ANALYZING MEDIUM AUTHORS

A
Project Report
Of
B.Tech and M.Tech
In
Mathematics and Data Science

Submitted By :
Sanyam Jain
(214104020)

Academic Year 2022-23

Department of Mathematics, Bioinformatics
and Computer Application

MAULANA AZAD
NATIONAL INSTITUTE OF TECHNOLOGY, BHOPAL

APRIL 2023

CONTENT

INTRODUCTION

Medium is a widely used online platform that offers authors the opportunity to publish articles on a diverse range of topics. It is a valuable source for data analysis due to its rich and diverse data, access to author profiles, real-time data, insights for content creation and marketing strategies, and potential impact on product optimization. Conducting data analysis on Medium data can provide valuable insights into author behavior, article content, and platform dynamics, which can inform decision-making and strategy development in various domains.

In this report, we will delve into the potential of author analysis in providing a comprehensive understanding of the characteristics of Medium article authors. The report will provide a detailed overview of the exploratory data analysis (EDA) process, encompassing key steps such as data collection, data cleaning, data analysis, and the insights gained from the analysis. Through this report, we aim to uncover significant findings that can shed light on the demographic profiles, writing styles, and publication frequency of authors on Medium. By conducting a thorough analysis of these factors, we can gain valuable insights that may inform content creation strategies, marketing trends and approaches for authors, publishers, and Medium as a whole.

OBJECTIVE

The objective of this report is to conduct an author analysis of Medium articles in order to gain a deeper understanding of the characteristics of authors who are most active on the platform, including their demographic profiles, writing styles, topics of interest and major publication they are active into. This analysis aims to provide valuable insights that can inform content creation strategies, platform policies, and marketing approaches for authors, publishers, and Medium as a whole.

Furthermore, this report aims to demonstrate the diverse range of potential use cases by leveraging simple sentiment analysis and summarization techniques on the content of Medium articles. This analysis will shed light on the valuable insights that can be derived from such text analysis methods, providing a glimpse into the potential applications of these techniques in the context of Medium article analysis.

PROCESS

The process for conducting an author analysis of Medium articles involves several key steps, including data collection, data cleaning, data analysis, and gaining insights from the analysis. We will outline each of these steps in detail to provide a comprehensive overview of the process.

Data Collection

The first step in the process is to collect data from Medium using Medium API. The Medium API allows us to access a wide range of information about articles, including author profiles, top contributing writers, article titles, publication dates, and more. We can use Python libraries, such as requests and json, to make API requests and retrieve data.

Since they provide a python package to fetch the results directly, we considered it as a better option and proceeded with it. The package allowed us to fetch articles, authors, and publication information from Medium. The retrieved data was saved in a Pandas DataFrame for further analysis.

Data Cleaning

Once the data was obtained, the next step involved cleaning and preprocessing it to ensure its accuracy and reliability for analysis. This included tasks such as removing duplicate records, handling missing values, dropping irrelevant fields, and correcting any inconsistencies in the data. Python library, Pandas, was used for data manipulation and cleaning tasks.

➔ Unnecessary columns were removed :

Before :

```
Index(['id', 'tags', 'claps', 'last_modified_at', 'published_at', 'url',
       'image_url', 'is_series', 'lang', 'publication_id', 'word_count',
       'is_locked', 'title', 'reading_time', 'responses_count', 'voters',
       'topics', 'author', 'subtitle'],
      dtype='object')
```

After :

```
Index(['id', 'tags', 'claps', 'published_at', 'lang', 'publication_id',
       'word_count', 'title', 'responses_count', 'voters', 'topics',
       'subtitle'],
      dtype='object')
```

➔ Rows having multiple “topics” were separated (using pandas explode command):

```
Before seperation by topics
0                    [artificial-intelligence]
1                        [design, photography]
2      [artificial-intelligence, programming]
3                                [programming]
4      [artificial-intelligence, programming]
                           ...
244                           [data-science, tv]
245                [productivity, self, work]
246                               [data-science]
247                    [productivity, gadgets]
248           [machine-learning, data-science]
Name: topics, Length: 249, dtype: object

 After seperation by topics
0      artificial-intelligence
1                       design
1                   photography
2      artificial-intelligence
2                   programming
                 ...
246                data-science
247                productivity
247                     gadgets
248            machine-learning
248                data-science
Name: topics, Length: 428, dtype: object
```

Data Analysis

Once the data was cleaned, we conducted a comprehensive analysis to gain insights into the characteristics of Medium article authors. This included-

1. First, finding top authors writing for a specific input topic. So that we can do our further analysis on the right sample data.
2. Finding user data including : bio, followers, following, social handle and more.
3. Finding each article's data published by the author.
4. Finding topics the author writes about.
5. Finding the top publication the author publishes in.
6. Finding popularity of articles, based on clap count and voter count. And how it is related to corresponding topics and publications.
7. Operations on the content of article:
    a) Finding information regarding the size of content itself.
    b) Doing sentimental analysis of the content.
    c) Performing Summarization.

Pandas was used to perform various data manipulation and analysis functions to derive meaningful insights from the data.

Insights and Visualization

Based on the analysis, we gained valuable insights into the profiles of Medium article authors (which were stated in the last section) which made us make assumptions like which topics are more trending in current time, which publications showed better results then others. The sentimental analysis showed us what kind of articles the author is more into. We also got the summarization results of the content which is a valuable insight, especially for people who don't have that much time to go through the whole article.

To enhance the understanding of the data, visualization techniques were utilized. Seaborn and Matplotlib were used to create various plots and charts to visualize patterns, trends, and relationships in the data, making it easier to interpret and communicate the findings visually.

RESULTS

1. Top writers

Following is the list of top writers we obtained under the topic "**data-science**" :

```
The PyCoach --- frank-andrade
Giorgos Myrianthous --- gmyrianthous
Sanjay Priyadarshi --- priyadarshisanjay
Jim Clyde Monge --- jimclydemonge
Molly Ruby --- molly.ruby
Bex T. --- ibexorigin
B. Chen --- bindichen
Youssef Hosni --- youssefraafat57
Lars Nielsen --- pythoslabs
Thuwarakesh Murallie --- thuwarakesh
```

Next, we performed our further analysis on the Rank-1 author of this list.

Username : **"frank-andrade" .**

2. User Data

Here's the user data we obtained:

```
Username:  frank-andrade
Fullname:  The PyCoach
Bio:  8M+ Views on Medium || Early Bird Discount: Make money by writing about AI, programming, data science or tech 👉 http://bit.ly/3zfbgiX
Profile Image:  https://miro.medium.com/1*veEX4-CiLz5jqUjwWfQo_Q.jpeg
Top Writer In:  ['artificial-intelligence', 'technology', 'science', 'entrepreneurship', 'business']
User Written Articles:  249
Followers:  47887
Following:  7
```

From the number of articles written by it can be concluded that this person is quite active and has written a lot about this field. He is successful in terms of followers - 47887!

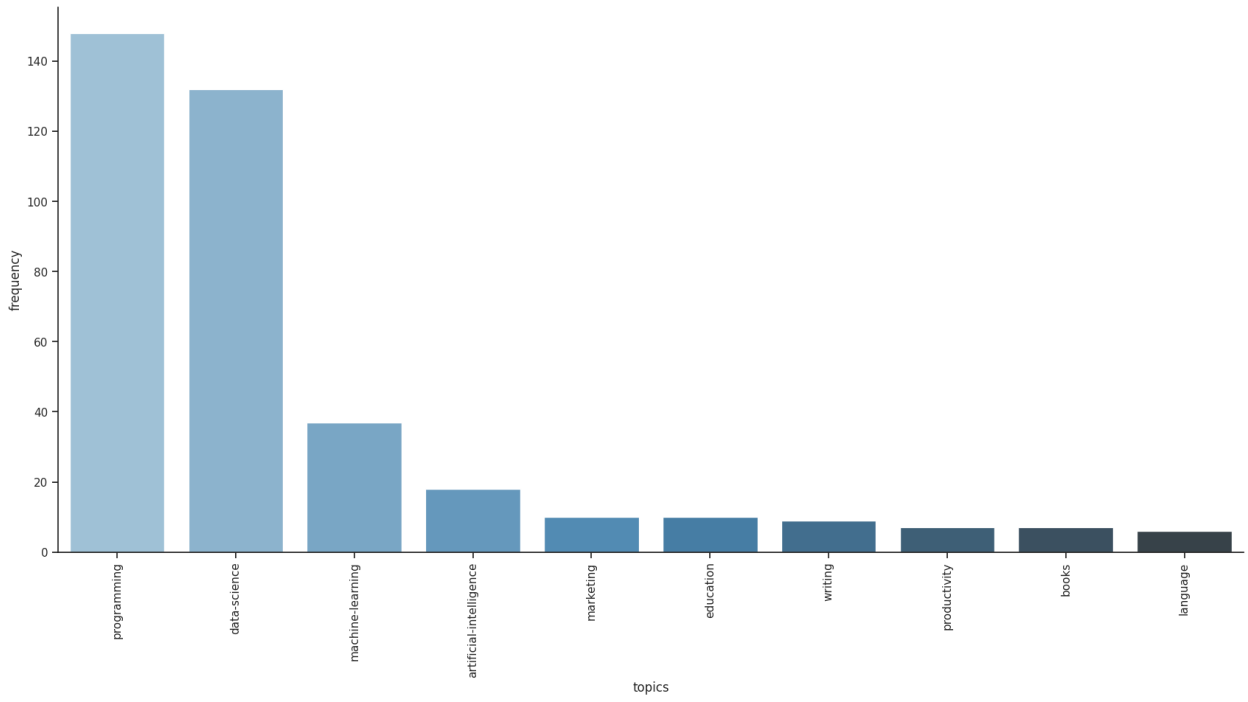
3. Article Data

After fetching results of each article's data, doing some preprocessing and saving it into pandas's dataframe, here's how the database looks like (starting 5 articles only):

	id	tags	claps	published_at	lang	publication_id	word_count	title	responses_count	voters	topics	subtitle
0	8ac0ffb14af9	[artificial-intelligence, chatgpt, technology,...	114	2023-04-14 09:01:08	en	*Self-Published*	310	Share Your AI Knowledge With The World: Write ...	4	20	artificial-intelligence	Calling all AI and tech enthusiasts.
1	c876fbe7915e	[artificial-intelligence, technology, chatgpt,...	784	2023-04-12 15:39:52	en	76436a11a2b0	1272	You're Using Midjourney Wrong! Here's How to C...	12	101	design	Generate amazing images by learning how to cre...
1	c876fbe7915e	[artificial-intelligence, technology, chatgpt,...	784	2023-04-12 15:39:52	en	76436a11a2b0	1272	You're Using Midjourney Wrong! Here's How to C...	12	101	photography	Generate amazing images by learning how to cre...
2	bda045eed47f	[technology, chatgpt, python, data-science, ar...	742	2023-04-11 10:30:02	en	76436a11a2b0	868	The ChatGPT Skill That Pays Up to $335,000 a Year	13	143	artificial-intelligence	AI is creating amazing new jobs.
2	bda045eed47f	[technology, chatgpt, python, data-science, ar...	742	2023-04-11 10:30:02	en	76436a11a2b0	868	The ChatGPT Skill That Pays Up to $335,000 a Year	13	143	programming	AI is creating amazing new jobs.

4. **Top topics the author has most wrote about (**Top 10) :

```
programming                148
data-science               132
machine-learning            37
artificial-intelligence     18
marketing                   10
education                   10
writing                      9
productivity                 7
books                        7
language                     6
Name: topics, dtype: int64
```

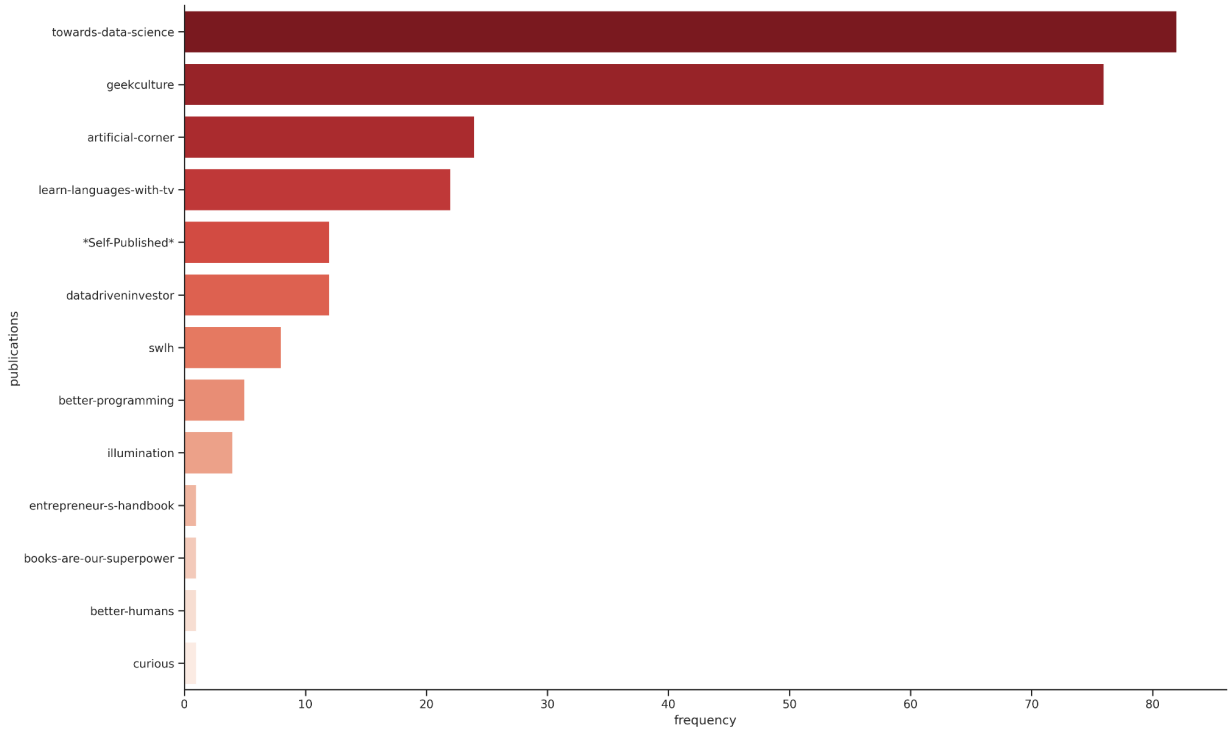


The frequency of topics graph obtained from the analysis provides valuable insights into the writer's writing patterns and preferences. It serves as a crucial tool for identifying the prominent themes in the writer's content and understanding their areas of expertise. He is more into programming and data science stuff, but also has tried writing about other topics like marketing and education as well.

The graph offers a visual representation of the data, making it easier to identify trends, biases, and audience engagement strategies. Overall, this analysis contributes to a comprehensive understanding of the writer's writing style and provides valuable information for content planning and strategic decision-making.

5. **Publications the author has published mostly:**

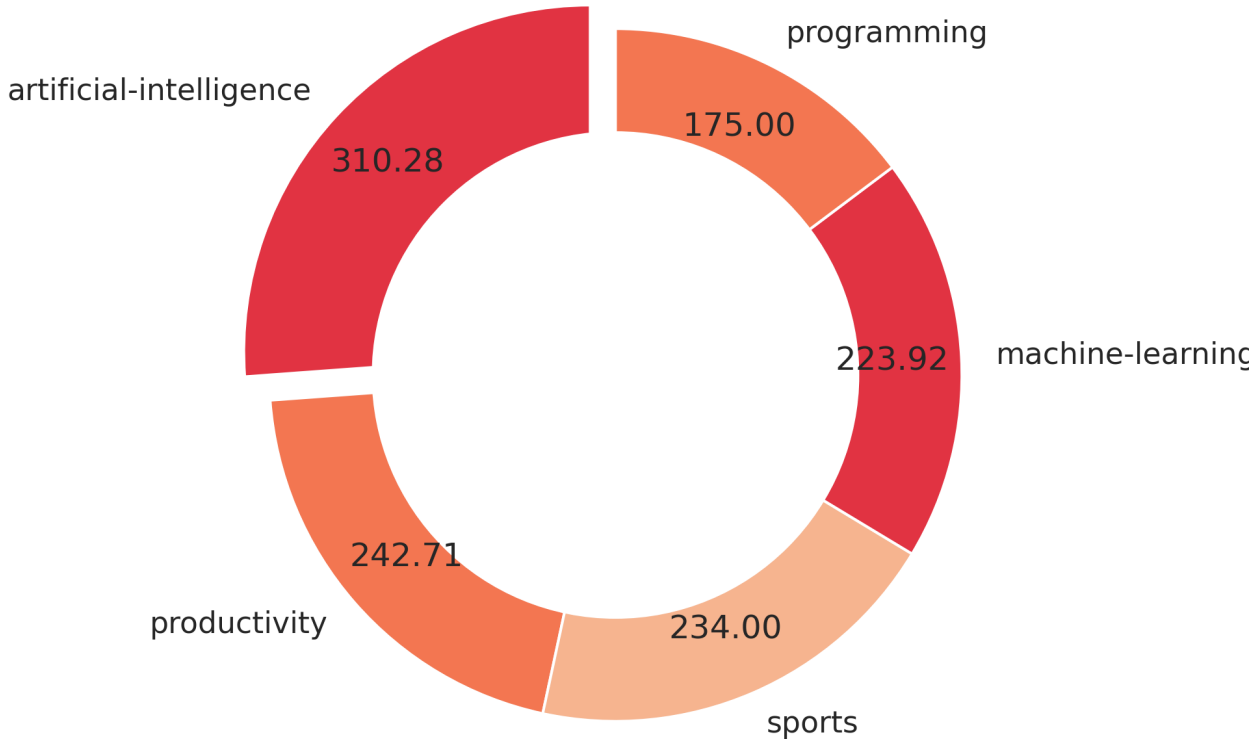
```
towards-data-science        82
geekculture                 76
artificial-corner           24
learn-languages-with-tv     22
*Self-Published*            12
datadriveninvestor          12
swlh                         8
better-programming           5
illumination                 4
entrepreneur-s-handbook      1
books-are-our-superpower     1
better-humans                1
curious                      1
Name: publication_slug, dtype: int64
```



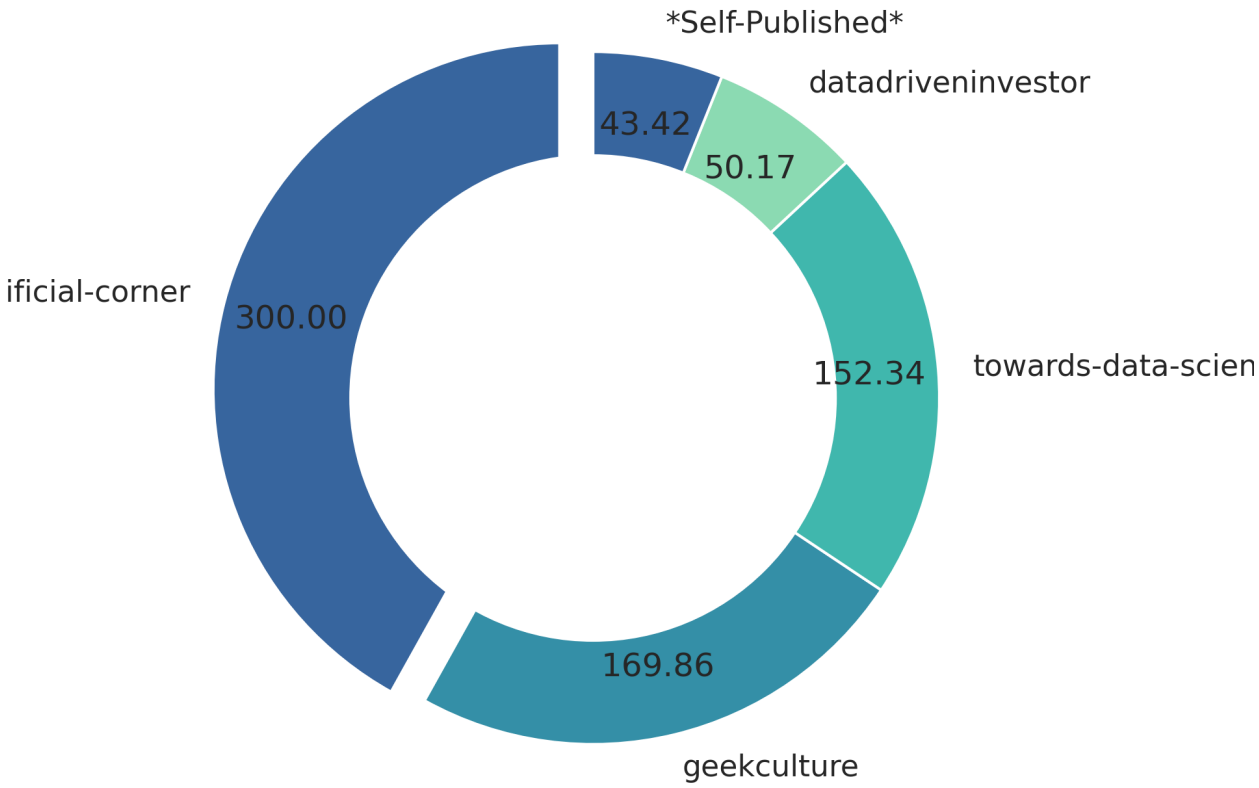
The graph showcases the frequency of articles published in each publication, allowing for easy identification of the publications where the writer has been most active. This analysis provides a comprehensive overview of the writer's engagement with different publications, shedding light on their preferred platforms for sharing their content.

6. Popularity of articles

Based upon the clap count and voter count we get a rough estimation of how well an article has received attention from its readers. And we have shown how this is related to the topic of the article and the publication in which it was published.



The first pie-chart showcases the engagement and readership of articles related to topics, allowing for easy identification of the topics that have garnered the most attention from readers (Here its Artificial Intelligence). This analysis provides a comprehensive overview of the writer's content performance and the interests of their audience.



Similarly, This graph showcases the relative popularity of articles in each publication, allowing for easy identification of publications where the articles have gained more traction. In this case, the articles author published in “artificial-corner” received the most votes. This analysis provides valuable insights into the readership and engagement patterns of articles across different publications.

7. Article Content Analysis

a) Sentimental Analysis

Following are the results of the sentimental analysis we got by using NLTK library:

	title	sentiment_score
0	Share Your AI Knowledge With The World: Write ...	0.9846
1	You're Using Midjourney Wrong! Here's How to C...	0.9990
2	The ChatGPT Skill That Pays Up to $335,000 a Year	0.9985
3	The Chatbot Competition: A Hands-On Comparison...	0.9990
4	All AI Tools And ChatGPT Prompts in 1 Article	0.9988
...	...	...
244	Learn a Foreign Language with Friends	0.9997
245	Learn the language hacking method, jump on Hel...	0.9999
246	The Simpsons is the Best TV show to Increase Y...	0.9993
247	6 Months Without an iPhone—This Is How My Li...	0.9976
248	The Best Movies to Learn a Foreign Language	0.9997

The sentiment analysis results offer valuable information for strategic decision-making, content planning, and brand management.

The majority of the articles exhibit a highly positive sentiment, with scores approaching 1. This outcome was anticipated, as the subject matter pertains to learning data, which typically involves a positive and explanatory approach.

b) Summarization

We utilized txt.ai's summarization model to effectively condense lengthy articles into concise summaries of less than 50 words, which were then stored for future use. This approach aimed to improve efficiency, provide a high-level overview, prioritize relevant information, enhance interpretability, and facilitate data reduction during the analysis process.

CONCLUSION

The utilization of the Medium API's Python package for importing author's data and subsequent analysis using pandas dataframe has proven to be an effective and efficient approach. The retrieved data from Medium provided valuable insights into the author's performance and engagement metrics, which enabled us to gain a deeper understanding of their content's reach and impact.

The cleaning and preprocessing steps applied to the data ensured that the analysis was based on accurate and reliable information. By leveraging the rich functionality of pandas, we were able to perform various data manipulations, such as filtering, aggregation, and merging, to derive meaningful insights.

The analysis of the data provided us with valuable information about the author's most popular topics, audience engagement patterns, and content performance over time. These insights can be utilized to inform content strategy, optimize performance, and identify opportunities for growth.

The visualization of the analyzed data using appropriate charts and graphs aided in effectively communicating the findings to stakeholders. Visualizations allowed for easier identification of trends, patterns, and outliers, making it simpler to understand the author's performance visually.

RESOURCES

Code for this Project:

https://colab.research.google.com/drive/1YIArScpRodVWvWgOo_xtioxjRmvLtEte?usp=sharing

Medium API Documentation:

https://docs.mediumapi.com/

Medium API Python Package:

https://pypi.org/project/medium-api/