**ÖZYEĞİN UNIVERSITY**

**CS 452/552 TERM PROJECT**

Extracting and Analyzing Authors in Turkish Universities using Google Scholar

Ege ORUC

Gamze ADIBELLI

**Abstract**

This project systematically collects, organizes, and analyzes data about authors affiliated with Turkish universities using Google Scholar. The study addresses challenges such as naming inconsistencies and the limitations of Google Scholar’s anti-scraping mechanisms. By doing so, it provides valuable insights into academic demographics, gender distribution, geographic representation, and departmental trends. Future research will explore integrating additional data sources and expanding the analysis to broader contexts.

**Introduction**

The academic landscape is a crucial indicator of societal progress and innovation. Understanding the demographics and contributions of researchers within Turkish universities reveals important patterns about research diversity and institutional strengths. However, obtaining accurate and comprehensive author data poses significant challenges. Data sources are often scattered, author names appear in inconsistent formats, and anti-scraping measures by Google Scholar hinder automation. This project aims to create a scalable framework for collecting and analyzing this data, focusing on demographic trends and institutional representation.

**Methodology**

The primary data sources for this project include the YOK website, which provides university details, Google Scholar for author profiles and metrics, and a gender identification dataset used for demographic analysis. Various Python libraries such as Selenium, pandas, and matplotlib were employed to automate searches, process data, and generate visualizations. The methodology involved extracting university names from the YOK website, automating searches on Google Scholar for author data, cleaning and standardizing the extracted data, and conducting statistical analyses to uncover meaningful insights.

Data cleaning was a critical step in this process. Names were standardized to address variations such as initials, abbreviations, and alternate spellings. Duplicate records were removed, and incomplete profiles were flagged for manual review. Wherever possible, affiliations were validated against official university data to ensure accuracy.

**Challenges**

**Data Collection**

One of the major obstacles was the restrictions imposed by Google Scholar, including CAPTCHA interruptions that slowed down the automated scraping process. Name variations further complicated data collection, as authors’ names appeared in inconsistent formats, leading to potential mismatches. Additionally, incomplete author profiles required significant manual intervention to fill data gaps and ensure comprehensive analysis.

**Technical and Analytical Difficulties**

Frequent CAPTCHA interruptions limited the scalability of the automation framework. Attempting to collect data for all universities in Turkey was time-intensive, forcing the project to focus on a subset of institutions. Analytical challenges included determining gender from external datasets, which introduced potential inaccuracies, and standardizing inconsistent departmental titles across universities to ensure uniformity in analysis.

**Results and Analysis**

An analysis of state and foundation universities ("devlet" and "vakıf" universities) further highlighted key differences in academic representation and output. State universities collectively produced 2,810,211 total citations and hosted 5,021 authors, demonstrating their dominance in academic representation. Foundation universities, while having fewer authors (2,480) and total citations (2,484,379), exhibited competitive h-index (8.69) and i10-index (12.78) averages, particularly in metropolitan areas like Istanbul and Ankara. These metrics reflect the robust research output of specific foundation universities despite having fewer resources. This trend is attributable to state universities’ longer histories and access to public funding, which facilitates research activities and infrastructure development. However, some foundation universities, particularly those established in metropolitan areas like Istanbul and Ankara, showed competitive metrics, with higher-than-average citations per author in certain fields.

A graph of different colored squares

Description automatically generated with medium confidence

The study uncovered significant patterns in academic demographics. Major cities such as Istanbul, and Antalya dominated the geographic representation, hosting over 60% of authors collectively. The majority of authors were affiliated with Computer Science and Engineering departments, which emerged as the most represented fields, reflecting their prominence in Turkish academia.

A graph of different colored bars

Description automatically generated

Cities, being metropolitan hubs, offer better access to research infrastructure and funding, resulting in higher academic activity. In contrast, authors from smaller or rural universities were underrepresented, indicating regional disparities in access to academic resources.

In terms of institutional representation, established universities such as Istanbul University and Middle East Technical University consistently led the rankings. These institutions had not only the highest number of affiliated authors but also demonstrated stronger research output, as evidenced by higher average citations and h-index values. Newer universities struggled with lower representation and fewer citations, highlighting the disparity in research capabilities and funding between older and newer institutions. A **university-level representation bar chart** could be included here to detail institutional differences.

A graph of different colored bars

Description automatically generated

A deeper analysis of publication trends across year ranges revealed interesting insights. Between 2010 and 2015, publication growth was steady but showed a significant uptick between 2016 and 2021. This trend highlights the increasing academic output in recent years, possibly driven by expanding research opportunities and funding. The rise was particularly notable in fields like Computer Science and Engineering, with some growth observed in interdisciplinary research areas such as Artificial Intelligence and Data Science.

A graph with blue line and orange line

Description automatically generated

The data also revealed that STEM fields overwhelmingly dominated publication counts, with Computer Science and Engineering leading the way. The representation of authors in Humanities and Social Sciences, though lower in comparison, showed a steady increase over time, suggesting a growing interest and contribution in these areas. A **departmental trends bar chart** is recommended here to illustrate this distribution effectively.

**Summary Table**

The following table summarizes key metrics from the analysis:

| **Metric** | **Value** |
| --- | --- |
| Total Authors | 32,649 |
| Unique Universities | 44 |
| Unique Cities | 19 |
| Top Department | Computer Science |
| Top City | Antalya |



**Discussion**

This project highlights the need for increased gender diversity in Turkish academia, particularly in STEM disciplines. The concentration of academic activity in metropolitan areas reflects disparities in resource allocation and access. Established universities dominate in terms of both representation and output, underscoring the need for equitable funding and support for emerging institutions.

Policymakers can leverage these findings to address geographic and institutional disparities. Universities can focus on fostering diversity and inclusion while enhancing support for underrepresented fields such as Humanities and Social Sciences. However, limitations such as the incomplete coverage of Google Scholar and occasional mismatches in author data suggest the need for additional data sources and improved algorithms for data matching.

**Conclusion and Future Work**

This project successfully developed a scalable pipeline for extracting and analyzing academic data, uncovering trends in gender diversity, departmental representation, and geographic disparities. Future efforts will focus on integrating broader data sources such as Scopus and Web of Science, developing advanced algorithms to resolve name ambiguities, and expanding the analysis to include more universities and regions. Collaborations with academic institutions can further validate and enrich the dataset, enhancing its applicability and impact.