



Build a Dataset of ResearchGate Members

Z. Mundher^{(1)*} , Y. Ali⁽²⁾ 

^(1,2) Department of Computer Science, College of computer science and mathematics, Mosul University, Mosul, Iraq

Article information

Article history:

Received: October 05, 2024

Accepted: November 26, 2024

Available online: January 01, 2025

Keywords:

Web Scraping

ResearchGate, Exploratory Data

Analysis (EDA)

K-means

ANOVA

Correspondence:

Zaid Mundher

zaidabdulah@uomosul.edu.iq

Abstract

In today's academic environment, social media platforms, ResearchGate in particular, play an essential role in facilitating collaboration and communication among researchers. Even with the wide usage of this site, there remains a significant gap in the availability of structured datasets that focus on researchers and their academic outputs. This lack of accessible data obstructs comprehensive analysis and evaluation of research trends and impact. This study seeks to address this gap by employing web scraping techniques to construct a dataset derived from ResearchGate, a leading platform for academic professionals. The introduced dataset consists of eight key features, including metrics related to publications, citations, and areas of research specialization. The availability of such a dataset not only provides a valuable resource for future research but also enables scholars to analyze research performance, identify collaboration opportunities, and uncover trends in academic productivity. Beyond dataset construction, this paper also details the exploratory data analysis (EDA) conducted to derive insights from the collected data. The K-means clustering algorithm was applied to categorize researchers according to their publication and citation patterns, offering a clearer understanding of academic achievements. The main contribution of this work is the establishment of a researcher's dataset that can be used in future studies to analyze and examine the efforts and trends of researchers in their scientific work.

DOI: [10.33899/edusj.2024.154161.1506](https://doi.org/10.33899/edusj.2024.154161.1506), ©Authors, 2025, College of Education for Pure Science, University of Mosul.

This is an open access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The large amount of data available today is a real treasure which has led to a new era of ideas in academic research, science and business. It is the age of data which can be obtained from different sources in the form of datasets. Simply, a dataset is a collection of data organized in rows and columns (tables). The dataset can be considered as the initial part of any data science, machine learning, and deep learning project. Data also plays an important role in the AI revolution, that we all experiencing today. ChatGPT, and other AI projects, are trained by using large datasets to provide the best results. Therefore, the effectiveness of any AI model, depends significantly on the availability of the datasets, meaning that the lack of data is a serious issue that the data scientists may face[14][11].

Mainly, there are two kinds of data: structured and unstructured. Traditionally, structured data is represented in tables, making it ready to use. However, with the widespread use of the Internet and social media, another form of data has emerged, known as unstructured data, that can be manually collected from various sources such as files, images, and web pages. Effort is needed to collect unstructured data by inspecting web pages and going through a specific automated process. The process of collecting data from various web pages and arranging it into files or tables is known as web scraping (also known as web data extraction) [5][6]. However, gathering data from web pages is not a new practice. Internet users are accustomed to using traditional "copy & paste" methods to manually obtain data. As technology develops and new concepts emerge, web scraping has become an automated process that can be achieved programmatically, providing more accurate results [12].

1.1 WEB SCRAPING

Scraping is gathering, so scraping the web means gathering data from the web. Web scraping is an automated process that saves effort and time by using scripts (programs) to gather data from websites [5]. Healthcare, finance, social media, and marketing are examples of domains that may use web scraping methods to improve their services. Python is one of the most popular programming languages for web scraping. Different libraries and techniques can be used to handle web scraping tasks. BeautifulSoup and Selenium are two very common libraries used for web scraping [3][9].

1.2 RESEARCHGATE

Our research team was tasked with analyzing data related to academic researchers worldwide. A literature review was made, and no dataset has been found containing researchers' information. Therefore, the initial task for our research team was to build the dataset from scratch. Consequently, web scraping techniques were used to gather the necessary data by scraping one of the most popular social media platforms for researchers: ResearchGate.

ResearchGate is a research and scientific social media platform, which has become a very important community of researchers' daily life. It is one of the most important websites that share researchers' information including names, institutes, papers, skills, and citations. ResearchGate's website allows researchers and scientists to create personal accounts containing their personal and scientific information [5][10]. Figure 1 illustrates a typical ResearchGate account with information such as name, institute, academic credentials, and citations.

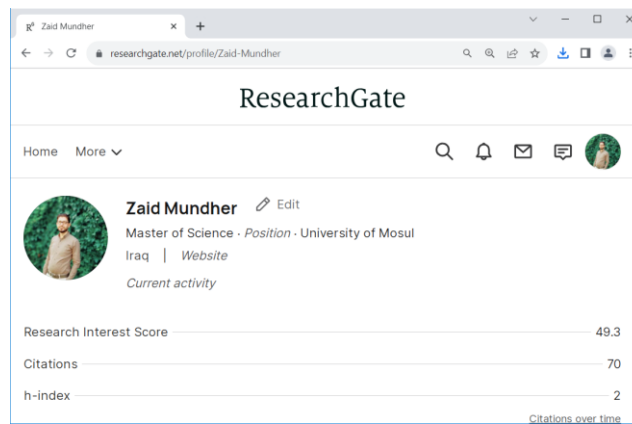


Figure (1): ResearchGate - a user account

Using the options available on the ResearchGate website, researchers can be clustered based on their affiliated institutes. For instance, there are approximately 32,500 members affiliated with Columbia University as illustrated in Figure 2 and Figure 3.

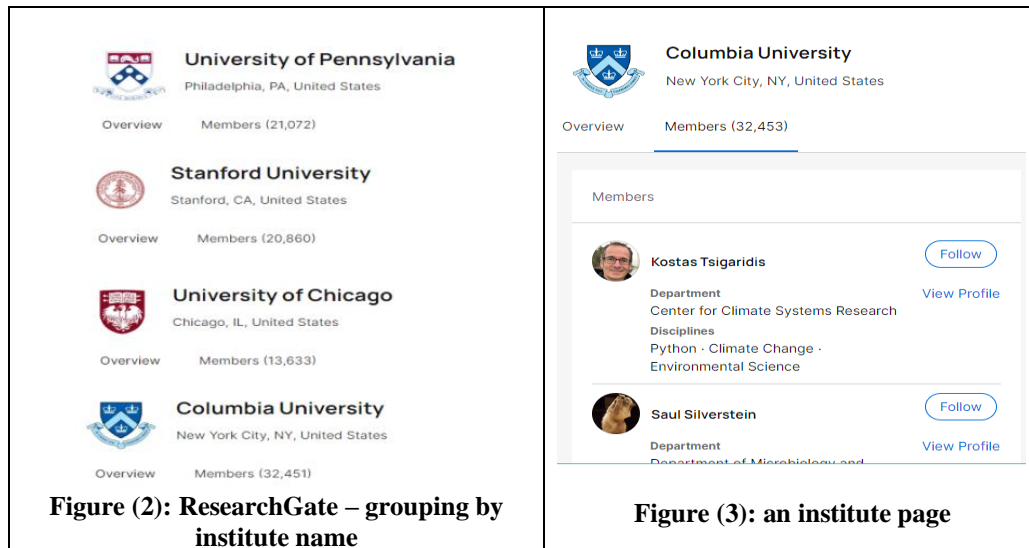


Figure (2): ResearchGate – grouping by institute name

Figure (3): an institute page

2. RELATED WORKS

Web scraping techniques have been widely adopted in diverse research projects. In [8], web scraping was applied in the e-commerce field to gather data on products and identify their updates. The study focused on the Tori.fi online shopping website as a case study. Authors of [12] utilized web scraping methods to scrape e-commerce websites for price comparison, aiming to assist customers in finding the best online deals automatically, thereby, saving time and effort. Finding academic journals can be challenging and time-consuming for researchers. Therefore, authors in [2] addressed this issue by introducing a solution to help researchers publish their papers in academic journals based on web scraping methods. SINTA, which is an Indonesian website for researchers and scientists, was used as a case study. In [4], a web scraping tool was introduced to gather geographic data, while [6] focused on gathering information from Instagram user accounts using web scraping techniques and storing the collected account information in a CSV file. On the other side, papers, such as [1][13] addressed issues related to the ResearchGate website. Analyzing the academic work of Iranian women on the ResearchGate platform was introduced in [1]. In [13], the most cited Indian researchers were identified based on the ResearchGate cited score. To the best of our knowledge, there is no previous work has introduced a scraper to gather researchers' information from ResearchGate. In this research, web-scraping techniques are used to scrape the ResearchGate website to create a dataset that contains information on researchers of a specific institute. Researchers' information was collected, classified, and stored in a local CSV file.

3. METHODOLOGY

In this work, BeautifulSoup—a widely used library for parsing HTML—is adopted to efficiently extract data from static HTML pages, aligning with the requirements of this research. According to [5], BeautifulSoup demonstrates superior results compared to other scraping methods, such as Selenium, in handling static content. This choice makes it particularly suitable for parsing structured data from ResearchGate profiles, given its stable HTML layout.

The implementation of the proposed work can be divided into three phases:

- Analysis phase
- Extraction phase
- Storing phase

3.1 ANALYSIS PHASE

Web scraping requires an initial and thorough examination of the HTML structure of the target site, as each website features unique layouts and tags. The primary goal of the Analysis Phase is to identify the relevant HTML elements that contain the target information. To accomplish this, the structure of ResearchGate's HTML was manually inspected, focusing on the profile pages of researchers affiliated with the University of Mosul (UoM). Using Chrome's "Inspect" tool, specific HTML tags, CSS selectors, and attributes associated with researcher profiles (such as class names and ID tags) were identified. This detailed inspection revealed the specific tags and container housing elements such as the researcher's name, department, and publication count. Figures 4 and 5 illustrate this process, while Figure 6 shows a sample of the extracted HTML code, marking the tags and properties that guide the parsing logic for the next phase.

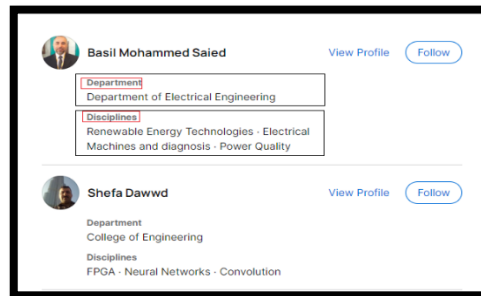


Figure (4): UoM institute

Figure 5 demonstrates the manual inspection of a webpage using Chrome web browser options.

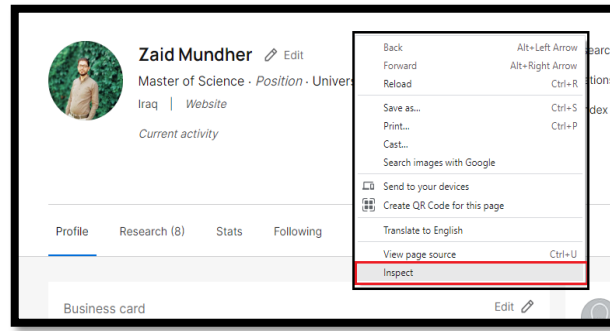


Figure (5): Inspect a web page manually

A sample of HTML code from the previous step is shown in Figure (6). IDs and class names of CSS properties were identified and memorized for use in the scraper code.



Figure (6): A sample HTML code

3.2 EXTRACTION PHASE

The Extraction Phase involves utilizing BeautifulSoup to systematically retrieve the targeted data fields. Researcher profiles on ResearchGate were collected based on their association with the chosen institute (UoM). Each profile page was parsed to gather the specific attributes listed in Table 1:

Table 1. - extracted info.

Index	Name
1	Name of the researcher
2	The link to the profile page
3	Profile photo
4	Department name
5	Number of publications
6	Number of citations
7	Disciplines
8	Skills

The extraction process required creating a parser with BeautifulSoup that targets these attributes based on the HTML analysis from the previous phase. Once the relevant tags were identified, the BeautifulSoup library was used to access these elements and extract data accurately. The following points outline the sequence of steps taken during scraping:

1. **Initialize URL** – Start with the UoM researchers' directory on ResearchGate.
2. **Scrape Initial Page** – Retrieve and parse all individual researcher profile links.
3. **Visit Profiles** – Access each profile URL.
4. **Extract Data** – For each profile, gather the researcher's name, profile photo, department, and other details as listed in Table 1.
5. **Repeat for All Profiles** – Continue for all researcher profiles associated with UoM.

3.3 STORING PHASE

Once the data is gathered, it is essential to save it in a structured format suitable for future analysis. In this study, data is stored as a Comma-Separated Values (CSV) file, a widely accepted format that supports efficient data manipulation and is compatible with numerous analysis tools.

After this phase, the created dataset is locally available, providing a valuable resource for exploratory analysis and machine learning tasks, including clustering researchers based on publication counts and citation metrics.

4. RESULT AND DISCUSSION

By implementing the proposed work, the researcher's data was successfully extracted and stored in a local CSV file, resulting in a dataset with 8 columns and 6806 rows. Figure 7 shows the dataset's dimensions (rows and columns) and the names of each column.

```
data.shape
(6806, 8)

data.columns
Index(['Name', 'link', 'profile_photo', 'department', 'Publications',
      'citation', 'desciplines', 'skill'],
      dtype='object')
```

Figure (7): created dataset – size and column names

A sample of the gathered dataset is presented in Figure 8.

	Name	link	profile_photo	department	Publications	citation	desciplines	skill
0	Basil Mohammed Saied	https://www.researchgate.net/profile/Basil-Saied	https://i1.rgstatic.net/ii/profile/image/11431...	Department of Electrical Engineering	107.0	237.0	Renewable Energy Technologies, Electrical Machi...	Renewable Energy Technologies, Electrical Machi...
1	Shefa Davvid	https://www.researchgate.net/profile/Shefa-Davvid	https://i1.rgstatic.net/ii/profile/image/86001...	College of Engineering	55.0	306.0	FPGA, Neural Networks, Convolution,	55 FPGA, Neural Networks, Convolution, Parallel F...
2	Theia'a Al-Sabha	https://www.researchgate.net/profile/Theiaa-Al-...	https://i1.rgstatic.net/ii/profile/image/27217...	Department of Chemistry (College of Education ...)	65.0	237.0	Spectrometry, Chromatography, Liquid Chroma...	65.3 Mass Spectrometry, Chromatography, Liquid C...
3	Dr. Marwan Muhib	https://www.researchgate.net/profile/Dmrvan-...	https://i1.rgstatic.net/ii/profile/image/74401...	College of Petroleum and Mining Engineering	46.0	8.0	Applied Geophysics, Exploration Geophysics, Stat...	46.6, Applied Geophysics, Exploration Geophysics...
4	Ziad Al Sarraf	https://www.researchgate.net/profile/Ziad-Al-S...	https://i1.rgstatic.net/ii/profile/image/70815...	College of Engineering	31.0	66.0	Design Engineering, CAD, Mechanical Properties,	31.9, Design Engineering, CAD, Mechanical Propert...

Figure (8): Top Five Rows of the dataset

The time taken to retrieve information for each researcher was approximately eight seconds, including a five-second delay that was deliberately added between requests to avoid IP blocking during data retrieval.

To gain basic insights about the data, some analysis was conducted. Firstly, as a pre-processing step, missing values were deleted to ensure that the dataset did not contain any null value. As a result, the final data set size is 2908 rows. EDA,

which is an essential step in understanding any dataset, was implemented to derive the initial insights and deep understanding of the data. The first question we aimed to answer was the number of researchers in each department. Figure 9 shows the top 10 departments by the number of researchers.

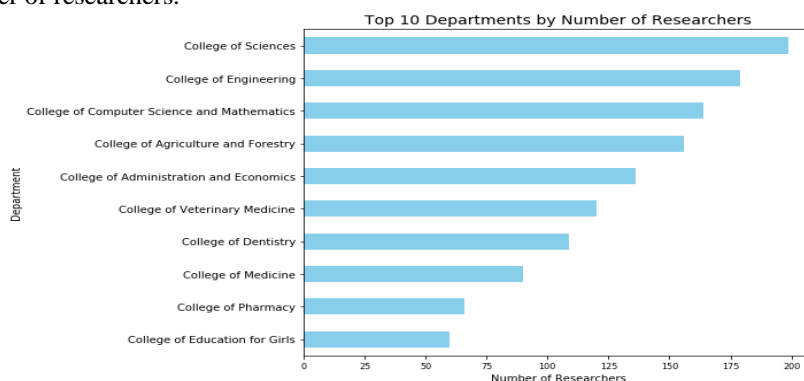
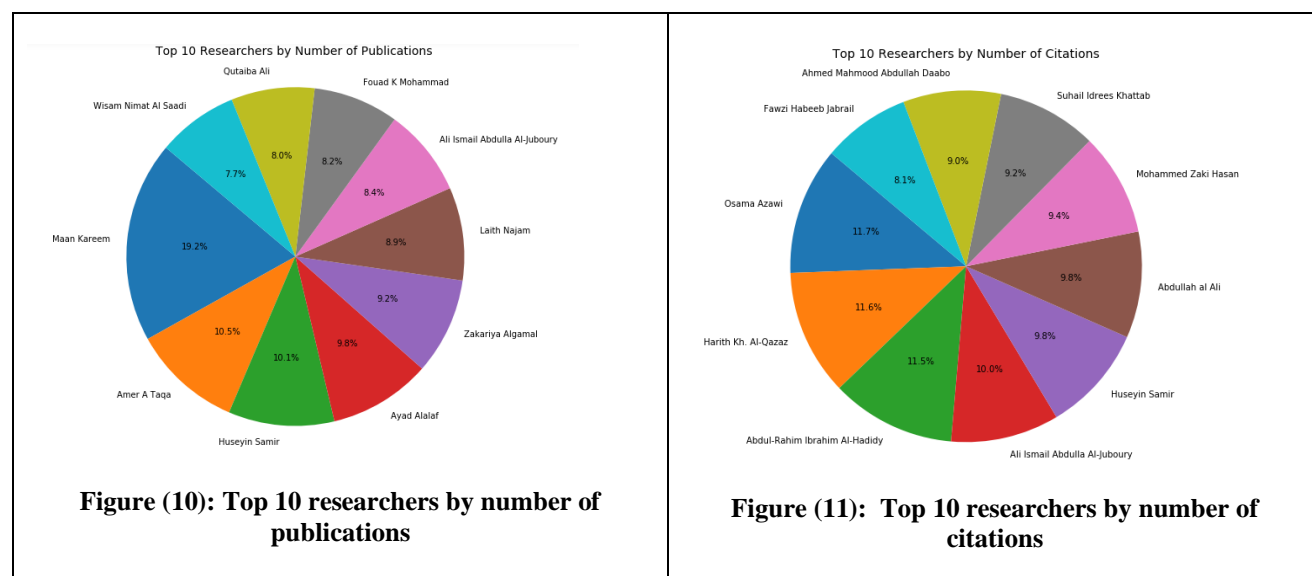
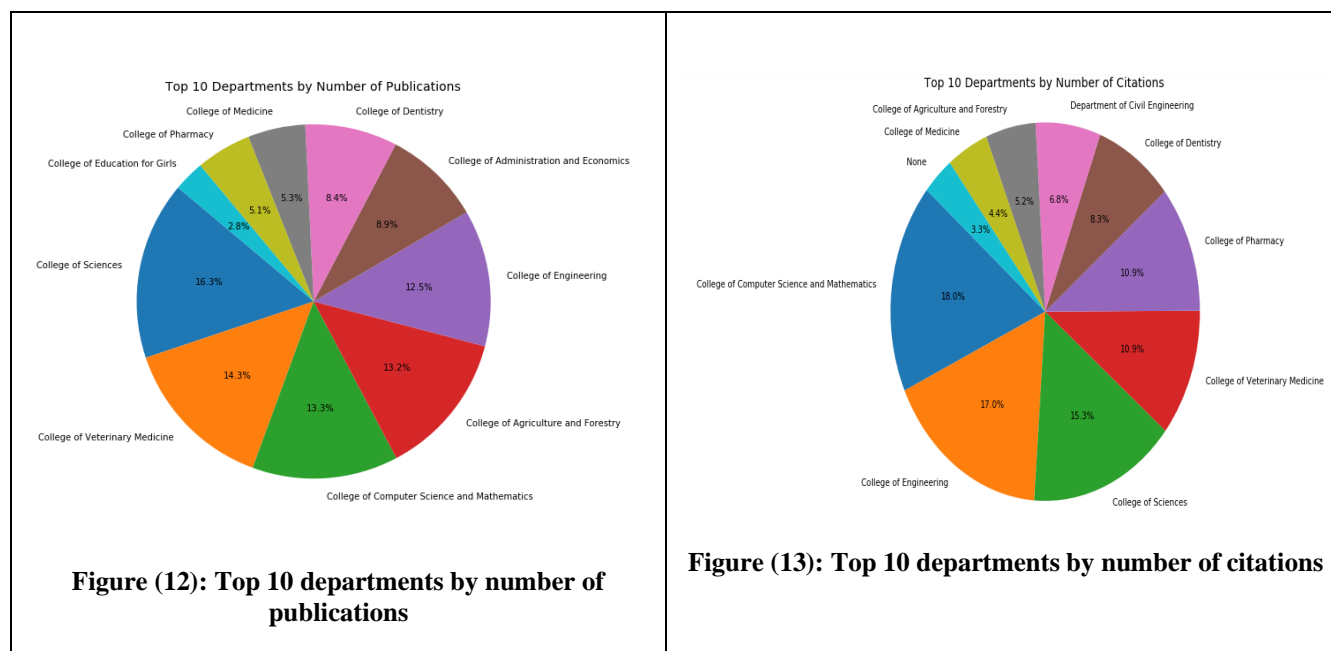


Figure (9): Top 10 Departments by Number of Researchers Five Rows of the dataset

The top 10 researchers by number of publications and the top 10 researchers by number of citations were also found. Figure 10 and Figure 11 show the results we obtained.



In addition, Figure 12 shows the top 10 departments by number of publications, while Figure 13 shows the top 10 departments by number of citations.



Moreover, the correlation between Publications and Citations was also calculated and figured in Figure 14. As noted, the correlation ratio was 0.464 which is an average value indicating a logical and reasonable positive correlation between a number of publications and a number of citations. Meaning that the more research is published by a particular researcher, the more citations there are on average.

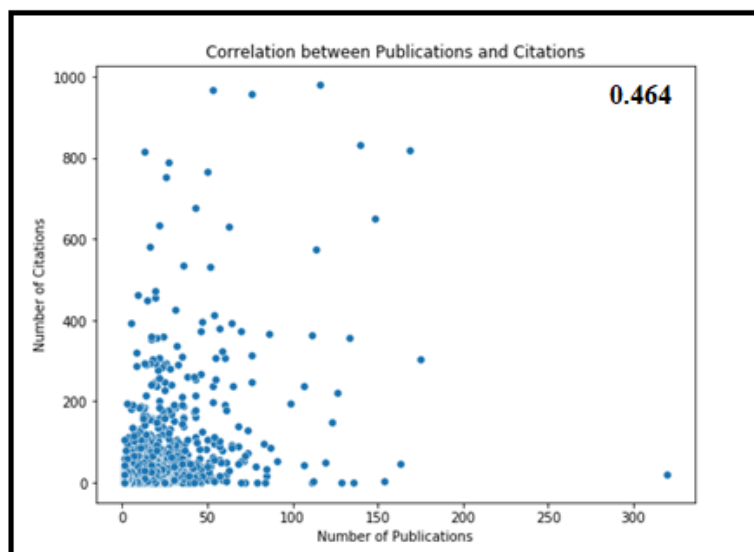


Figure (14): The Correlation between Publications and Citations

Researchers were also divided into groups based on the number of their publications and citations. Table 2 and Figure 15 show the three categories.

Table 2. – Researchers Groups

No. of Citation	Category	Result
Less or equal 20	Low	2259
Less or equal 50	Medium	329
More than 50	High	320

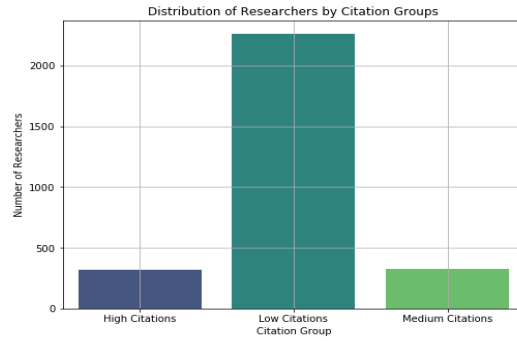


Figure (15): distribution of researchers by citation group

We also determined the efficiency of researchers. Obviously, publications can be considered as the measure of the productivity of a researcher. On the other hand, citations are the main factor to measure the quality of a researcher. Thus, with the above two points, the efficiency of any researcher can be calculated by the following:

$$\text{researcher_efficiency} = \text{res_citation} / \text{res_publication}$$

Where res_citation of a particular researcher refers to the number of citations her research has received, and the res_publication refers to the number of research she has published. Figure 16 displays the top 10 most efficient researchers in our dataset.

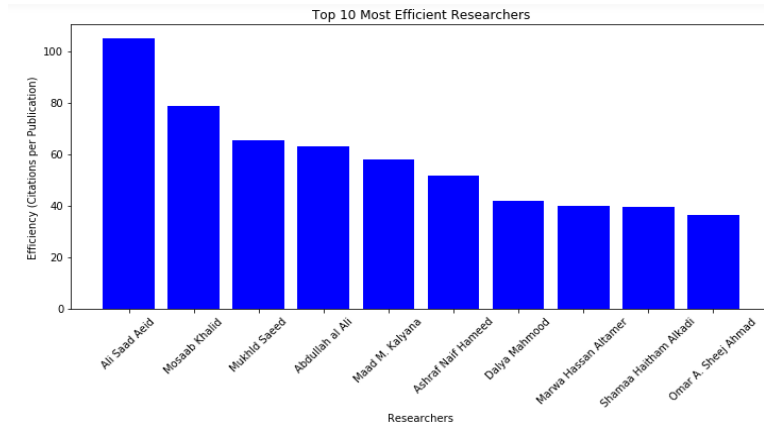


Figure (16): top 10 most efficient researchers

Finally, to understand the data more deeply and gain a better insight into researchers, the K-means algorithm was used to divide the researchers into three groups based on their publications and citations. Figure 17 illustrates the K-means clustering with three clusters.

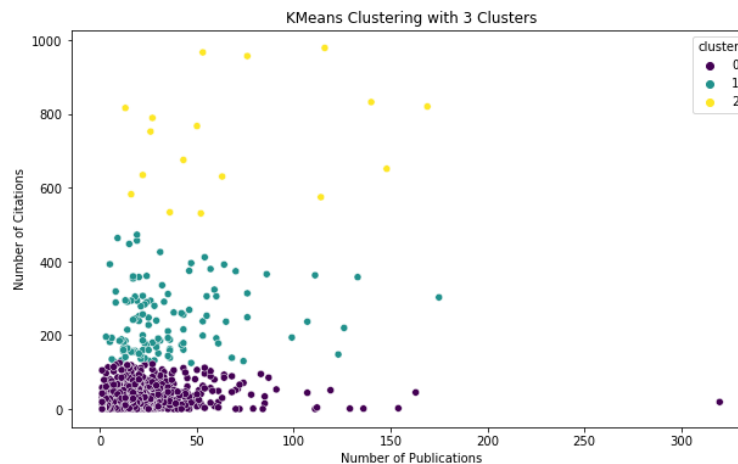


Figure (17): k-means clustering with three clusters

Statistically, the ANOVA (Analysis of Variance) method was used to compare the mean values of publication and citation across the three clusters. The ANOVA results for Publications are shown in Figure 18, while the results of Citations are shown in Figure 19.

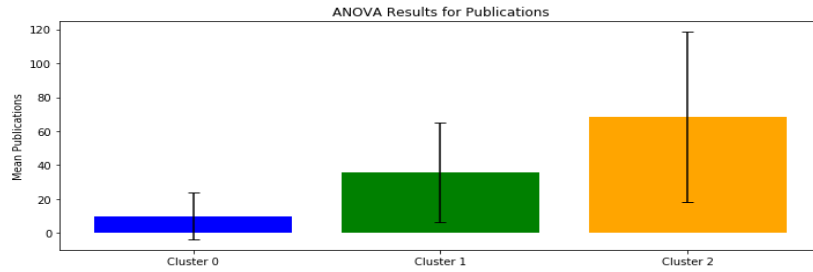


Figure (18): ANOVA results for Publications

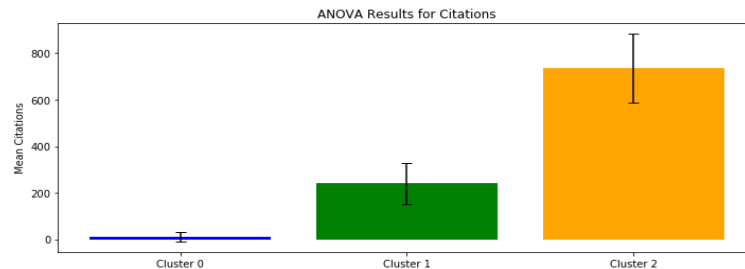


Figure (19): ANOVA results for Citations

Results in the above two figures provide statistical evidence that the results obtained from the K-means algorithm are not random but reflect a real variation in the data.

5. CONCLUSION

The field of data analysis is undeniably driven by data. Therefore, the unavailability of datasets poses a significant challenge for data analysts. Using web-scraping, data engineers can collect data from different web pages and store it as a dataset to overcome the problem of unavailability of data. The need for a dataset containing researchers' information was the reason behind this work. Therefore, the ResearchGate website was scraped to build the needed dataset using the BeautifulSoup library. Data extraction was successfully conducted, laying the groundwork for future exploration and analysis of the dataset using Exploratory Data Analysis (EDA) advanced techniques. By constructing this dataset from scratch, this work not only provides an organized repository of academic profiles but also opens doors for further research on academic networks and researcher profiling based on structured social media data. The challenges and fundamental limitations in this work include the fact that some researchers have not fully completed their accounts on the ResearchGate platform. For instance, college affiliation was not specified by some researchers, which constitutes a barrier to establishing a comprehensive dataset. We hope in the future to communicate with the relevant authorities to guide researchers to complete their profiles fully and accurately. This will allow us to finalize and share a comprehensive researcher dataset on platforms, such as Kaggle, for the public benefit.

REFERENCES

- [1] Abdollahi, Fatemeh & Moosavi, Sara & Riahinia, Nosrat. "Analyzing the scientific presence of Iranian women inventors in LinkedIn and ResearchGate social networks" 10.22070 /RSCI.2023.16945.1624, 2023.
- [2] Adila, Nelawati. "Implementation of Web Scraping for Journal Data Collection on the SINTA Website". Sinkron. 7. 2478-2485. 10.33395/sinkron.v7i4.11576, 2022.
- [3] BeautifulSoup, [Online] Available: <https://www.crummy.com/software/BeautifulSoup>, 2023
- [4] Brenning, A. and Henn, S., "Web scraping: a promising tool for geographic data acquisition", doi:10.48550/arXiv.2305.19893, 2023.
- [5] C, Priya & Salvi, Pawan & Pawar, Shravani. "ISSUES AND CHALLENGES OF WEB SCRAPING: HEALTHCARE INDUSTRY CASE STUDY APPROACH". The Online Journal of Distance Education and e-Learning, January 2023 Volume 11, Issue 1, 2023.
- [6] Himawan, Arif & Priadana, Adri & Murdiyanto, Aris. "Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application", IJID (International Journal on Informatics for Development). 9. 59-65. 10.14421/ijid.2020.09201, 2020.

- [7] Morina, Vesa & Sejdiu, Shqipe. "Evaluating and comparing web scraping tools and techniques for data collection". 11th UBT Annual International Conference On Computer Science And Engineering, 2022.
- [8] Onyenwe, Ikechukwu & Gr, Onyedinma & Nwafor, Chidinma & Agbata, Obinna. "Developing Products Update-Alert System for e-Commerce Websites Users Using HTML Data and Web Scraping Technique", 2021.
- [9] R. S. Chaulagain, S. Pandey, S. R. Basnet and S. Shakya, "Cloud Based Web Scraping for Big Data Applications," IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 2017, pp. 138-143, doi: 10.1109/SmartCloud.2017.28, 2017.
- [10] ResearchGate, [Online] Available: www.researchgate.net, 2024.
- [11] S. Bale, N. Ghorpade, R. S, S. Kamalesh, R. R and R. B. S, "Web Scraping Approaches and their Performance on Modern Websites," 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 956-959, doi: 10.1109/ICESC54411.2022.9885689, 2022
- [12] Shaikh, Arman & Khan, Raihan & Panokher, Komal & Ranjan, Mritunjay & Sonaje, Vaibhav. "E-commerce Price Comparison Website Using Web Scraping". International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences. Volume 11. 1-13. 10.37082/IJRMPS.v11.i3.230223, 2023.
- [13] Panda, Subhajit and Kaur, Navkiran, Research Performance of Top Cited Indian Researchers on ResearchGate Platform: An Altmetric Analysis (August 29, 2023). Journal of Information and Knowledge, 60(4), 267–280, 2023.
- [14] Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1328-1347, 1, doi: 10.1109/TKDE.2019.2946162, April 2021.

انشاء قاعدة بيانات لأعضاء موقع بوابة الباحث

زيد عبد الاله منذر⁽¹⁾ ، ياسر علي محمود⁽²⁾

^(1,2) قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

المستخلص:

في البيئة الأكاديمية اليوم، تلعب منصات التواصل الاجتماعي، مثل موقع ResearchGate، دورًا محوريًا في تسهيل التعاون والتواصل بين الباحثين. وعلى الرغم من الاستخدام الواسع لهذا الموقع، لا تزال هناك فجوة كبيرة في توفر مجموعات بيانات منظمة تركز على الباحثين وإنتاجهم الأكاديمي. إن هذا النقص في البيانات المتاحة يعوق إمكانية إجراء تحليل شامل وتقييم لتوجهات البحث وأثرها. تهدف هذه الدراسة إلى سد هذه الفجوة من خلال استخدام تقنيات استخراج البيانات لبناء مجموعة بيانات مشتقة من موقع ResearchGate، والذي يعد من أبرز المنصات المتخصصة للمحترفين الأكاديميين. تتكون مجموعة البيانات المقدمة من ثمانية ميزات رئيسية، تشمل مقاييس تتعلق بالنشر، والاستشهادات، ومجالات التخصص البحثي. يتيح توفر مثل هذه البيانات مصدرًا قيمًا للبحوث المستقبلية، ويمكن الباحثين من تحليل الأداء البحثي، وتحديد فرص التعاون، والكشف عن توجهات الإنتاجية الأكاديمية. إلى جانب بناء مجموعة البيانات، يعرض هذا البحث أيضًا تحليل البيانات الاستكشافي (EDA) الذي أجري لاستخلاص رؤى من البيانات المجمعة. كما تم تطبيق خوارزمية التجميع K-means لتصنيف الباحثين وفقًا لأنماط النشر والاستشهادات، مما يوفر فهمًا أوضح للإنجازات الأكاديمية. تتمثل المساهمة الرئيسية لهذا العمل في إنشاء مجموعة بيانات للباحثين يمكن استخدامها في الدراسات المستقبلية لتحليل وفحص جهود وتوجهات الباحثين في أعمالهم العلمية. وتتمثل المساهمة الرئيسية لهذه الدراسة في إنشاء مجموعة بيانات شاملة تعالج فجوة مهمة في البحث الأكاديمي. ومن خلال تسليط الضوء على إمكانيات موقع ResearchGate كمصدر بيانات، واستخدام تقنيات استخراج البيانات من الويب بطرق مبتكرة، تهدف هذه الدراسة إلى تعزيز استكشاف وتحليل البيانات الأكاديمية بشكل أعمق، مما يساهم في تطوير منهجيات البحث الأكاديمي.