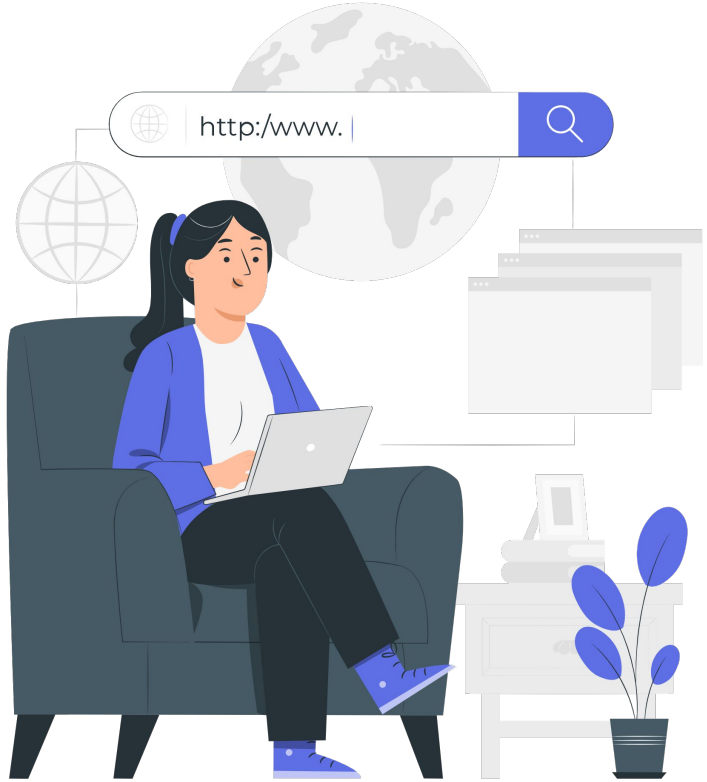
CS 328 - Intro to Data Science

ACAD SEARCH

A crowd-sourced search engine to find leading professors

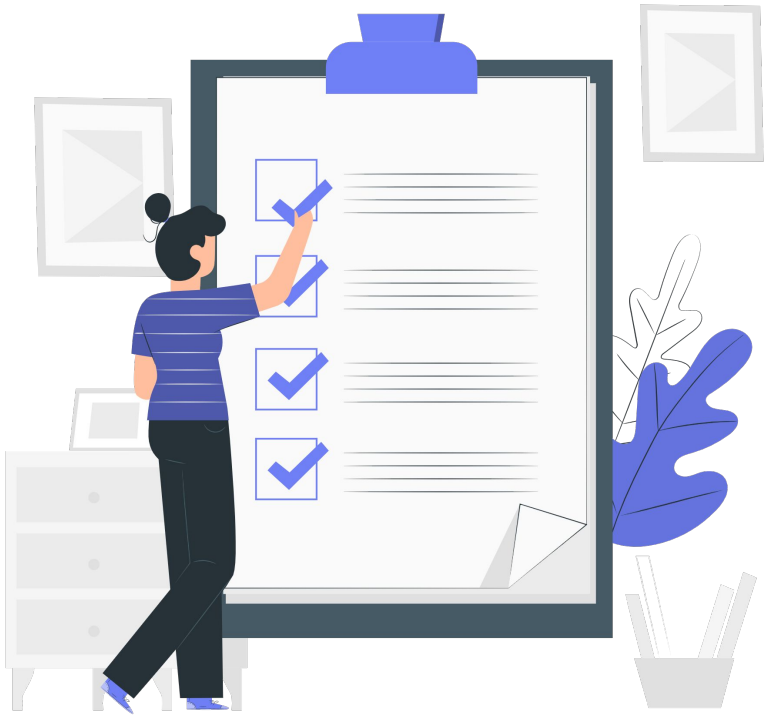


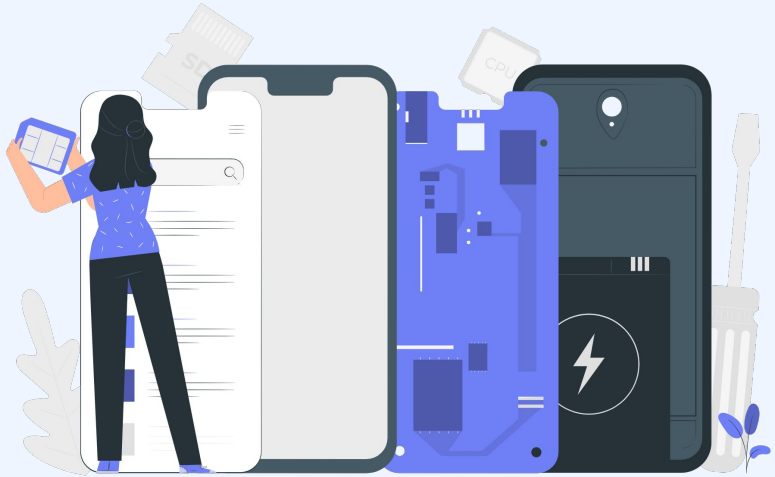
MOTIVATION

- ➔ Students often need to search for professors.
- ➔ Google Scholar provides less control over search results.
- ➔ A tool focused towards students searching for professors.

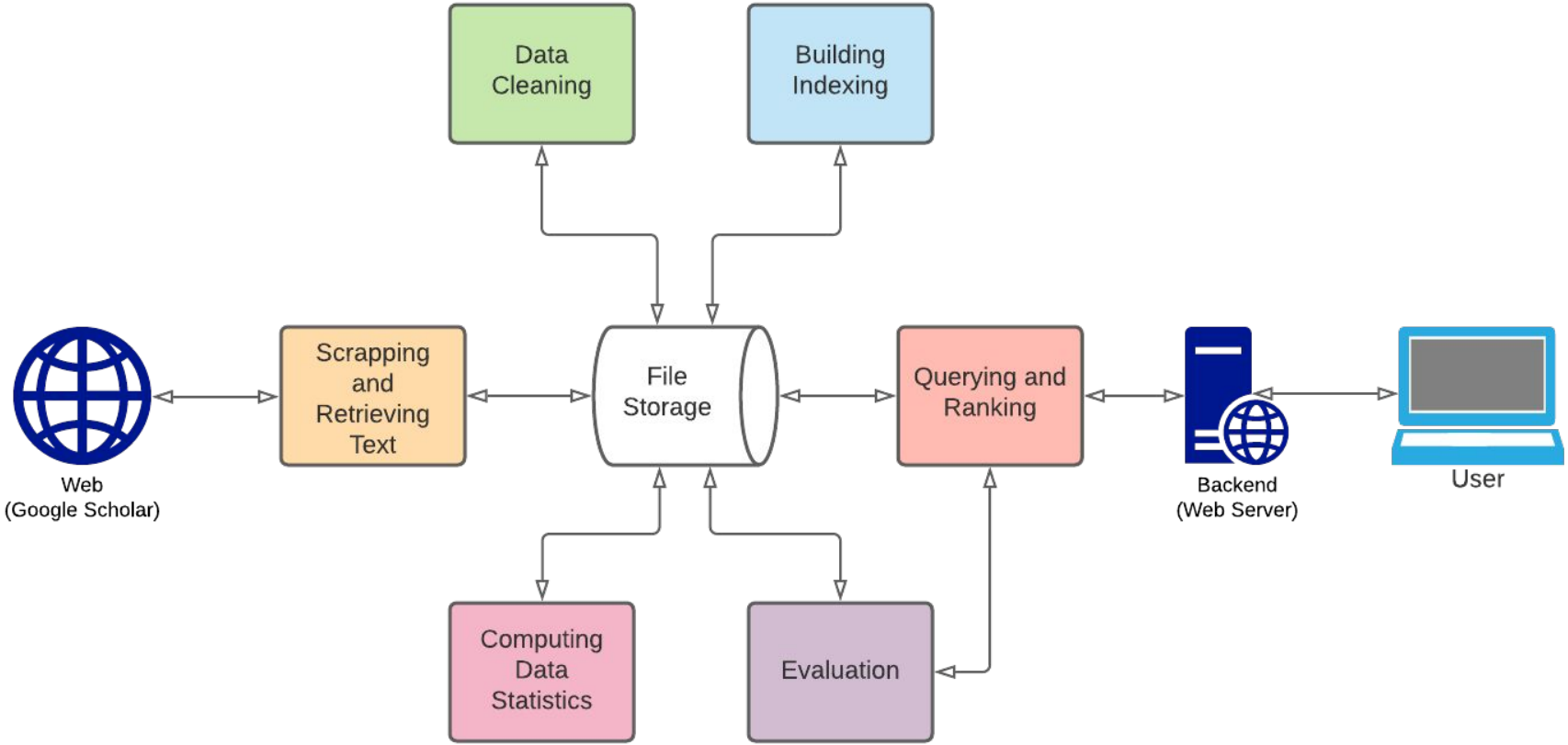
OBJECTIVES

- Search Engine for professors based on name, topic, paper title, etc.
- Ranking and filtering based on h-index, citations over last 5 years, conferences, etc.
- Ranking metric based on the data.





HIGH LEVEL ARCHITECTURE

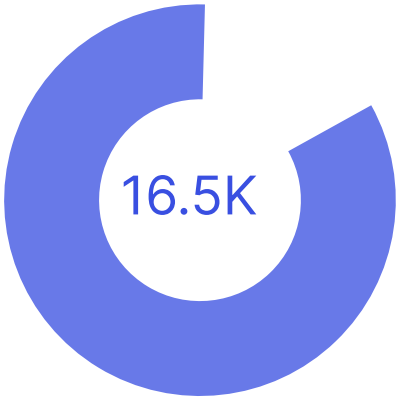

Data Cleaning
Building Indexing
Web (Google Scholar)
Scrapping and Retrieving Text
File Storage
Querying and Ranking
Backend (Web Server)
User
Computing Data Statistics
Evaluation

DATA COLLECTION

SOURCE

List of Google Scholar IDs from [CSRankings GitHub Repository]

Repo



Scraped

Cleaned

SCRAPING

Data scraped per professor:

- Name
- Affiliation
- Profile Image URL
- Verified Email at
- Homepage URL
- Research Topics List

- Citation (overall & 5 yrs)
- H-Index (overall & 5 yrs)
- I-Index (overall & 5 yrs)
- Citation List Year-wise
- Titles: Top 100 Cited Papers
- URLs: Top 100 Cited Papers

AVOIDING GETTING BLOCKED

- ➔ Pass request headers to appear as a browser.
- ➔ Random delays of 1-3 seconds between requests.
- ➔ Divide across multiple IP addresses.

PRE PROCESSING

CLEANING

- ➔ Removed redundant entries.
- ➔ Some Google Scholar pages did not follow the same HTML conventions as the rest. Cleaned such corruptions using regular expressions.

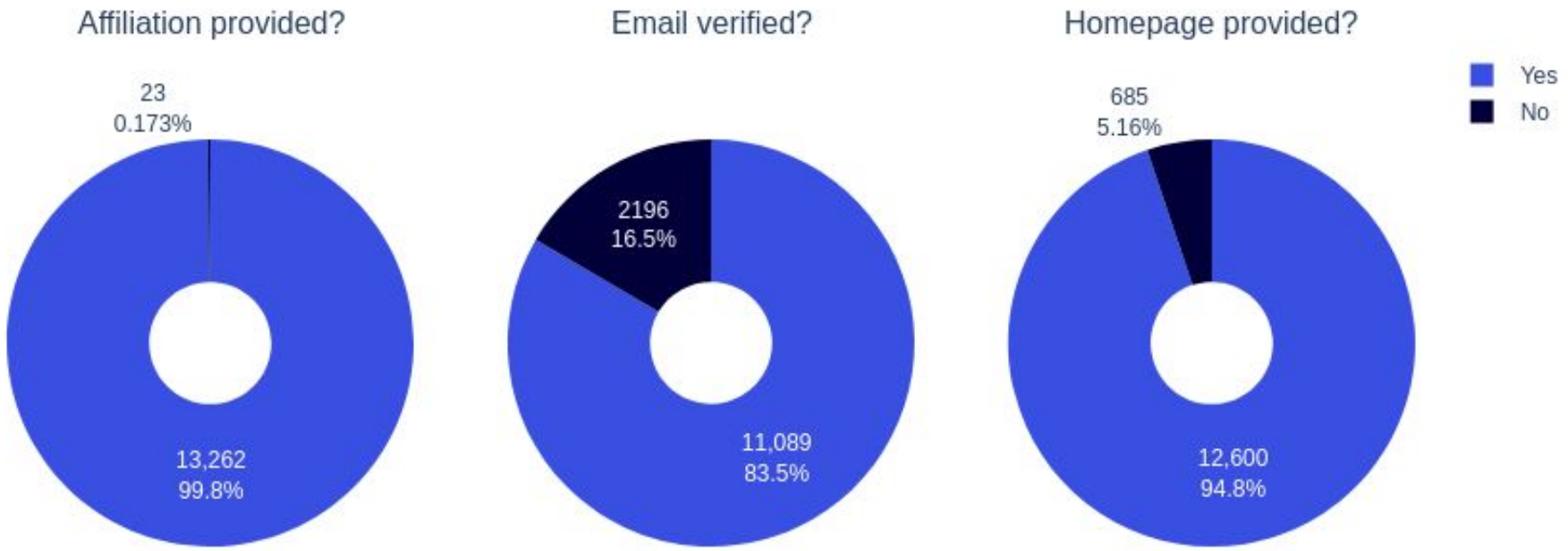
TEXT PRE-PROCESSING

- ➔ Convert to lower-case.
- ➔ Perform stemming.
- ➔ Remove stop words.

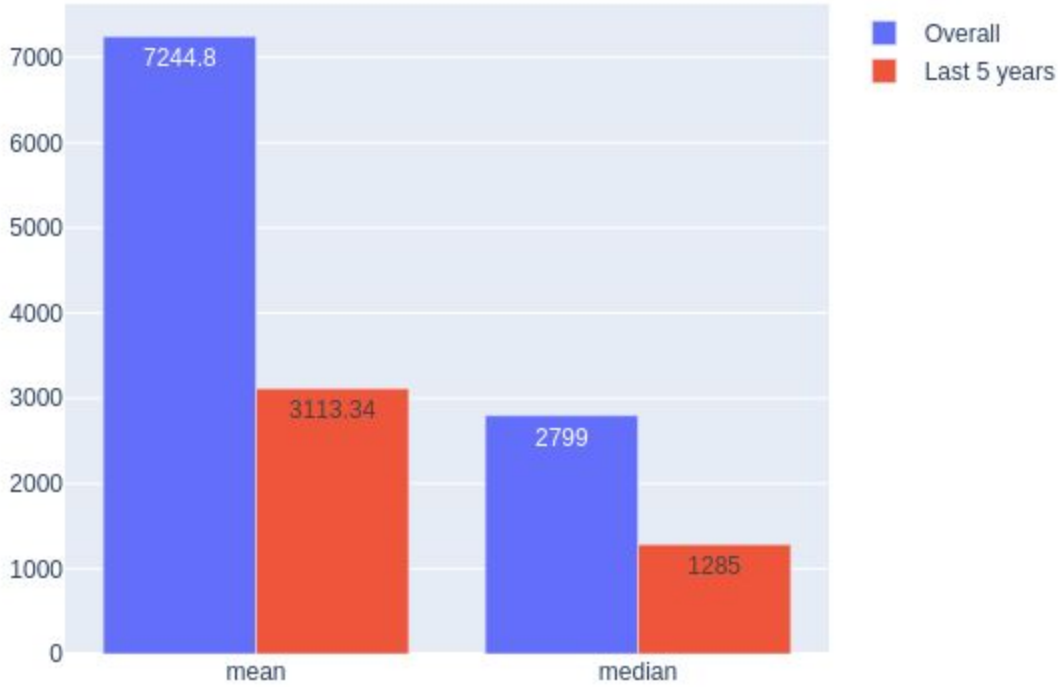
DATA STATISTICS

BASICS

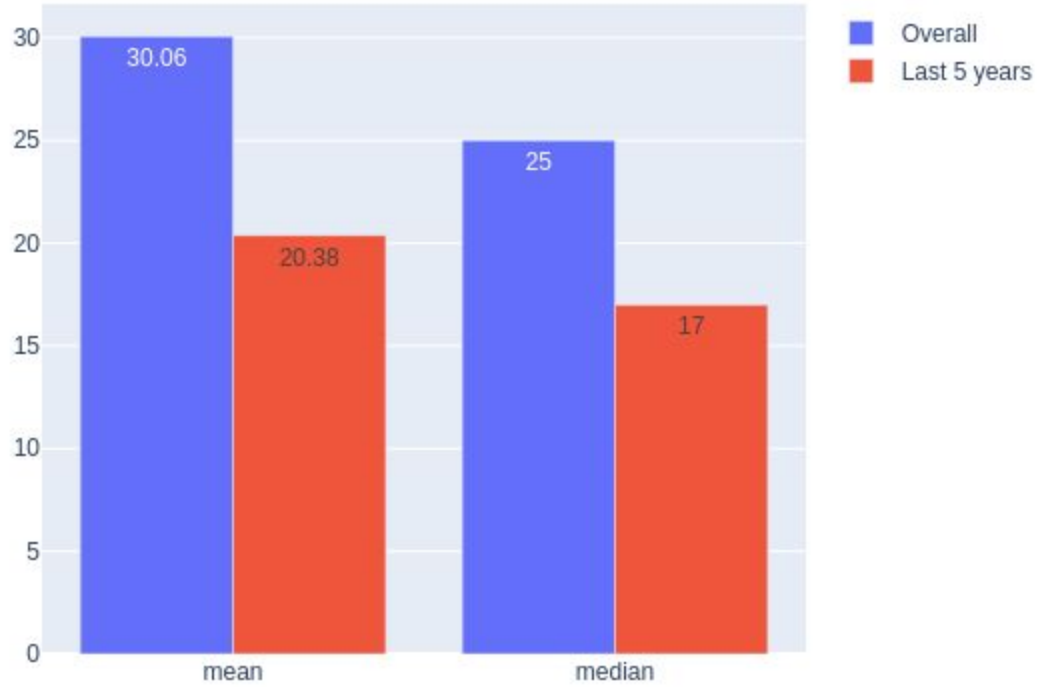
Number of professors	13285
Number of institutions	8716
Number of publications	893575
Average number of publications per professor	79.80
Size of data after cleaning	232 MB


Affiliation provided?
23
0.173%
13,262
99.8%
Email verified?
2196
16.5%
11,089
83.5%
Homepage provided?
685
5.16%
12,600
94.8%
Yes
No

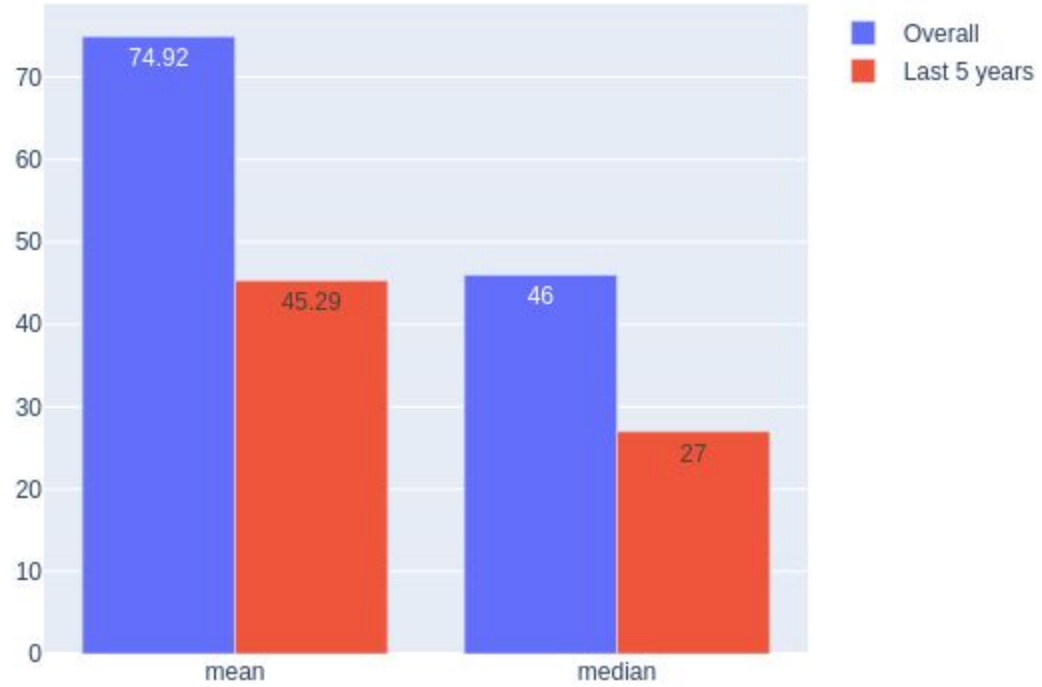
No. of citations

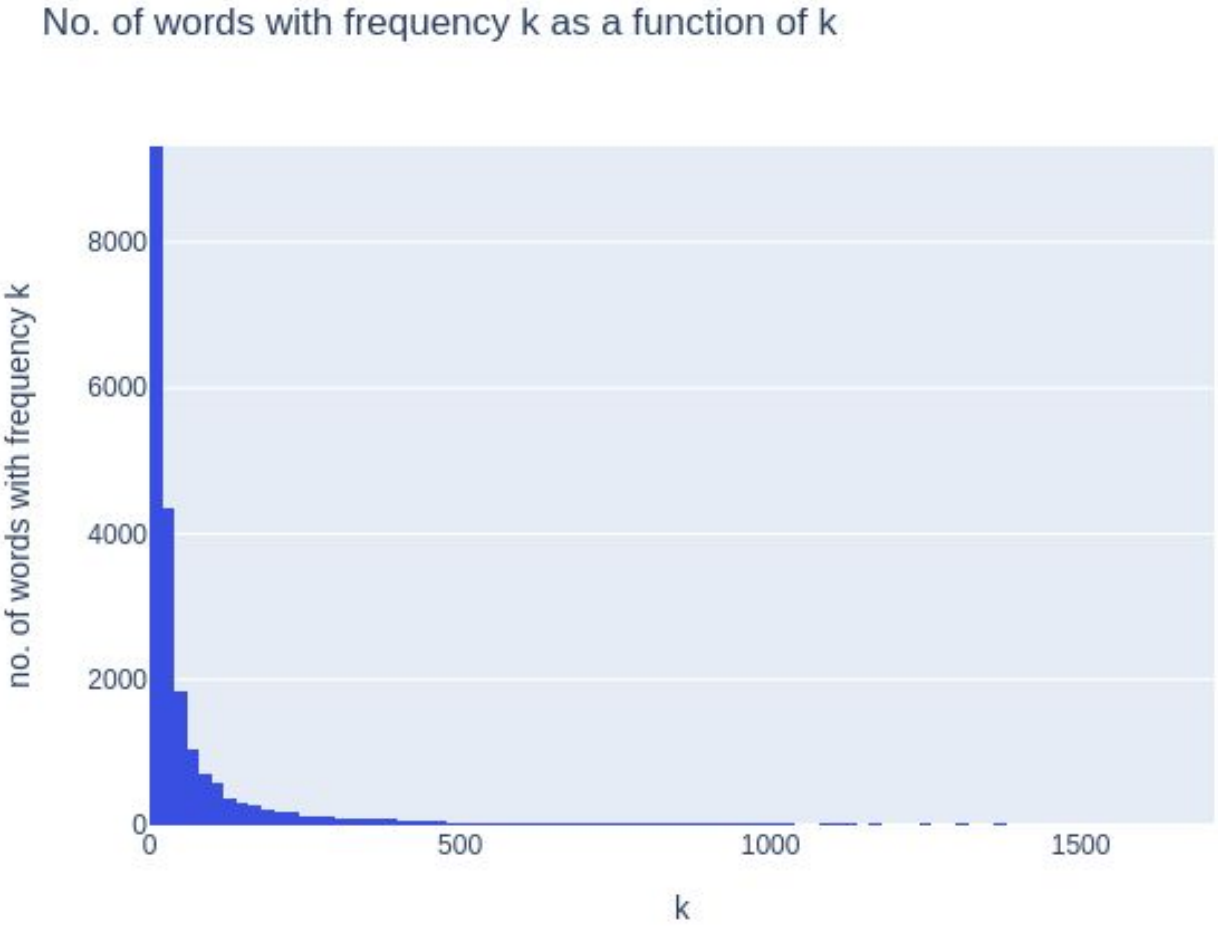


h-index


30
25
20
15
10
5
0
30.06
20.38
25
17
mean
median
Overall
Last 5 years

i10-index


74.92
45.29
46
27
70
60
50
40
30
20
10
0
mean
median
Overall
Last 5 years


No. of words with frequency k as a function of k
no. of words with frequency k
8000
6000
4000
2000
0
0
500
1000
1500
k

INDEXING

2 INDICES

- ➔ Research topics and paper titles
- ➔ Name and Affiliation(Institution)

INVERTED INDEX

{

‘machin’ : [[0, 12], [3, 16], [6, 9]],

‘learn’ : [[0, 13], [3, 17], [8, 18]]

}

QUERYING & RANKING

BOOLEAN RETRIEVAL

- ➔ **AND**: Return professors that contain all the words from the query
- ➔ **OR**: Return professors that contain at least 1 word from the query.
- ➔ Optimisation

PHRASE RETRIEVAL

- ➔ Allows users to search for exact matches for phrases.
- ➔ The word positions stored in the inverted index is used.

TF-IDF (Term Frequency-Inverse Document Frequency)

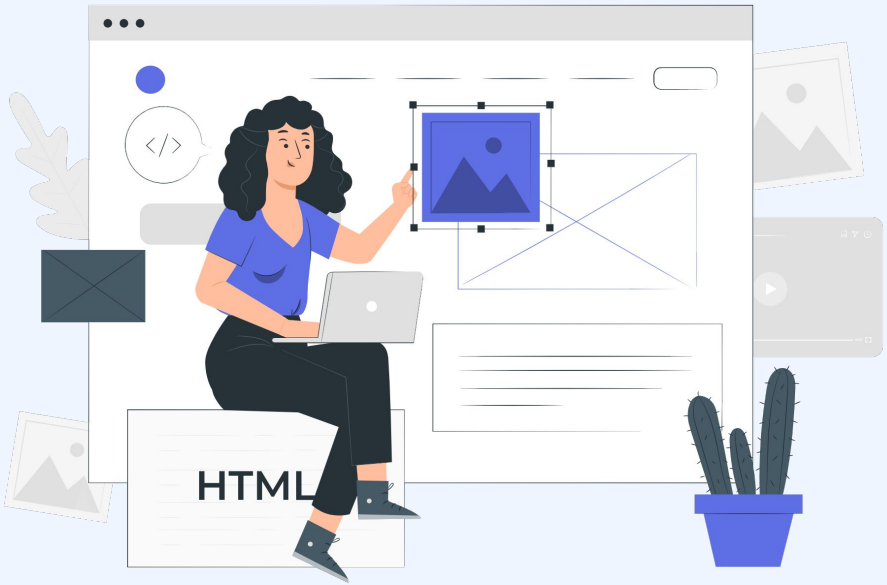
- TF-IDF(t, d, D) = TF(t, d) x IDF (t, D)
- TF: measure of how frequently a word/term appears in a document, normalised by the document size.
- TF(t, d) = no. of times t appears in d ÷ total no. of words in d
- IDF: measure of how much information a word provides, i.e., if it's common or rare across all documents.
- IDF(t, D) = log[no. of documents ÷ (1 + no. of documents containing t)]

SORTING RESULTS

- ➔ Users can sort search results based on no. of citations, h-index, no. of citations in last 5 years etc.
- ➔ Boolean and phrase retrieval results are sorted using a default ranking metric.

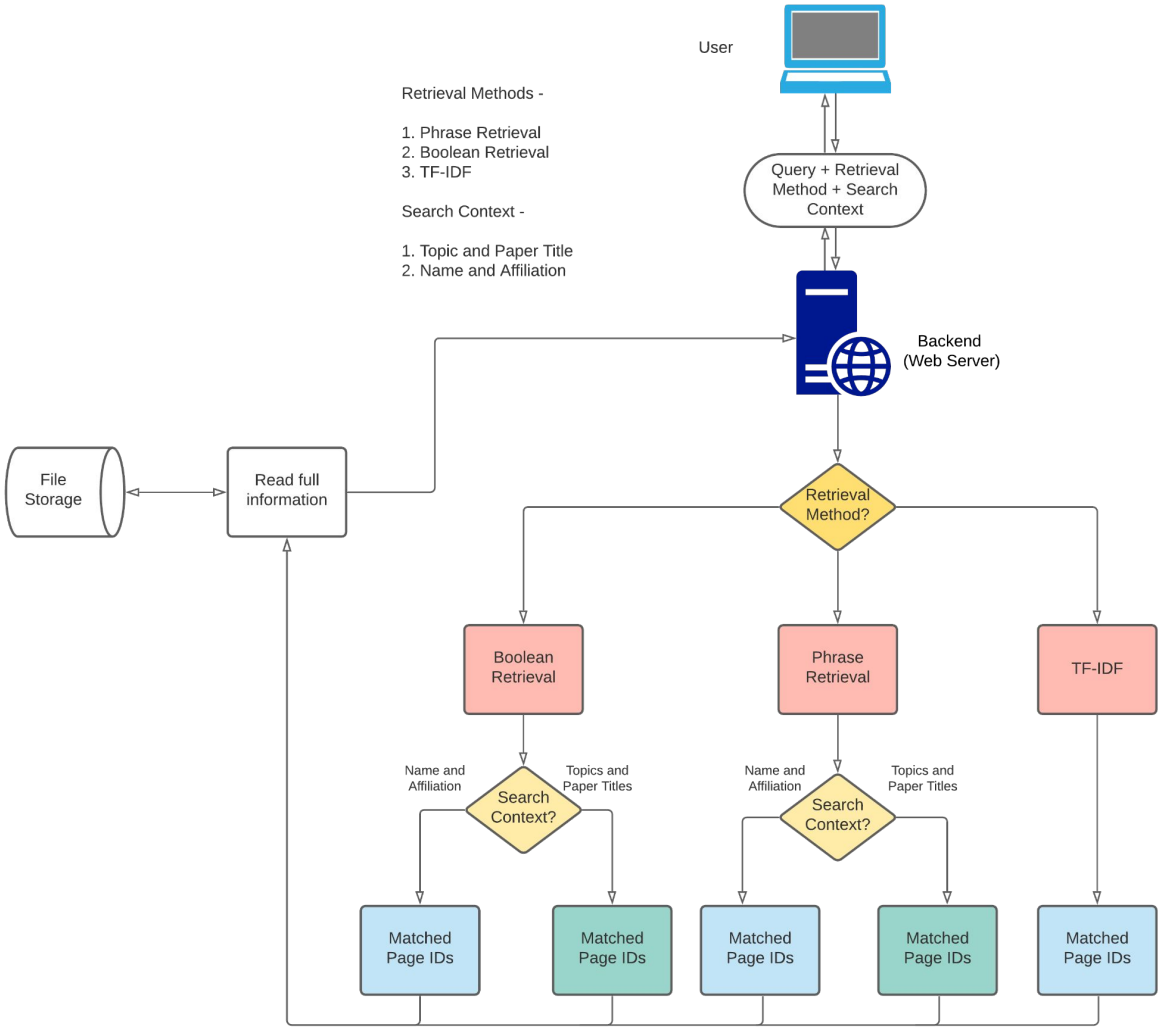
DEFAULT RANKING METRIC

- ➔ Linear combination of h-index, i10-index, no. of citations, h-index(last 5 yrs), i10-index(last 5 yrs), citations(last 5 yrs)



WEB SERVER

- Built using Flask.
- An interface for users to use the search engine.
- Hosted here.


User
Retrieval Methods -
1. Phrase Retrieval
2. Boolean Retrieval
3. TF-IDF
Search Context -
1. Topic and Paper Title
2. Name and Affiliation
Query + Retrieval Method + Search Context
Backend (Web Server)
File Storage
Read full information
Retrieval Method?
Boolean Retrieval
Phrase Retrieval
TF-IDF
Name and Affiliation
Topics and Paper Titles
Search Context?
Name and Affiliation
Topics and Paper Titles
Search Context?
Matched Page IDs
Matched Page IDs
Matched Page IDs
Matched Page IDs
Matched Page IDs

EVALUATION

MEDIAN RANK

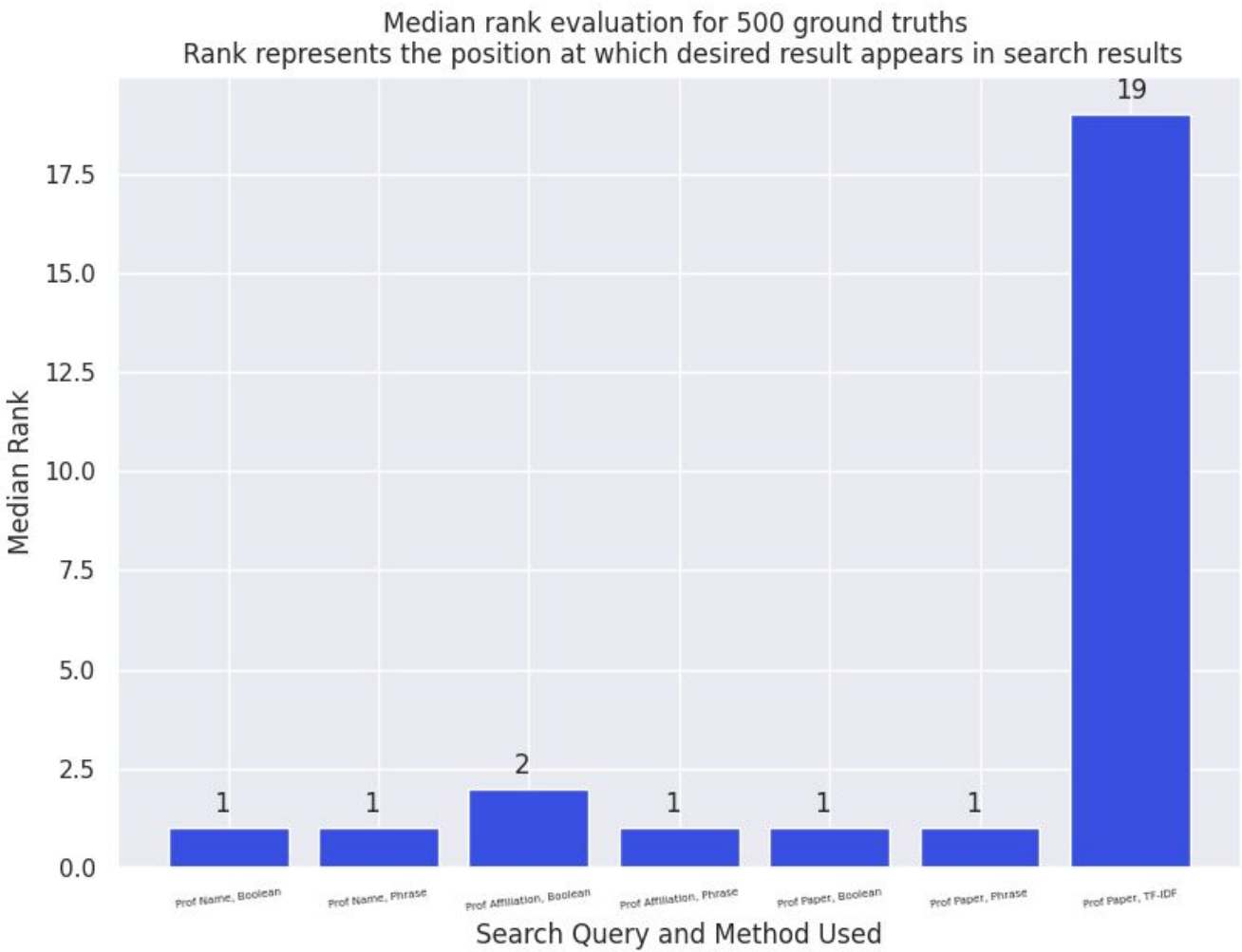
➔ Median of the ranks achieved by the ground truth in the search results, over a large number of searches.

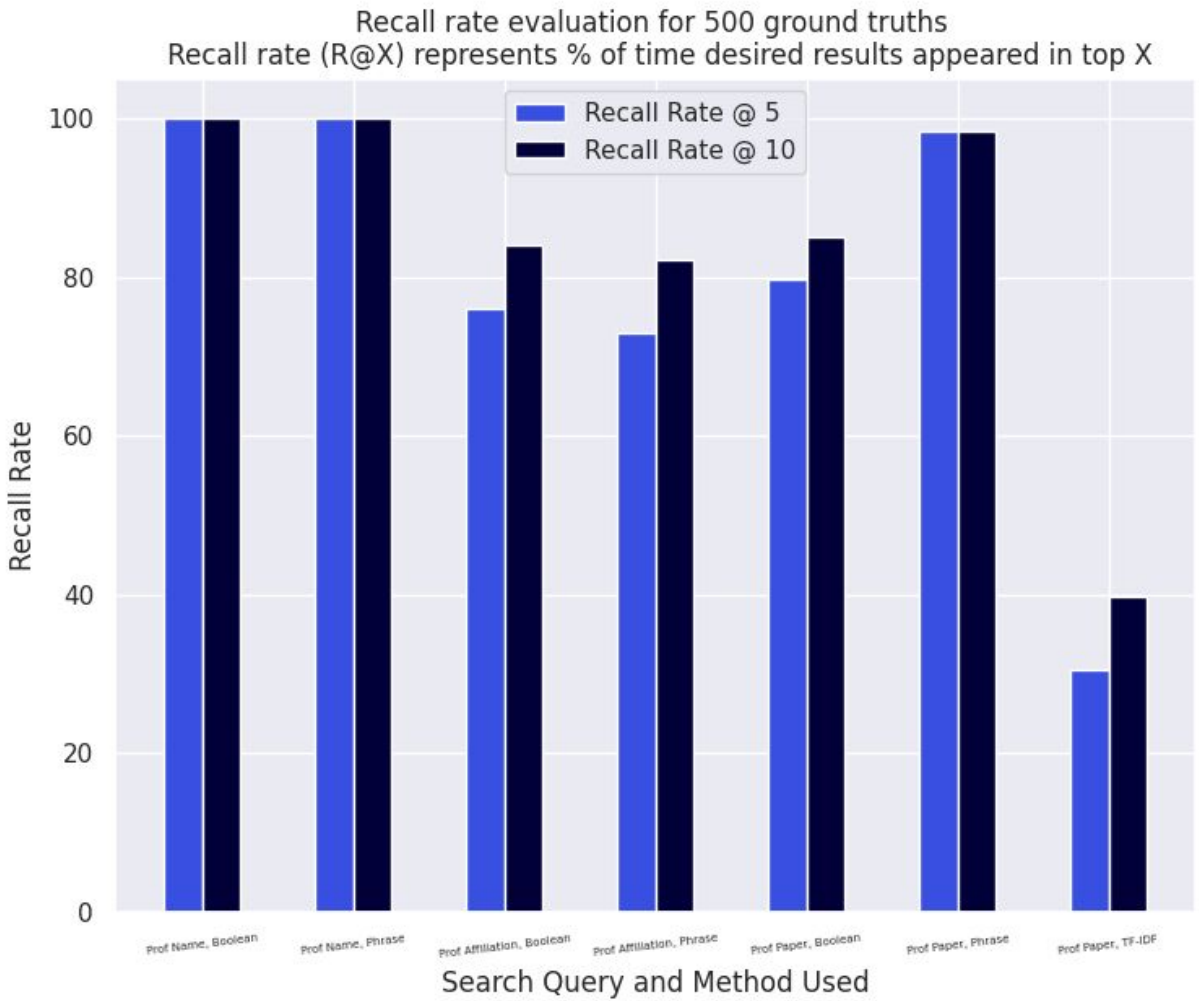
RECALL RATE, R@X

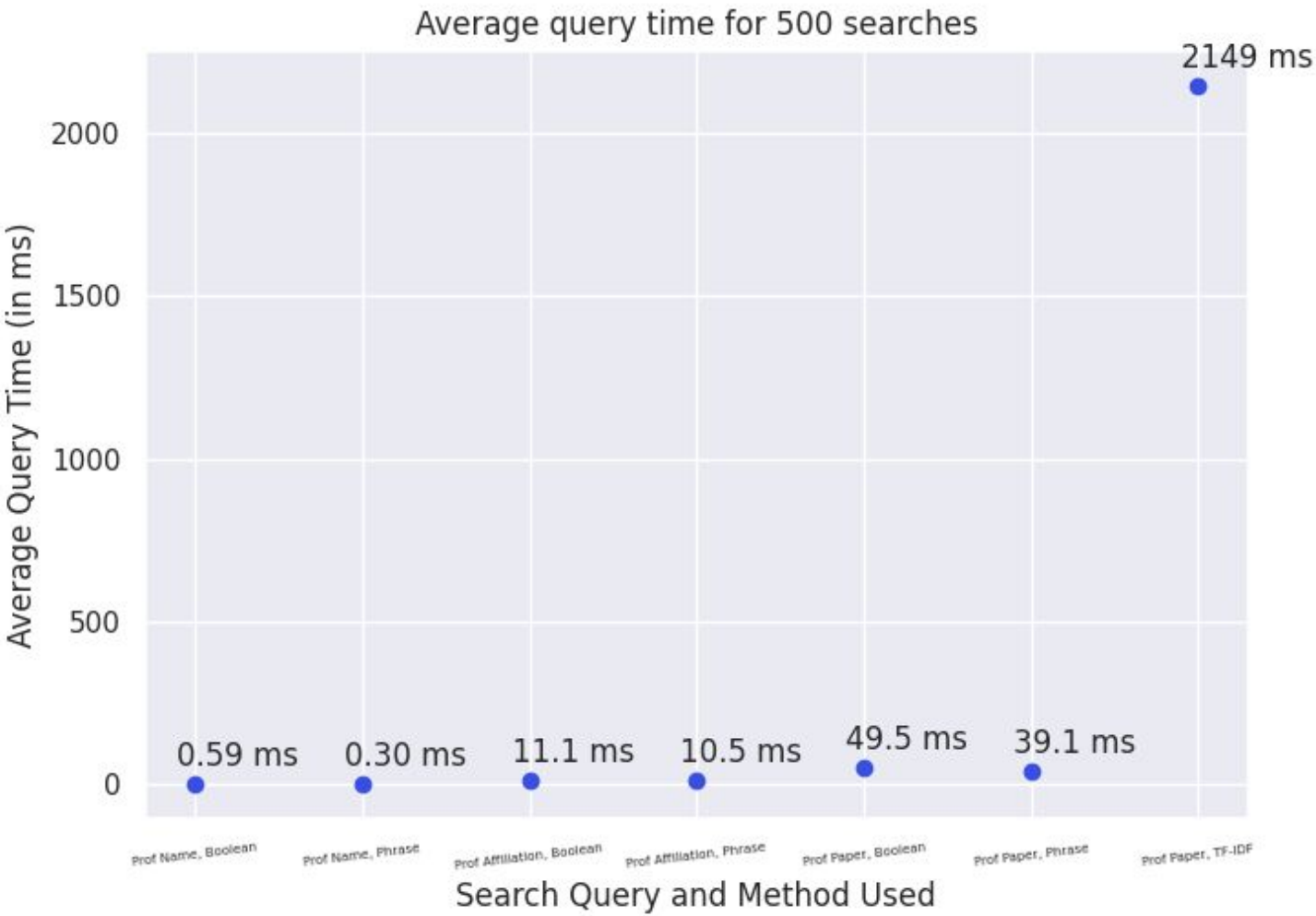
➔ Percentage of times that the ground truth appears in the top X results.

AVERAGE TIME PER QUERY

➔ Time taken by retrieval algorithms to return matched document IDs.


Median rank evaluation for 500 ground truths
Rank represents the position at which desired result appears in search results
Median Rank
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
1
1
2
1
1
1
19
Prof Name, Boolean
Prof Name, Phrase
Prof Affiliation, Boolean
Prof Affiliation, Phrase
Prof Paper, Boolean
Prof Paper, Phrase
Prof Paper, TF-IDF
Search Query and Method Used


Recall rate evaluation for 500 ground truths
Recall rate (R@X) represents % of time desired results appeared in top X
Recall Rate @ 5
Recall Rate @ 10
Recall Rate
100
80
60
40
20
0
Prof Name, Boolean
Prof Name, Phrase
Prof Affiliation, Boolean
Prof Affiliation, Phrase
Prof Paper, Boolean
Prof Paper, Phrase
Prof Paper, TF-IDF
Search Query and Method Used


Average query time for 500 searches
2149 ms
2000
1500
1000
500
0
Average Query Time (in ms)
0.59 ms
0.30 ms
11.1 ms
10.5 ms
49.5 ms
39.1 ms
Prof Name, Boolean
Prof Name, Phrase
Prof Affiliation, Boolean
Prof Affiliation, Phrase
Prof Paper, Boolean
Prof Paper, Phrase
Prof Paper, TF-IDF
Search Query and Method Used

DEMO

FUTURE WORK

FUTURE WORK

- ➔ Scraping more data
- ➔ Improving user experience
- ➔ Personalizing search results
- ➔ Performing user tests
- ➔ Using page ranking on citation data
- ➔ Learning the default ranking metric from user feedback

GitHub Repository: link
Web App: link

GROUP MEMBERS

Amey Kulkarni (18110016)
Chris Francis (18110041)
Nishikant Parmar (18110108)

ACKNOWLEDGEMENTS

Special thanks to Prof. Anirban Dasgupta for his valuable guidance and feedback.

Repo

Web App