authors might have slightly fewer citations, but the effect is so small it is practically negligible.

2.1.6 1.1.6 Bar Chart of Title Length Over the Years

A bar chart was drawn to analyze how the average paper title length changes by year.

- **Observation:** The title length appears to increase over time, except for some fluctuations in the earliest years.

2.1.7 1.1.7 Word Cloud of the Abstracts

A word cloud was generated based on the abstracts in the dataset.

- **Observation:** The resulting word cloud, shown in the notebook, highlights frequently used terms across all abstracts.

2.1.8 1.1.8 Correlation Between Title Length and Referenced Papers' Title Length

Pearson and Spearman correlations were computed between the title length of a paper and the average title length of the papers it cites.

- **Pearson:** Suggests that papers with longer titles tend to cite papers with longer titles (positive linear correlation).

- **Spearman:** Also indicates a relatively consistent monotonic relationship between citing and cited papers in terms of title length.

2.1.9 1.1.9–1.1.12 Top 10 Authors With the Most Publications, Citations, References, and In-Dataset Citations

The lists of top 10 authors by these criteria were identified.

- **Results:** Refer to the notebook for the detailed tables and rankings.

2.1.10 1.1.13 Predicting Total Citations from Number of Publications

To examine how well an author's number of publications predicts their total citations, correlation and linear regression analyses were performed.

Correlation Analysis.

- **Pearson Correlation:** 0.582 ($p\text{-}value = 0$). This suggests a moderate to strong positive linear relationship: authors who publish more papers generally receive more citations.
- **Spearman Correlation:** 0.602 ($p\text{-}value = 0$). This indicates a slightly stronger monotonic relationship, showing a consistent pattern in how authors with more publications also tend to have higher citation counts.
- **Statistical Significance:** Both results were highly significant (p-value = 0).

Linear Regression Analysis. A simple linear regression model was used to predict total citations from publication count:

- **Slope:** 43.534, suggesting each additional publication corresponds to approximately 43 more citations on average.
- **Intercept:** -26.562, which is not meaningful in a real-world context (no one with zero publications would have negative citations), but included in the model structure.
- **R-squared:** 0.339, indicating that publication count explains about 33.9% of the variance in total citations.

Interpretation and Conclusion.

- The moderate to strong correlation (Pearson: 0.582, Spearman: 0.602) confirms publication count is a meaningful predictor of citation count.
- However, the linear regression shows that other factors contribute significantly, given an R^2 of 0.339. Journal impact, collaboration networks, and paper quality also influence citation rates.

2.2 1.2 Network Analysis

2.2.1 1.2.1 Citation Network (Paper–Paper Network)

Clustering Coefficient Over Time. The clustering coefficient was plotted over time to measure the interconnectedness of the citation network.

- **Observations (1950s–2010):** The coefficient remains close to zero, indicating sparse interconnections and a tree-like structure.
- **Around 2010:** A sharp increase suggests a more interconnected network, reflecting the rise in collaborative research and the formation of denser citation clusters.

Average Path Length and Diameter.

- **Average Path Length:** 2.11, indicating that any two papers in the largest connected component are typically just over two citation steps apart.
- **Diameter:** 4, meaning no paper is more than four steps away from any other within this component.
- **Conclusion:** The citation network is well-connected and compact, characteristic of a small-world network.

Influential Papers (PageRank). Papers with higher PageRank values are considered more influential in the network. Refer to the notebook for detailed results.

2.2.2 1.2.2 Co-Authorship Network (Author–Author Network)

Network Density Per Year. Network density (the ratio of actual to possible connections) was computed by year.

- **Pre-1980:** Density can spike to 1.0 due to small, fully interconnected author groups.
- **Post-1980:** Density decreases and stabilizes as more authors and collaborations diversify.

Influential Researchers (Centrality Measures). Central authors were identified using:

- **Degree Centrality:** Number of direct co-author connections.
- **Betweenness Centrality:** Importance as a bridge in the network.
- **Closeness Centrality:** Efficiency in reaching other authors.

Detailed results are available in the notebook.

Communities of Authors. Communities representing closely collaborating or thematically similar author groups were detected. Refer to the notebook for community structures.

2.2.3 1.2.3 Venue Network (Conference–Journal Network)

Interdisciplinary Collaborations Between Venues. Measures such as betweenness centrality were used to detect top interdisciplinary venues. Additionally, community detection was applied to identify venues that appear in multiple smaller clusters. Note that sampling issues might affect result quality.

Most Influential Venues. Centrality metrics (e.g., PageRank, betweenness) identified key journals and conferences in the dataset.

Emerging Fields. Newly formed connections between venues can indicate emerging research areas. Conclusions are shown in the notebook, though partial dataset coverage may limit reliability.

2.2.4 1.2.4 Temporal Evolution of the Citation Network

Network Density Over the Years. Citation network density was analyzed annually:

- **Before 1990:** Density is nearly zero, implying sparse connections.
- **Late 1980s–Early 1990s:** Sharp increase due to rapid citation growth.
- **Post-1990:** Density declines steadily as the number of papers (nodes) grows faster than the citation links (edges).

Burst Detection of Influential Papers. Certain papers experience sudden citation spikes. Detailed results are provided in the notebook.

Integration of New Papers into the Network. The average in-degree (citations received) for new papers was plotted:

- **Before 1990:** Nearly zero in-degree for new papers.
- **After 1990:** Sharp increase in how quickly new research is cited, reflecting enhanced dissemination (e.g., digital libraries).

3 Section 2: Data Extrapolation via Clustering

3.1 2.1 Community Detection

Multiple algorithms were tested, and they produced the same Average Intra-Cluster Clustering Coefficient (up to 15 decimal places). The Louvain algorithm was selected for the full graph. Results are shown in the notebook.

3.2 2.2 Naming the Communities

An aggregation method was used to assign labels to communities by calculating keyword frequencies within each community. The final labels and detailed outputs can be found in the notebook.

3.3 2.3 Paper–Paper Clustering via Embedding

Clustering quality was evaluated using the Davies–Bouldin Index (DBI) and the Silhouette Score.

- **DBI:** 4.20 (lower is better; a high DBI suggests significant overlap among clusters).
- **Silhouette Score:** 0.0316 (close to 0 indicates poor separation).

These metrics suggest poorly defined, overlapping clusters. Potential reasons include high-dimensional embeddings, noisy data, or insufficient sample size. Re-evaluation with more optimized embeddings or different clustering methods could improve performance.

The limited (seubset) dataset size may have contributed to the poor performance metrics.

The additional questions related to this section are addressed within the notebook.

4 Section 3: Citation Regressor

Evaluation on the Test Dataset

Metrics:

- **RMSE:** 325.95
- **MAE:** 61.32
- **R^2:** -0.0339

These values suggest large prediction errors and indicate that the regressor performs worse than a simple mean-based baseline (as evidenced by the negative R^2). Several factors contribute to this outcome:

- A *very small subset* of the dataset was used for testing, limiting generalizability.
- Citation counts are heavily influenced by external factors not included in the model, such as journal quality, disciplinary trends, or co-authorship networks.

Future work could involve using larger training sets, incorporating additional features, or experimenting with more advanced embedding techniques.

The additional questions related to this section are addressed within the notebook.

5 Section 4: Product

No special report was produced for this section.