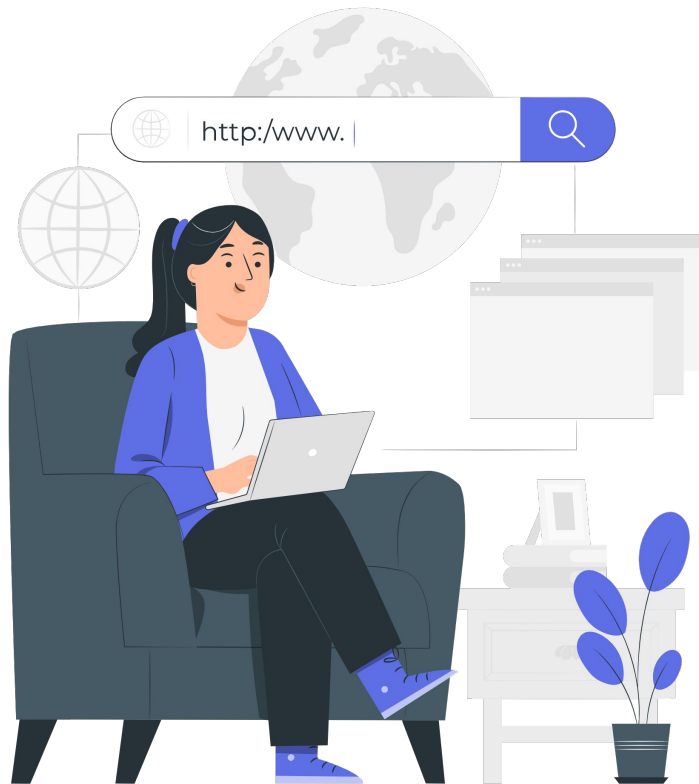


# ACAD SEARCH

A crowd-sourced search engine  
to find leading professors



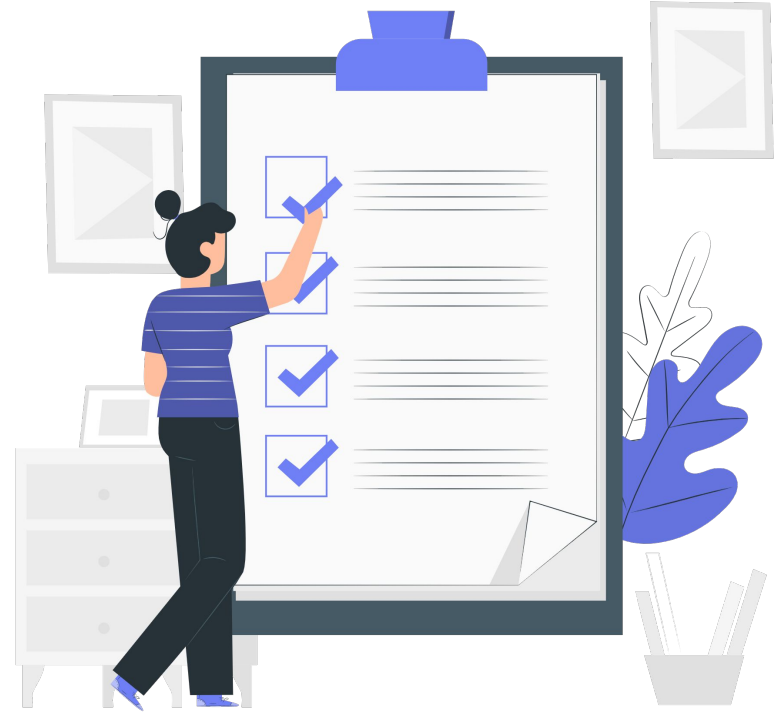


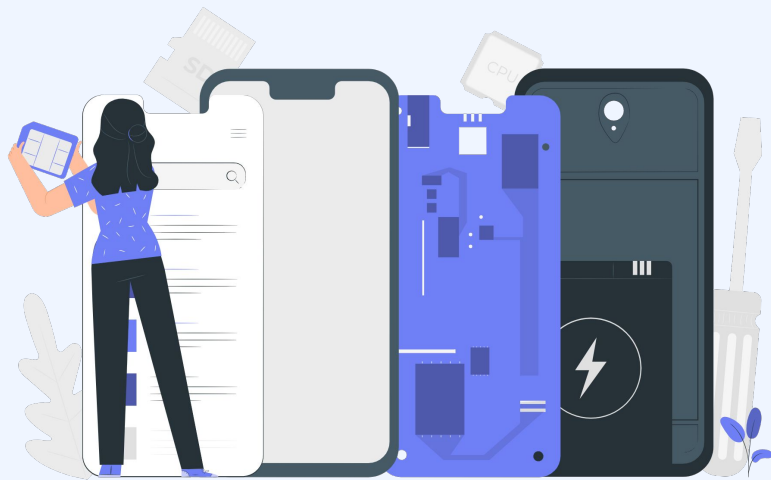
# MOTIVATION

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

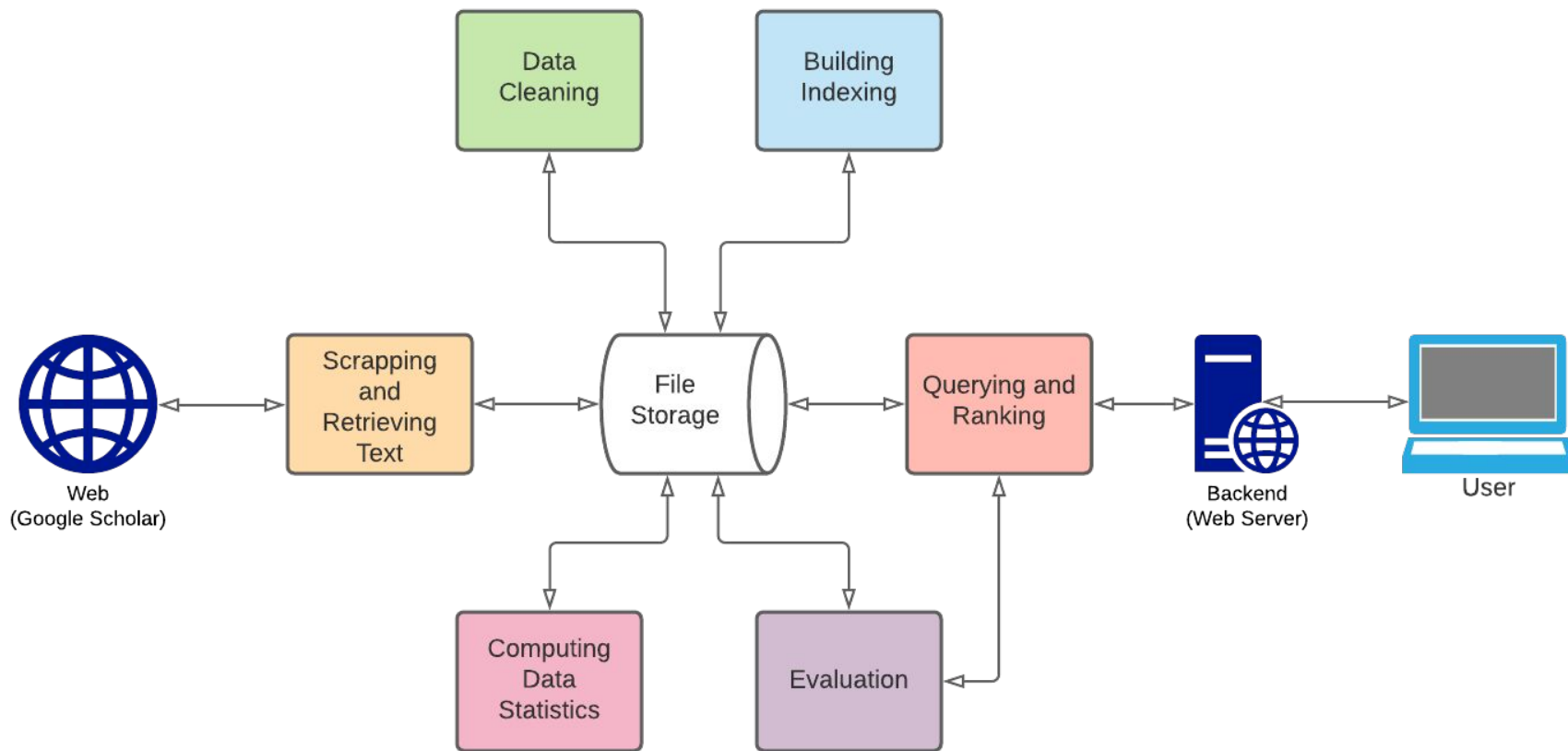
# OBJECTIVES

- 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 COLLECTION

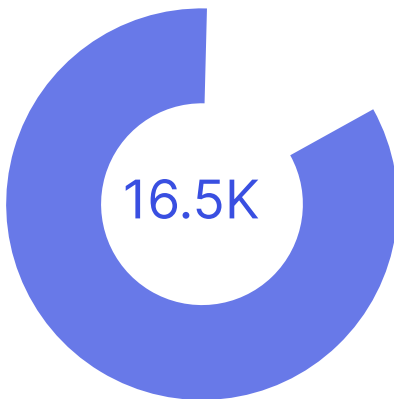


# SOURCE

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.
- 

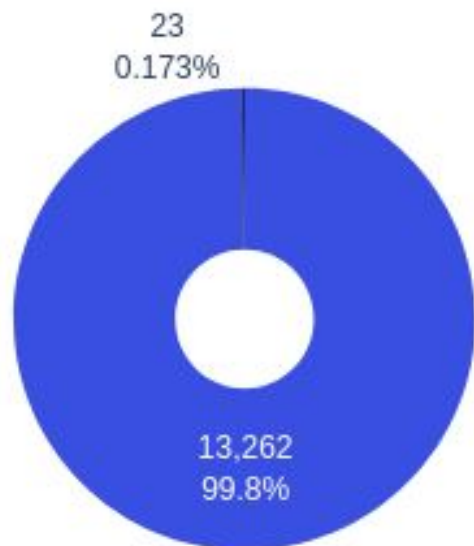
# DATA STATISTICS



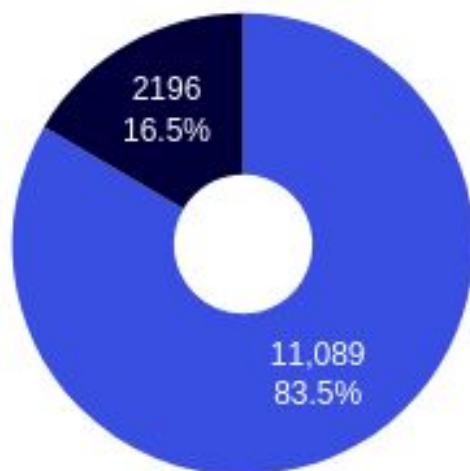
# BASICS

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

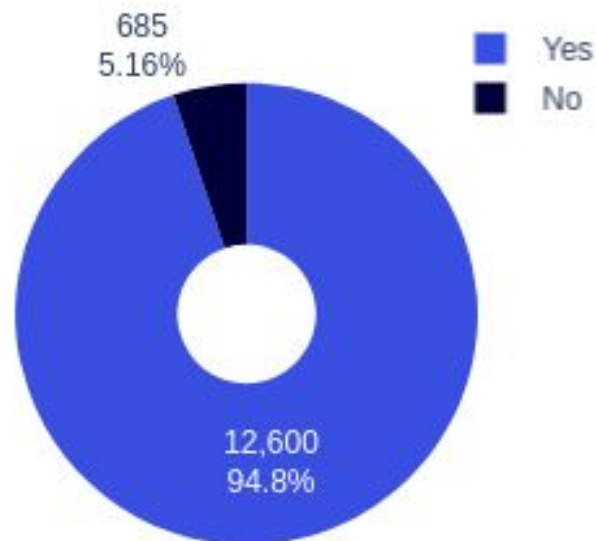
Affiliation provided?



Email verified?

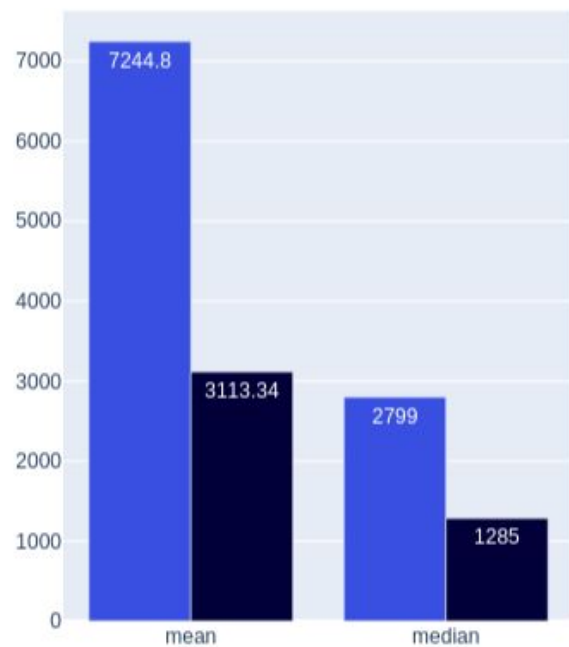


Homepage provided?

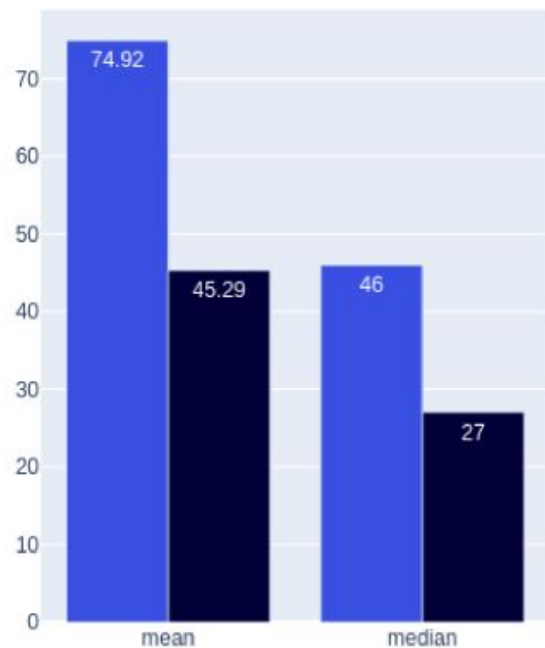


■ Yes  
■ No

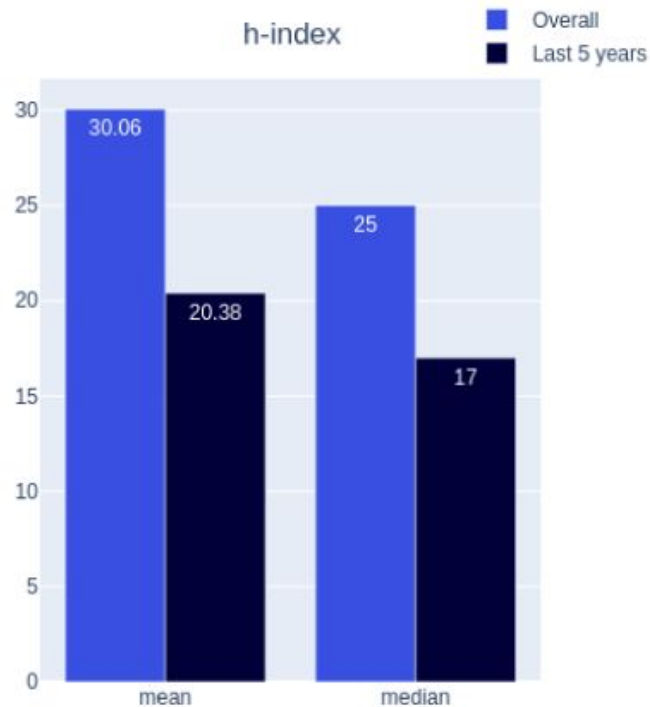
No. of citations



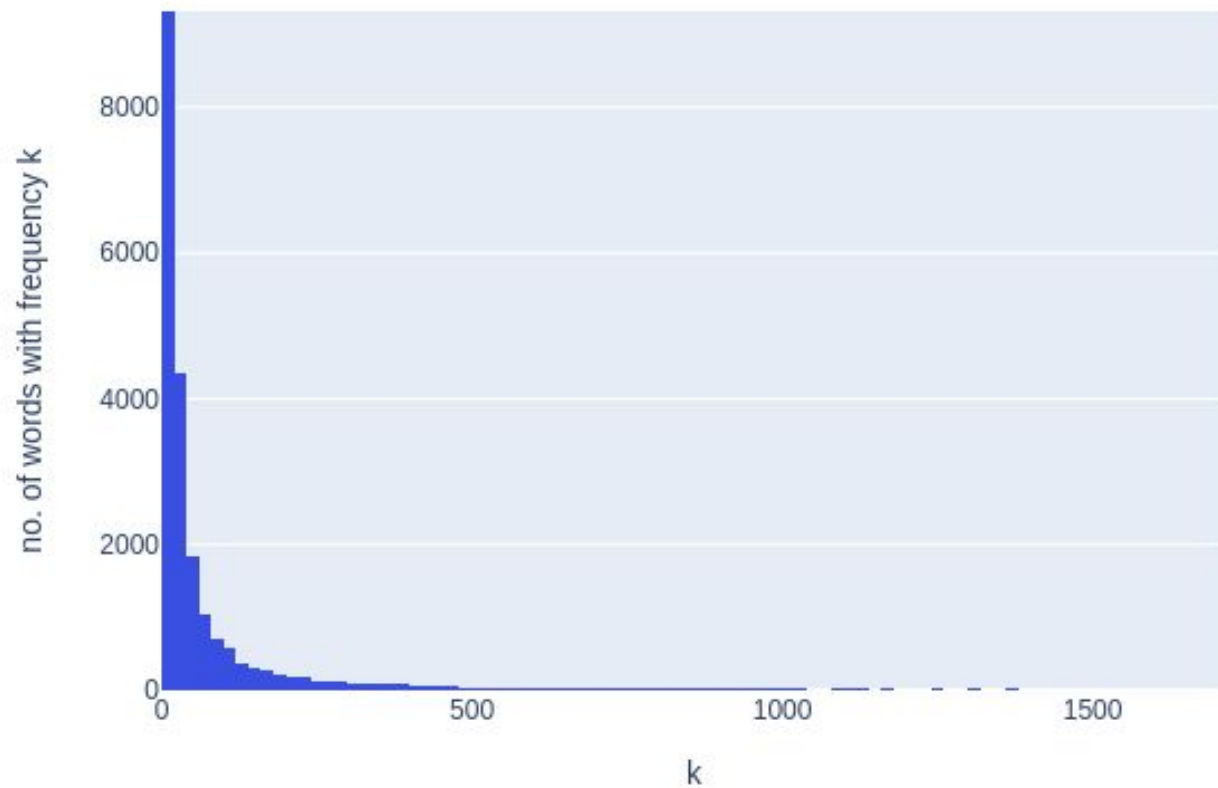
i10-index



h-index



No. of words with frequency  $k$  as a function of  $k$







# 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.

# 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)

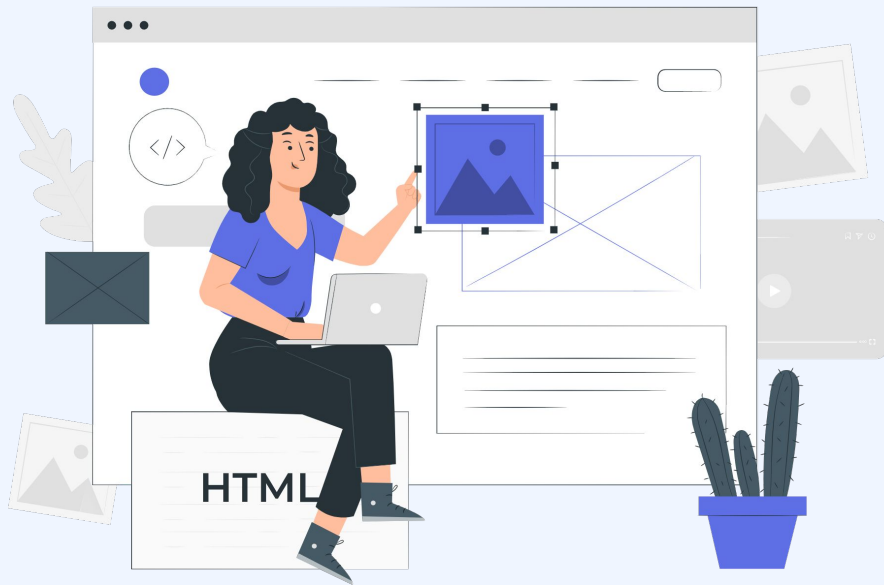
- $TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D)$
- TF: measure of how frequently a word/term appears in a document, normalised by the document size.
- $TF(t, d) = \text{no. of times } t \text{ appears in } d \div \text{total no. of words in } d$
- IDF: measure of how much information a word provides,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

# **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.

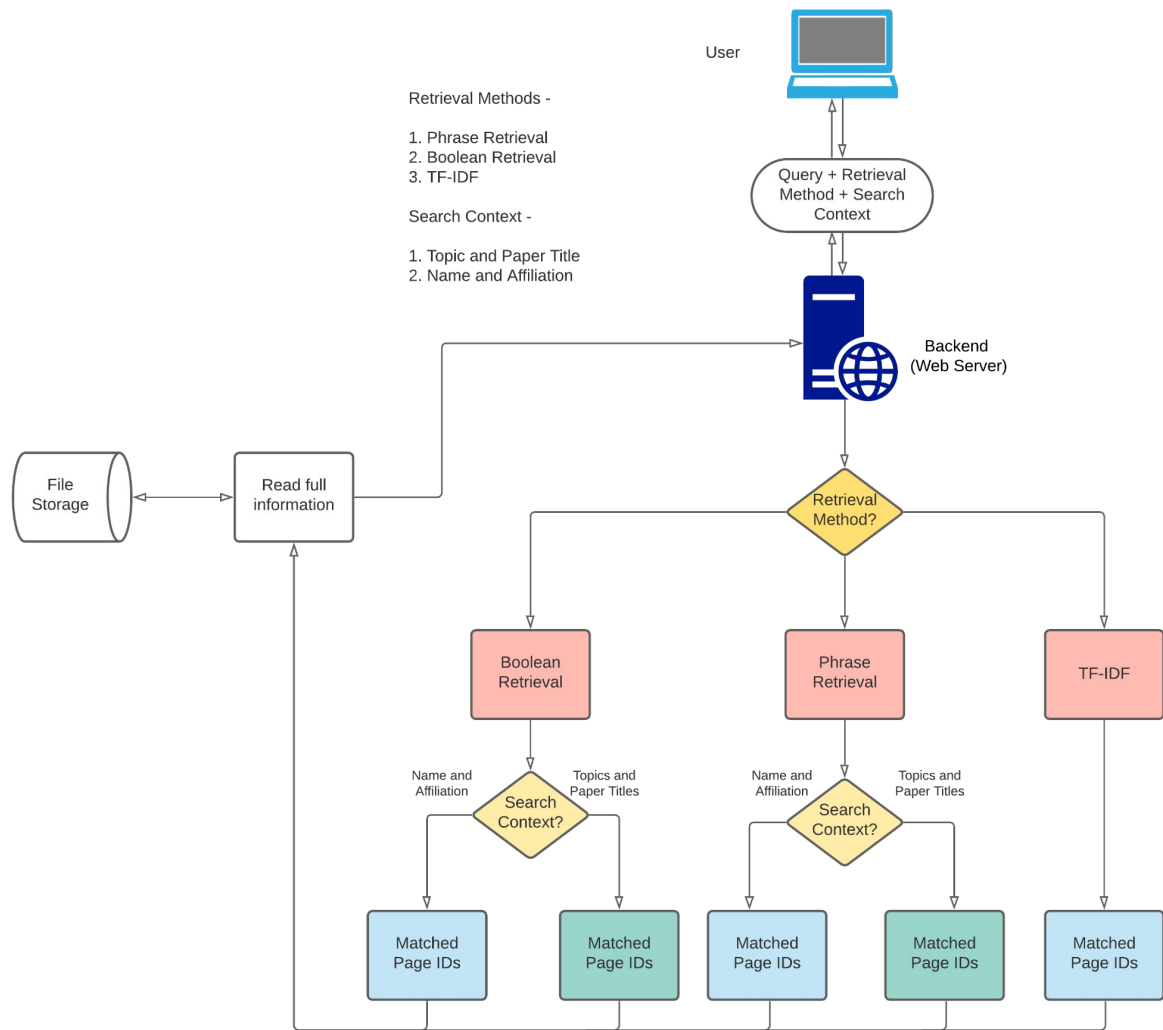
# **DEFAULT RANKING METRIC**

- 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](#).





# EVALUATION



## MEDIAN RANK

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

## RECALL RATE, $R@X$

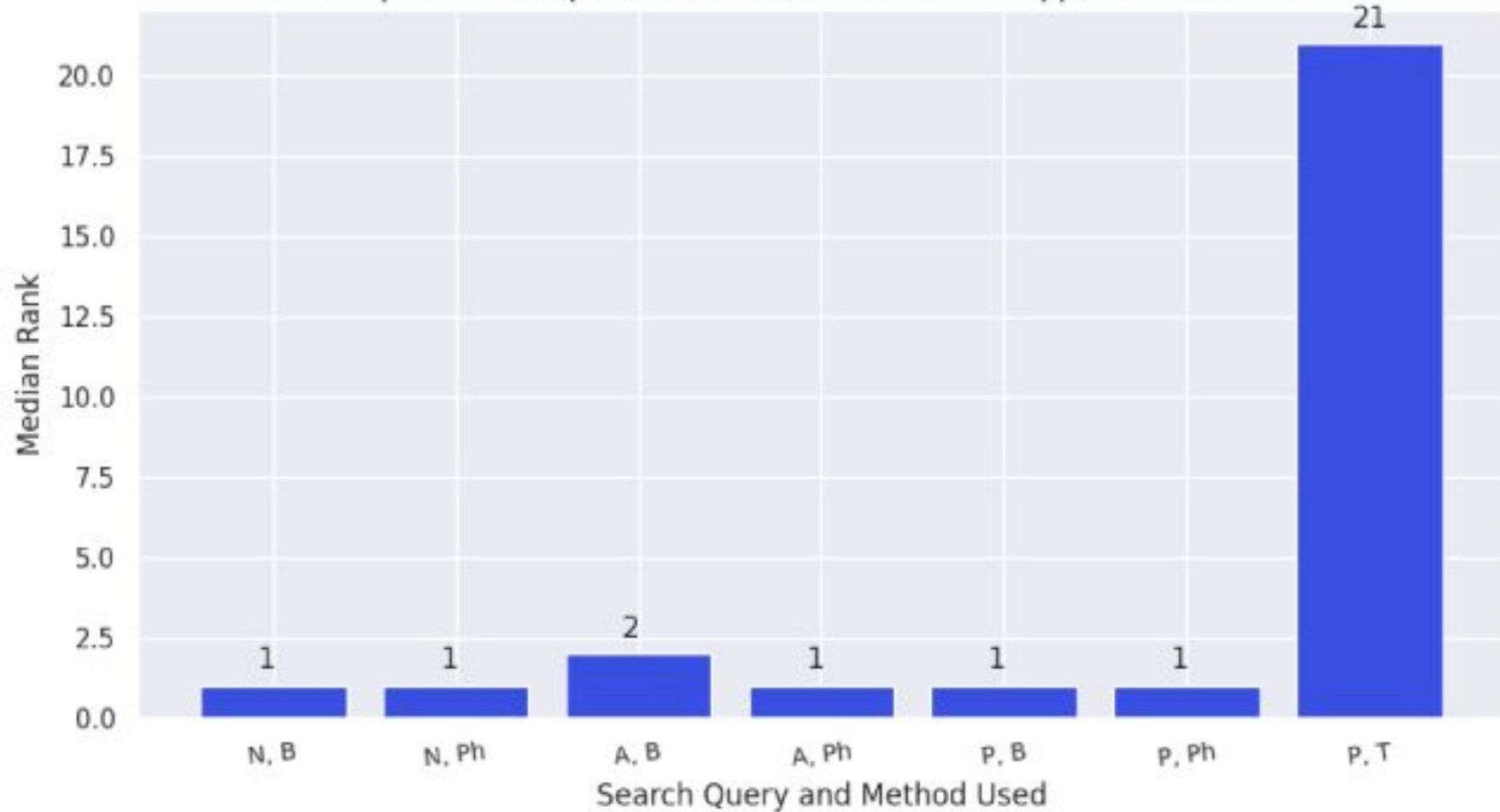
- Percentage of times that the ground truth appears in the top  $X$  results.

## AVERAGE TIME PER QUERY

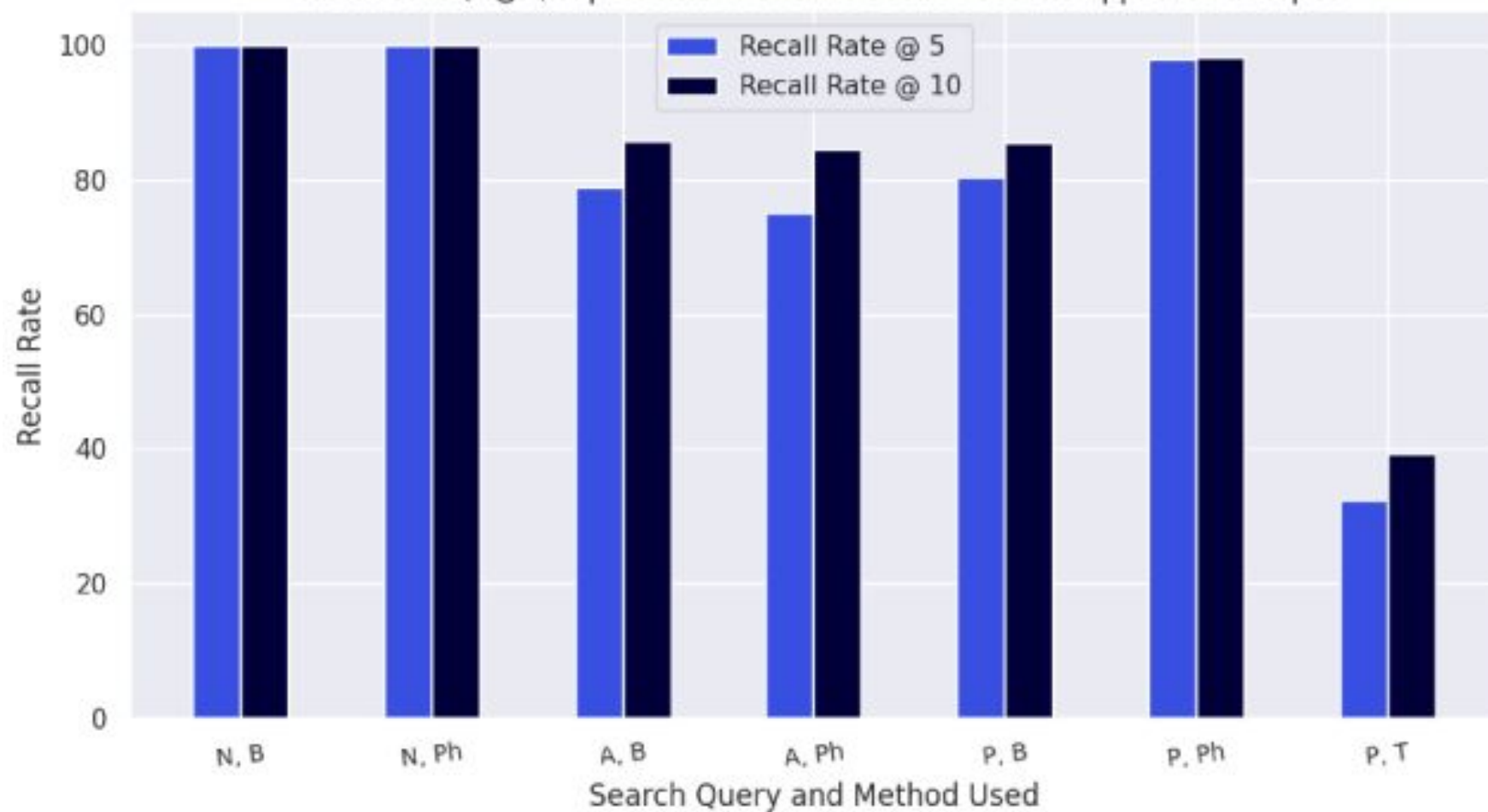
- Time taken by retrieval algorithms to return matched document IDs.

# Median rank evaluation for 500 ground truths

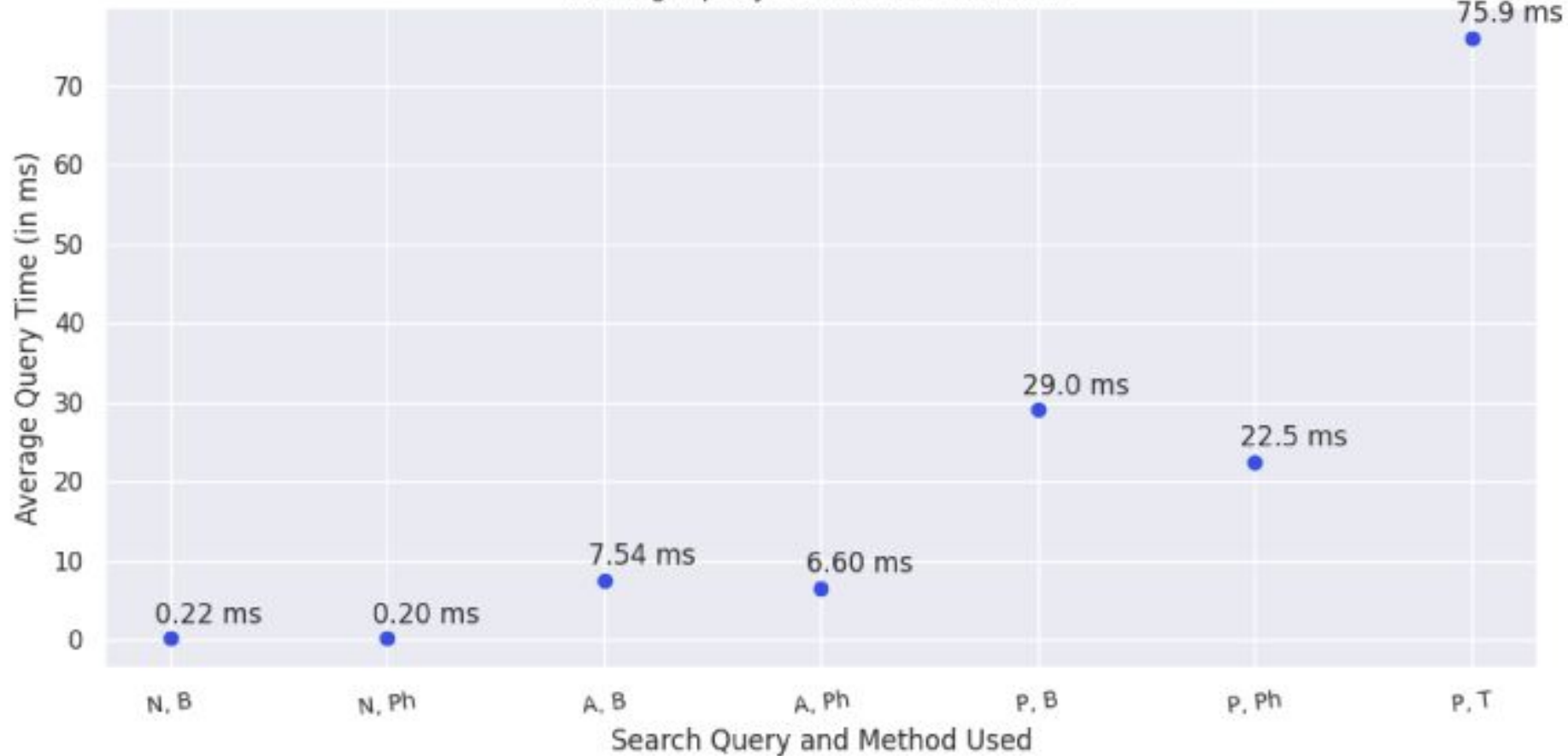
Rank represents the position at which desired result appears in search results



Recall rate evaluation for 500 ground truths  
Recall rate ( $R@X$ ) represents % of time desired results appeared in top X



Average query time for 500 searches





# DEMO

[Link to Web App](#)

# 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