

# Scriburg: A Configurable Preferential Web Search Engine

Master's thesis

Alhajras Algdairy

Albert-Ludwigs-Universität Freiburg



**UNI  
FREIBURG**

# Contents

1. Problem
2. Solution
3. Evaluation

# Problem: Examples

## Example1: Chatbot



- Q&A (python)
- Requires data
- Stack OverFlow, Reddit, or Quora

## Example2: Market Research



- Cheapest product
- Requires data
- Amazon, Ebay, or Alibaba

How to optimally crawl data from the internet for a given set of websites?

# Problem: Classical Solutions

## Commercial Search Engine



- Sponsored
- Customization ranking
- Third-party dependency

## Commercial Tools

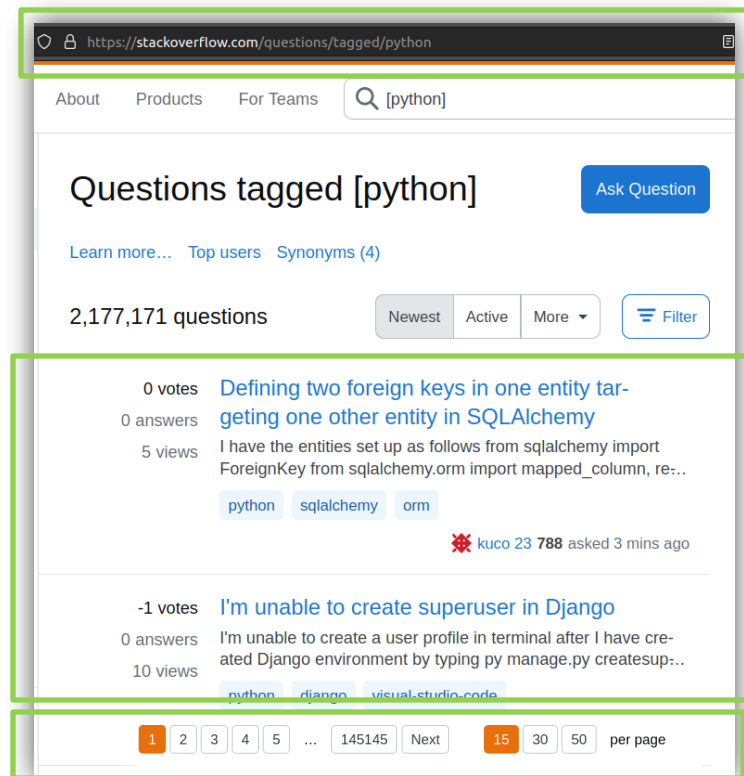


- Limited configurations
- Expensive
- Clouds
- No indexing

**Configurable, generic, easy to use, free search engine**

# Problem: Formal Problem Definition

1. Given a set of URLs  $U = \{U_1, U_2, U_3, \dots U_n\}$
2. Find the next URLs set  $\hat{U} = \{\hat{U}_1, \hat{U}_2, \hat{U}_3, \dots \hat{U}_n\}$
3.  $\forall x \in U$  download all documents  $D = \{D_1, D_2, D_3, \dots D_n\}$
4. Given a user query  $q$ , find relevant documents  $R$ ,  $R \subset D$
5. User-friendly configurable interface



# Problem

Questions?

# Solution: Scriburg

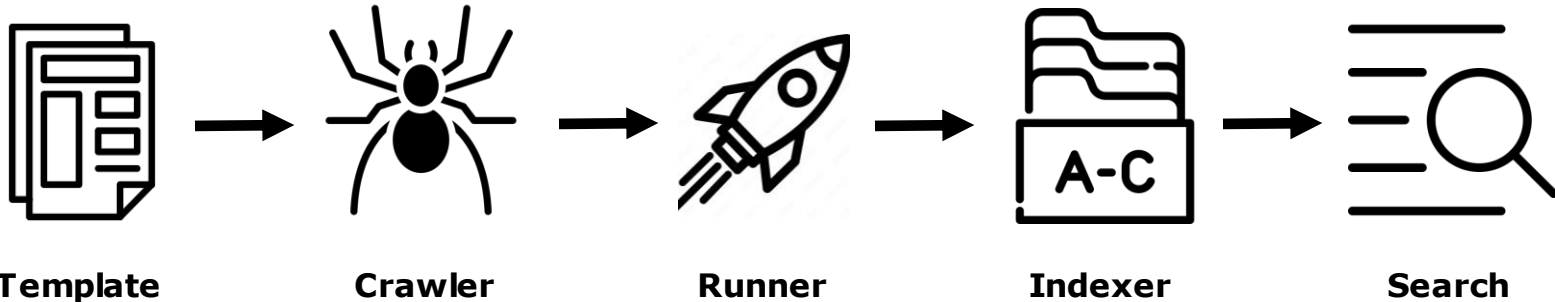
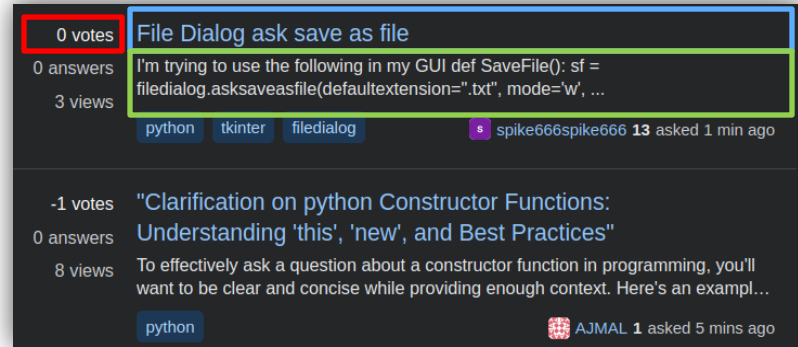
- Free open-source search engine
- No programming knowledge needed (GUI)
- Configurable (crawling, indexing and ranking)
- Scalable (vertically and horizontally)

The screenshot displays the Scriburg web interface. At the top, there is a navigation bar with links: Documentation, Templates, Crawlers, Runners, Indexers, and Search. A user profile icon is in the top right corner. The main content area is titled 'Runners' and includes a '+ Create a runner' button. Below this is a table with the following columns: ID, Status, Crawler, Progress, and Actions.

ID ↑↓	Status ↑↓	Crawler ↑↓	Progress ↑↓	Actions
Stack OverFlow #134	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	Stack OverFlow #10	Started at: 3 months ago / Completed at: 3 months ago #Documents: 2415 Current URL: <a href="https://stackoverflow.com/questions/tagged/python?tab=newest">stackoverflow.com/questions/tagged/python?tab=newest</a>	<div>⋮</div> <div>📄</div>
Crawler-test-15-09 #133	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	crawler-test #4	Started at: 3 months ago / Completed at: 3 months ago #Documents: 578 Current URL: <a href="https://crawler-test.com/urls/with_session_id?sessionID=13...">crawler-test.com/urls/with_session_id?sessionID=13...</a>	<div>⋮</div> <div>📄</div>
Uni ranking Local #132	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	Uni ranking #9	Started at: 4 months ago / Completed at: 3 months ago #Documents: 2389 Current URL: <a href="https://timeshighereducation.com/world-university-rankings...">timeshighereducation.com/world-university-rankings...</a>	<div>⋮</div> <div>📄</div>

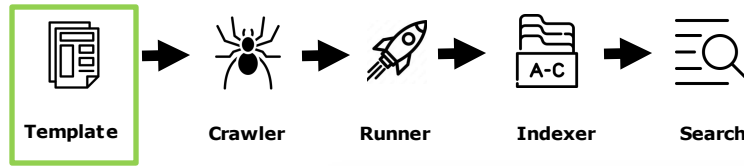
# Solution: Demo

- Objective:
  - Crawl and index Stack OverFlow questions
  - Title, Summary, Votes.
- Workflow:
  - Template (Title, Summary, Votes)
  - Crawler (Template, Seed URL, Threads)
  - Runner (Crawler)
  - Indexer (Title, Summary)
  - Search (Indexer)

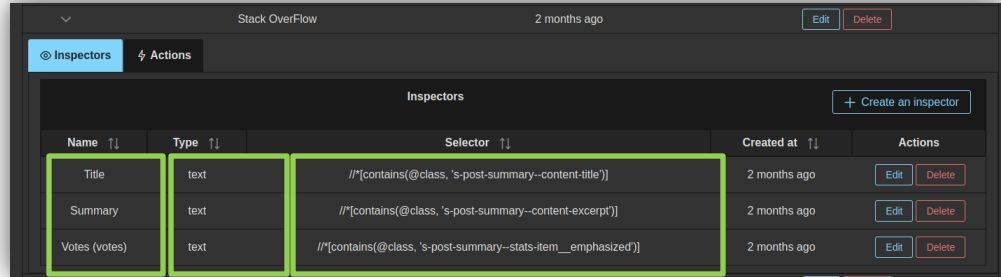




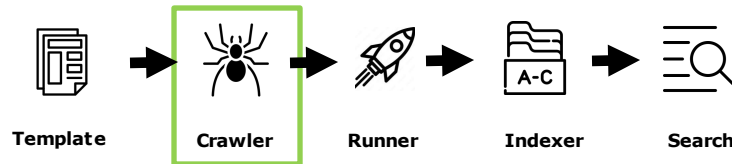
# Solution



- Inspectors (fields)
  - Title, Summary and Votes
  - Image, Text, or Link
  - XPath



# Solution



- Configurations

- Mandatory
- Advanced

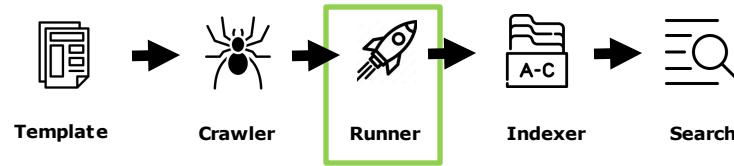


Name*	A user-defined identifier for the crawler
Template*	The blueprint for collecting documents with
Seed URL*	The starting point of crawling



Max Pages	The upper limit for the number of pages to be visited
Max Docs	The maximum number of documents to be collected
Max Depth	Maximum jumps between pages (crawling depth)
Robots.txt	The URL where the <code>robots.txt</code> file can be located
Threads	Number of threads used in the crawling process
Pagination	Scope to collect the following URLs
Excluded URLs	URLs that the crawler must refrain from visiting
Walltime (ms)	Sets the duration for which the crawler should continue crawling
Show Browser	Deactivate the headless mode in Selenium

# Solution

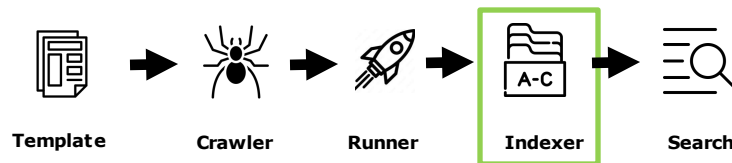


- Runners / Jobs / Tasks
  - Parallel
  - Different machines
  - Status
  - Progress


The screenshot shows the 'Runners' section of a web application. The interface includes a navigation bar with links for Documentation, Templates, Crawlers, Runners, Indexers, and Search. A '+ Create a runner' button is located in the top right. The main content area displays a table of runners with columns for ID, Status, Crawler, Progress, and Actions. The first runner, 'Stack OverFlow #134', is highlighted with a green box around its status indicators and progress details.

ID ↑↓	Status ↑↓	Crawler ↑↓	Progress ↑↓	Actions
Stack OverFlow #134	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	Stack OverFlow #10	Started at: 2 months ago / Completed at: 2 months ago #Documents: 2415 Current URL: <a href="#">stackoverflow.com/questions/tagged/python?tab=newe...</a>	<div><div></div><div></div></div>
Crawler-test-15-09 #133	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	crawler-test #4	Started at: 3 months ago / Completed at: 2 months ago #Documents: 578 Current URL: <a href="#">crawler-test.com/urls/with_session_id?sessionID=t3...</a>	<div><div></div><div></div></div>
Uni ranking Local #132	<div><div>New</div><div>Running</div><div>Exit</div><div>Completed</div></div>	Uni ranking #9	Started at: 4 months ago / Completed at: 2 months ago #Documents: 2389 Current URL: <a href="#">timeshighereducation.com/world-university-rankings...</a>	<div><div></div><div></div></div>

# Solution



- Configurations
  - Mandatory
    - Inspectors: Title and Summary
  - Advanced
- Overview
  - Status
  - Inspectors



Name*	A user-defined identifier for the indexer
Inspectors*	Checklist of all the available inspectors used by the crawlers.
b Parameter	b parameter for the BM25 formula.
k Parameter	k parameter for the BM25 formula.
Stop Words List	List of words that should be excluded during the indexing process.
Small Words Threshold	The threshold of which the word can be considered small and will be skipped from the indexing process.
Words Weight List	Boost some words by giving them weight, e.g. "Freiburg=5" will add more 5 points to the score when the "Freiburg" word is found.
Boosting Formula	This formula result will be added to the final score. It uses inspectors variable.
Dictionary File Name	The dictionary file name that helps the suggestions list by using synonyms.
Use Synonyms	Enable using synonyms in the suggestions list. For example, typing "USA" will result in "United States of America".
Q-Gram	The $n$ value of the q-grams for creating a q-gram inverted index, see Section 3.5.1

Stack Overflow  
Created: 2 months ago Completed: a month ago

New cross indexer  
Created: 4 months ago Completed: a month ago

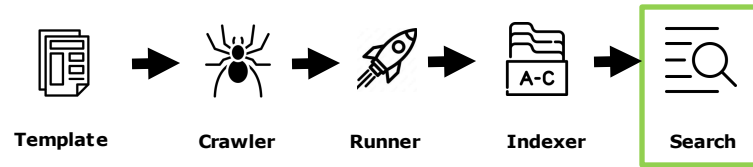
• Title (Stack OverFlow)  
• Summary (Stack OverFlow)

Start indexing Edit Delete

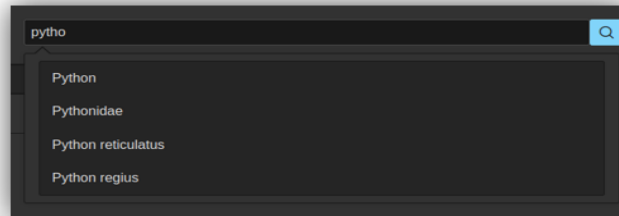
• Title (Douglas flat list)

Start indexing Edit Delete

# Solution



- Search Engine Result Page (SERP)
  - Auto suggestions
  - Dynamic search result



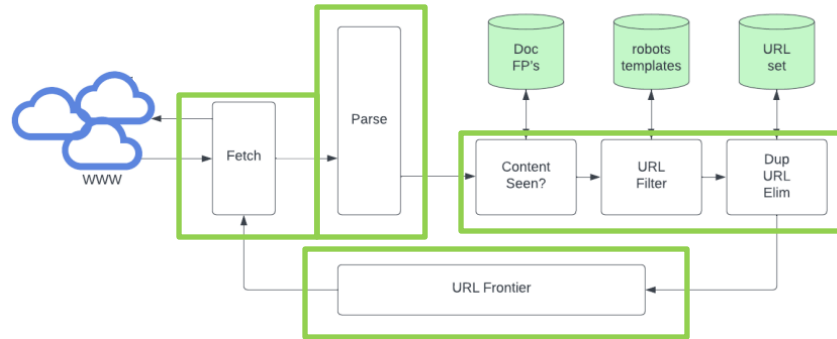
Summary	Title	Votes
When creating a <code>virtual environment</code> using the <code>virtualenv</code> module, using python 3.11 as default in the Path, in WINDOWS 10, I encounter the following Creation of <code>virtual environment</code> with python 3.9 version is executed successfully but python version 2.7 or python version 2.2 shows error		0 votes
pyenv <code>virtualenv</code> puts <code>virtual environment</code> somewhere besides the directory they were called in. This means that when I come back to a project months later, I'm left to guess at what <code>virtual</code> ...	How to keep track of which pyenv <code>virtualenv</code> goes with my project?	-3 votes
Matplotlib is not working in Jupyter Notebook on Mac. As can be seen below, I think it is installed in the <code>environment</code> . I have tried upgrading both <code>Jupyter</code> and <code>matplotlib</code> and am able to use ...	Matplotlib is not working in Jupyter Notebook on Mac	0 votes
Here's my <code>setup</code> a Mac, running OS X Tiger. Windows XP running in a <code>virtual</code> machine (Parallels). Windows XP has my Mac home directory mapped as a network drive. I have two files in a directory of ...	Bizarre python ImportError	3 votes

Solution: Demo

Questions?

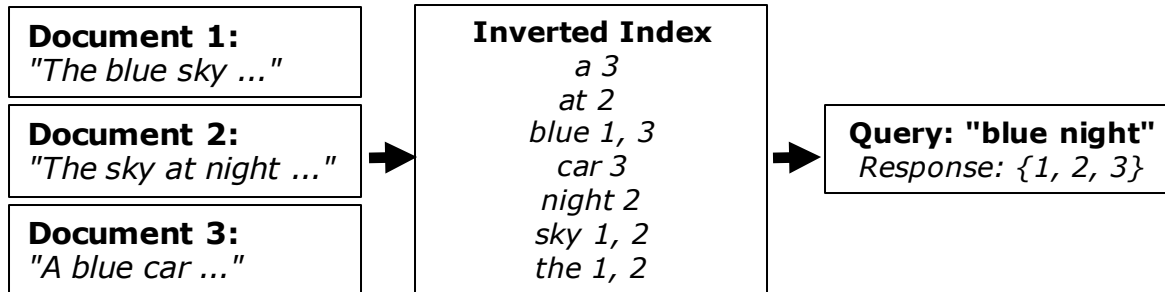
# Solution: Crawlers Overview

- URL Frontier (Stack or Queue)
- Fetch
- Parse
- Filter
  - Content seen
  - URL filter (Duplicated, cross-origin)



# Solution: Indexers

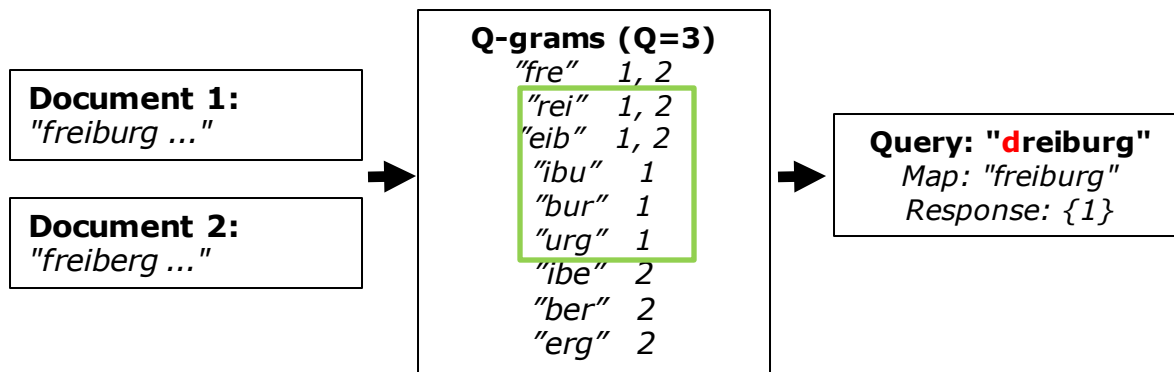
- Inverted index





# Solution: Fuzzy Search with Q-grams

- Forgive misspelling
  - Generate Q-grams for each token
  - $Q3(\text{"freiburg"}) = \{\text{"fre"}, \text{"rei"}, \text{"eib"}, \text{"ibu"}, \text{"bur"}, \text{"urg"}\}$



# Solution: Ranking

$$Total(d,q) = \mathbf{BM25} + \mathbf{QM} + TW + BF$$

- BM25
  - $N$  : total number of documents.
  - $tf$  : term frequency, the number of times a word occurs in a document.
  - $df$  : document frequency, the number of documents containing a particular word.
  - $DL$ : document length (number of words).
  - $AVDL$ : average document length (number of words).
  - $b$ : controls documents normalization.
  - $k$ : controls the impact of terms frequency.
- Query Matching Score (QM)
  - Document contains more shared tokens with the query ranks higher
- Optional: Token Wight (TW) List, "*freiburg=0.5*"
- Optional: Boosting Formula (BF), "*log(votes)*"

$$score(d,q) = \widehat{tf} \cdot \log_2 \frac{N}{df}$$

$$\widehat{tf} = \frac{tf \cdot (k + 1)}{k \cdot \alpha + tf}$$

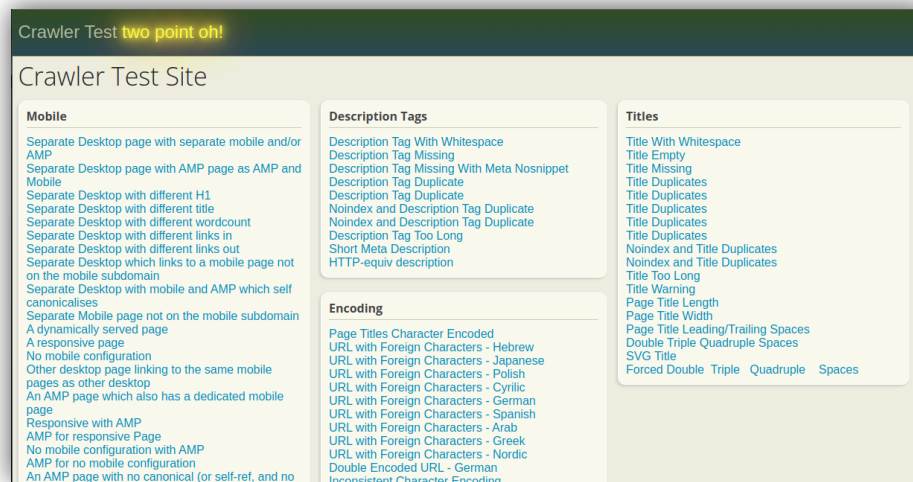
$$\alpha = 1 - b + \frac{b \cdot DL}{AVDL}$$

# Solution

Questions?

# Evaluation: Crawler

- crawler-test.com
  - Different HTTP status
  - Different content (Dynamic loading)
  - Robots.txt protocol
  - Redirection and more edge cases



Site	Pages	Scriburg (1 thread, 1 node)	ParseHub (free)	Factor
crawler-test	402	11 min	40 min (200 page and crashed)	7.27

## Evaluation: Crawler

Site	Pages	Scriburg (1 thread, 1 node)	ParseHub (free)	Factor
Uni ranking	94	<b>5.85</b> min	18.8 min	<b>3.2</b>
Stack OverFlow	100	<b>5.9</b> min	20 min	<b>3.3</b>
Douglas	7	<b>6.58</b> min	Crashed	xxx

Do not get too greedy! Using four threads resulted in DoS!

## Evaluation: Indexer

- Precision at  $k = 5$  (**P@k**)

$$P@k = \frac{|Q_{valid}(q) \cap Q_{result}^k(q)|}{k}$$

$Q_{valid}$ : Set of expected documents

$Q_{result}^k$ : Set of top  $k$  documents returned by Scriburg

- Average Precision (**AP**)

$$AP = \frac{\sum_{i=1}^n P@_{r_i}}{n}$$

$r_1, r_2, \dots, r_n$ : List of positions at which relevant documents from  $Q_{valid}$  appear in  $Q_{result}$

# Evaluation: Indexer

- Stack Overflow dataset

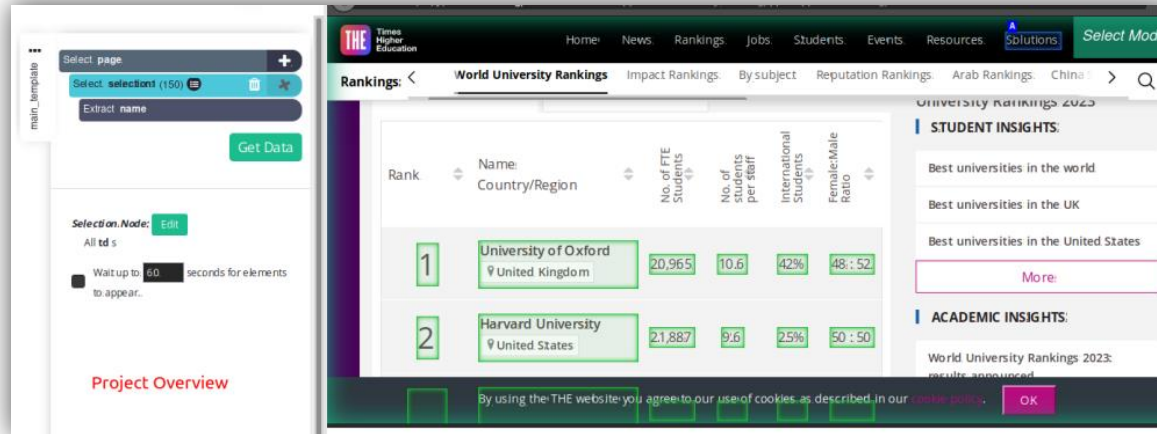
File Size	1.4 MB
Entries Count	2,415
Words Count	108,122
Fields	Title, Summary, Votes

- Benchmark is made of **6** queries with an average of **6 relevant** documents per query
- Configurations and result:

Inspectors*	Title, Summary			
BM25 Parameters	$b = 0.1, k = 0.81$			
Stop Words	None			
Small Words Threshold	2			
Q-gram	3			
Boosting Formula	None			
Result	MP@5 0.8	MP@R 0.84	MAP 0.89	Time(s) 0.31

# Evaluation: User Experience

Aspect	Scriburg	ParseHub
Workflow	2.5/5	3.5/5
Robustness	3.5/5	2.5/5
Configuration	3.5/5	2/5





# Future Work

- Enhance Scriburg by using an automatic CSS detection.
- Investigate on how to update the documents periodically (detect changes).
- Experiment with different languages (Not only English).
- Experiment with different websites (social media).
- Using IP Rotation.

# Evaluation

Thank you for your attention!