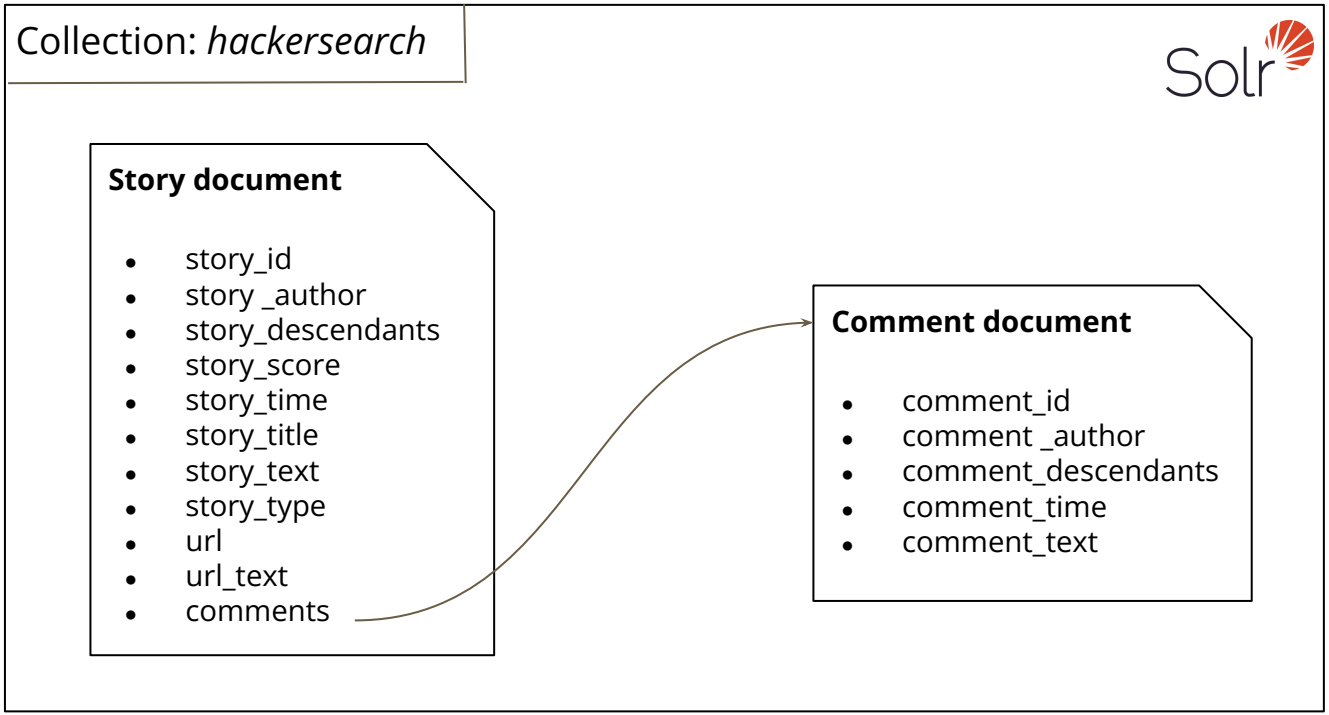
HackerSearch

Information Retrieval

Ana Barros - up201806593
João Costa - up201806560
João Martins - up201806436

Documents


Collection: *hackersearch*
Solr
Story document
story_id
story _author
story_descendants
story_score
story_time
story_title
story_text
story_type
url
url_text
comments
Comment document
comment_id
comment _author
comment_descendants
comment_time
comment_text

Indexing process

Field	Type	Indexed
store_id	int	false
story_author	singleToken	false
story_descendants	int	true
story_score	int	true
story_time	date	true
story_title	text	true
story_text	text	true
story_type	category	true
url	url	true
url_text	text	true
comments.comment_id	int	false
comments.comment_author	singleToken	false
comments.comment_descendants	int	false
comments.comment_time	date	false
comments.comment_text	text	true
newssite_filter	newsfilter	true
search	text	true

Table 1: Document Fields

Source	Dest
url	newssite_filter
story_type	search
story_title	search
story_text	search
url	search
url_text	search
comments.comment_text	search

Table 2: Copy Fields

Schema Definition

Field-type: title

The type of the stories' titles. Very similar to the text type.

- Tokenizer:
  - Splits on whitespace, punctuation and @ characters.
- Filters:
  - Non ASCII characters are converted to their ASCII equivalents.
  - Characters are converted to lower case.
  - Synonym filter to increase the detection of related topic/words.
  - English words reduce to their root form and possessive cases disappear.
- Stream treatment with this field-type is the same during index and query time.

Schema Definition

Field-type: text

The type of all the text fields (story text, URL content, comment text, and search).

- Tokenizer
  - URLs and email addresses are kept as a full token.
  - Splits on whitespace, punctuation, and hyphens.
- Filters:
  - Remove HTML tags (Before Tokenizer)
    - `<a href=https://solr.apache.org>Solr</a>` becomes Solr
  - Non ASCII characters are converted to their ASCII equivalents.
  - Characters are converted to lower case.
  - Synonym filter to increase the detection of related topic/words.
  - Removal of stop words (e.g. the, a, an, etc...).
  - English words are reduced to their root form and possessive cases disappear.
- Stream treatment with this field-type is the same during index and query time.

Schema Definition

Field-type: url

The type of the stories' URLs. Allows the user to focus on their search on specific domains.

- Tokenizer:
  - *Index Time*: one token for each (sub-)domain of the URL.
  - *Query time:* Splits on whitespace, punctuation, etc...
- Filters:
  - Characters are converted to lower case.

Field-type: category

The type of the stories' categories.

- Tokenizer:
  - *Index Time*: Categories stay as a single token.
  - *Query time:* Splits on whitespace, punctuation, etc...
- Filters:
  - Characters are converted to lower-case.

Schema Definition

Field-type: newsfilter

The type of the *newssite_filter* field. Allows the user to tailor their results towards news or non-news content.

- Tokenizer:
  - *Index Time*: extracts the domain (not the subdomains) of a URL.
  - *Query time:* Splits on whitespace, punctuation, etc...
- Filters:
  - *Index Time*: Only the domains present on a news website list are kept. The tokens output from this process can only be **empty** (not a news website) or the word **news** (a news website).
  - *Query time:* only the tokens new and news will match with the URLs of news-related websites.

Field-type: other

These field types are just wrappers around built-in Solr field types, without any defined tokenizers or filters.

- **singleToken** - The type of text fields that aren't indexed (e.g.: a story's author).
- **date** - The type of fields representing points in time. Useful for sorting/filtering for recent stories.
- **int** - The type of integer fields (e.g.: the number of descendants of a story).

Recovery process

Query: version control tools -svn

Goal:

- Find tools for software version control or news related to the topic.
- Filter *svn* related results.

Results:

- Overall good results.

Query: China +(newssite_filter:news)

Goal:

- Find stories from news websites discussing topics relating to China.

Results:

- Only relevant documents were found.

Recovery process & Evaluation

Query: version control tools -svn

Goal:

- Find tools for software version control or news related to the topic.
- Filter SVN related results.

Results:

- All systems produce similar precision and recall results
- Most results refer to GitHub and GitLab

System	AP	P@10	R@10	F_β
Schemaless	90.85%	70.00%	70.00%	70.00%
Schema	89.04%	90.00%	52.94%	78.95%
Weighted Schema	85.71%	90.00%	52.94%	78.95%
Schema w/o synonyms	87.07%	60.00%	75.00%	62.50%

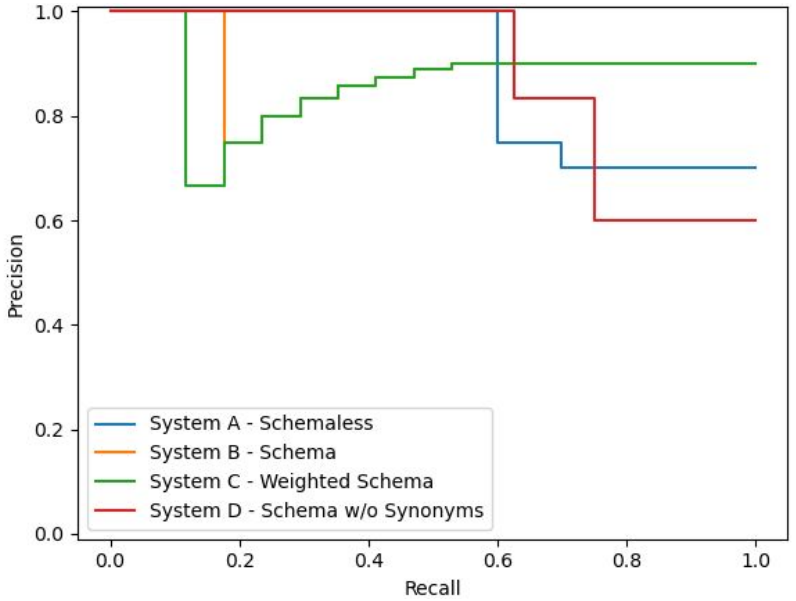


Table Caption

AP - Average Precision

R@10 - Recall for the first 10 retrieved documents

P@10 - Precision for the first 10 retrieved documents

Fβ - Fβ-score (β = 0.5) calculated as follows: $(1+\beta^2)\cdot\frac{P\cdot R}{(\beta^2\cdot P)+R}$

Recovery process & Evaluation

Query: eclipse

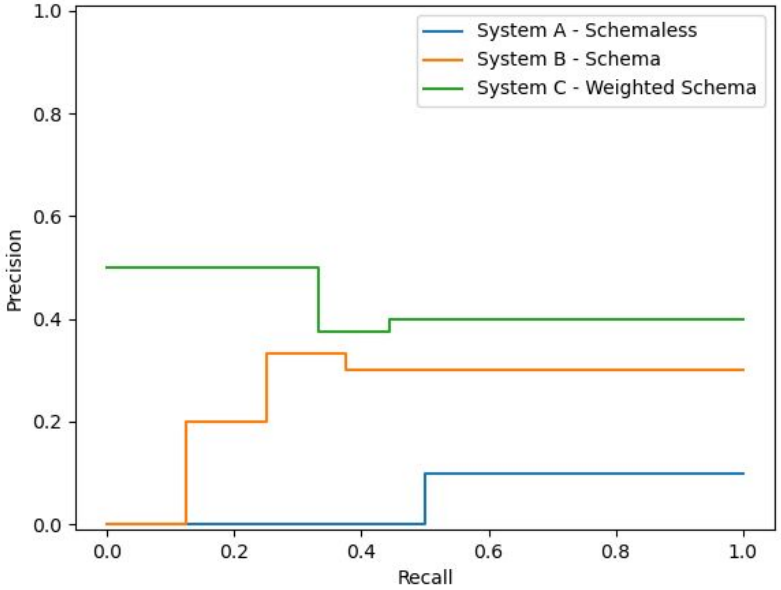
Goal:

- Find news/blog posts that talk to the Eclipse IDE or plugins/tools that interact with it.

Results:

- Ambiguous query → almost all retrieved documents aren't relevant.
- Stories about astronomy and the solar/lunar eclipse phenomena.
- Texts where eclipse is mentioned as a way to refer to something that has fallen into disuse.

System	AP	P@10	R@10	F_β
Schemaless	1.00%	10.00%	50.00%	11.90%
Schema	22.20%	30.00%	37.50%	31.25%
Weighted Schema	54.15%	40.00%	44.44%	40.82%



Recovery process & Evaluation

Query: JSON command line tools

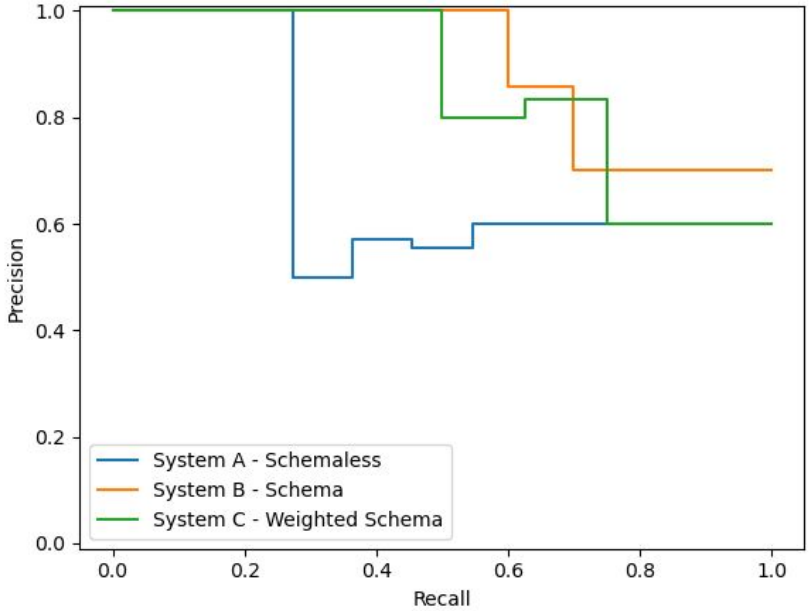
Goal:

- Find tools that manage files/stream in the JSON format, on the command line.

Results:

- The schema system had 30% more precision than the *schemaless* system.
- The usage of the term "command line" seems to reduce the number of irrelevant documents retrieved.
- Results of the same query with these terms matching separately proved to produce significantly worse results (20% less precision and recall).

System	AP	P@10	R@10	F_β
Schemaless	72.02%	60.00%	54.55%	58.82%
Schema	92.10%	70.00%	70.00%	70.00%
Weighted Schema	85.07%	60.00%	75.00%	62.50%



Recovery process & Evaluation

Query: China +(newssite_filter:news)

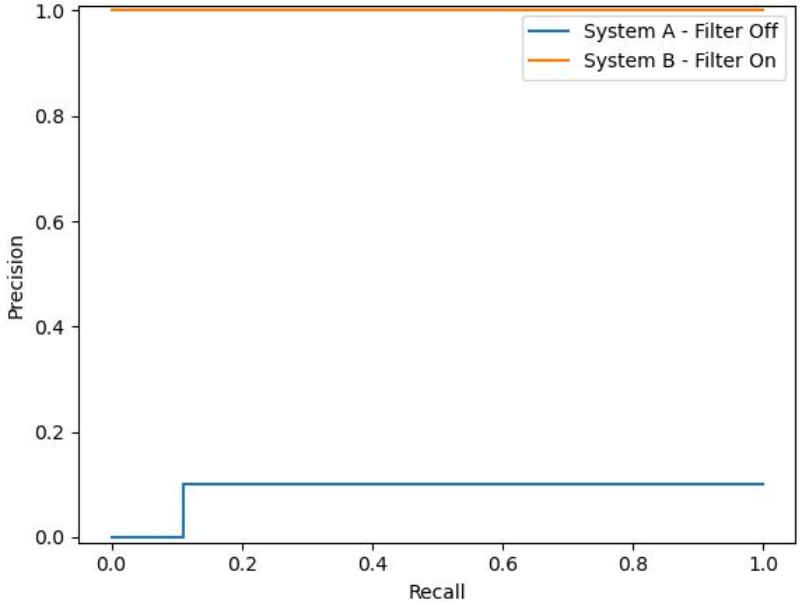
Goal:

- Find stories from news websites discussing topics relating to China.

Results:

- With filter → only relevant document retrieved.
- No filter → 1 relevant document retrieved.
- No filter → Most stories are personal blog posts, guides on how to optimize content for the Chinese market and persona takes on Chinese policies/culture.

System	AP	P@10	R@10	F_β
Schema w/ news filter	2.11%	10.00%	11.11%	10.20%
Schema w/o news filter	100.00%	100.00%	50.00%	83.33%



Future Work

- Better **weight system**:
  - Performed below expectations.
  - Could improve by weighing the exact match of the original term higher than the synonyms
- **N-grams** can be used for auto-completion features for the front-end.
- Find a way to identify *niche* stories:
  - Would allow users to **explore topics**.
- Use Solr's **faceting mechanism**:
  - Tune queries using *story type* field and *news story* filter.
  - Allows users to filter their search according to how many results match each category.