Group 53
Diana Freitas,
Diogo Samuel Fernandes,
Juliane Marubayashi

Monitorizado por João Damas

Goodreads books & reviews

PRI - Information Retrieval

goodreads

Table of Contents

01. Retrieval Tool

02. Information Needs

03. Documents and Collections

04. Schema & Indexing

05. Retrieval and Evaluation

06. Future Work

01

Retrieval Tool

02

Information Needs

Information Needs

The quality of the search was improved by providing the possibility to search books by thematic, specific events and named-entities, and also by distinguishing negative and positive reviews.

The information needs to be approached are the following:

- ➔ Search for books about a **specific theme**;
- ➔ Search for **positive and negative feedback** on a book review;
- ➔ Search for books about a **historical events** and **named-entities**;
- ➔ Search for **scientific books**;

03

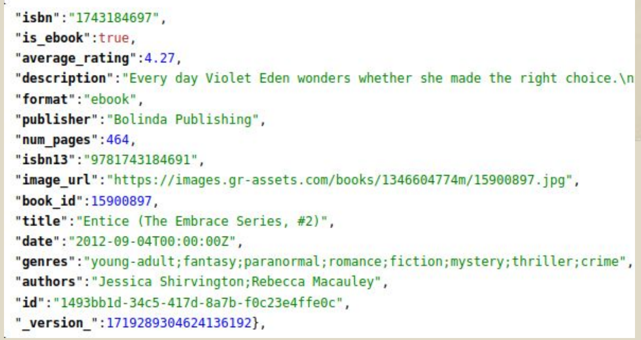
Documents & Collections

Two Documents

Both documents are saved on a single CSV file.
Subsets for each query.

Books

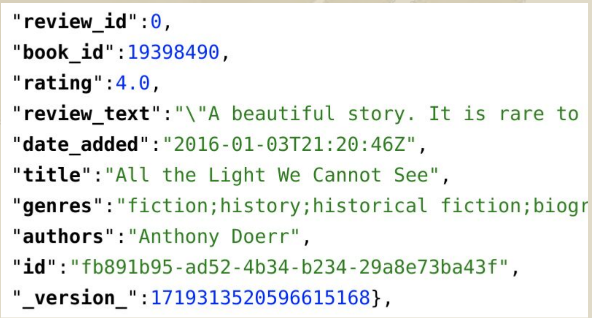
Contains relevant information about the books, the most relevant fields being: average_rating, description, title & genres.



```
"isbn":"1743184697",
"is_ebook":true,
"average_rating":4.27,
"description":"Every day Violet Eden wonders whether she made the right choice.\n
"format":"ebook",
"publisher":"Bolinda Publishing",
"num_pages":464,
"isbn13":"9781743184691",
"image_url":"https://images.gr-assets.com/books/1346604774m/15900897.jpg",
"book_id":15900897,
"title":"Entice (The Embrace Series, #2)",
"date":"2012-09-04T00:00:00Z",
"genres":"young-adult;fantasy;paranormal;romance;fiction;mystery;thriller;crime",
"authors":"Jessica Shirvington;Rebecca Macauley",
"id":"1493bb1d-34c5-417d-8a7b-f0c23e4ffe0c",
"_version_":1719289304624136192},
```

Reviews

Merges the reviews with some book details:
title, genre, authors & date



```
"review_id":0,
"book_id":19398490,
"rating":4.0,
"review_text":"\"A beautiful story. It is rare to
"date_added":"2016-01-03T21:20:46Z",
"title":"All the Light We Cannot See",
"genres":"fiction;history;historical fiction;biogr
"authors":"Anthony Doerr",
"id":"fb891b95-ad52-4b34-b234-29a8e73ba43f",
"_version_":1719313520596615168},
```

04

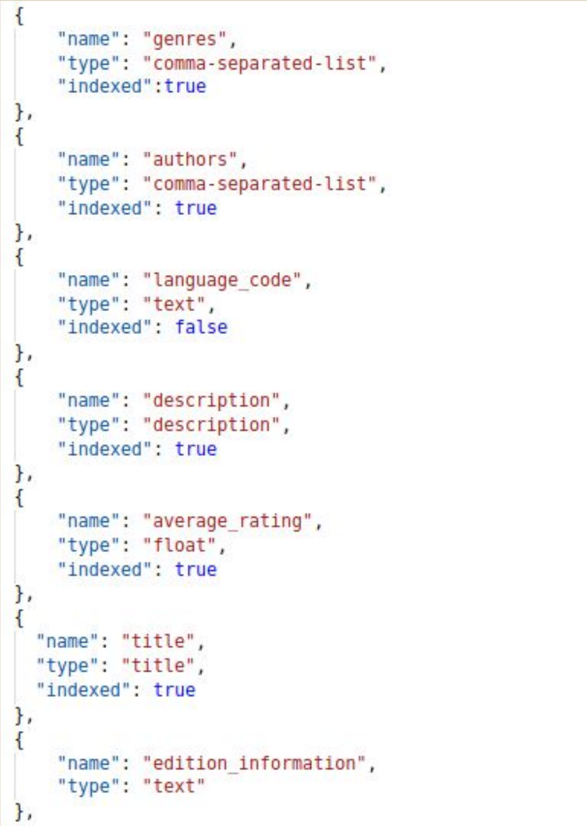
Schema & Indexing

Indexing

The fields that are going to be used to retrieve data and to build the queries, which will satisfy the information needs, are the ones to be indexed.

The indexed fields have one of the following types:

- ➔ **Numeric** - Solr predefined types like *TrieFloatField* - **rating**;
- ➔ **Text fields** - Solr *TextField* was not enough as each text field would require specific tokenizers and filters. Thus, custom field types were created for **review**, **description** and **title.**
- ➔ **Custom fields** - The **authors** and **genres**, which were represented as lists of words separated by semicolon, use the type *Comma Separated List*.



```
{
    "name": "genres",
    "type": "comma-separated-list",
    "indexed":true
},
{
    "name": "authors",
    "type": "comma-separated-list",
    "indexed": true
},
{
    "name": "language_code",
    "type": "text",
    "indexed": false
},
{
    "name": "description",
    "type": "description",
    "indexed": true
},
{
    "name": "average_rating",
    "type": "float",
    "indexed": true
},
{
  "name": "title",
  "type": "title",
  "indexed": true
},
{
    "name": "edition_information",
    "type": "text"
},
```

Tokenizers & Filters

Comma Separated List

- ➔ *SimplePatternTokenizerFactory*

Title, Description & Review

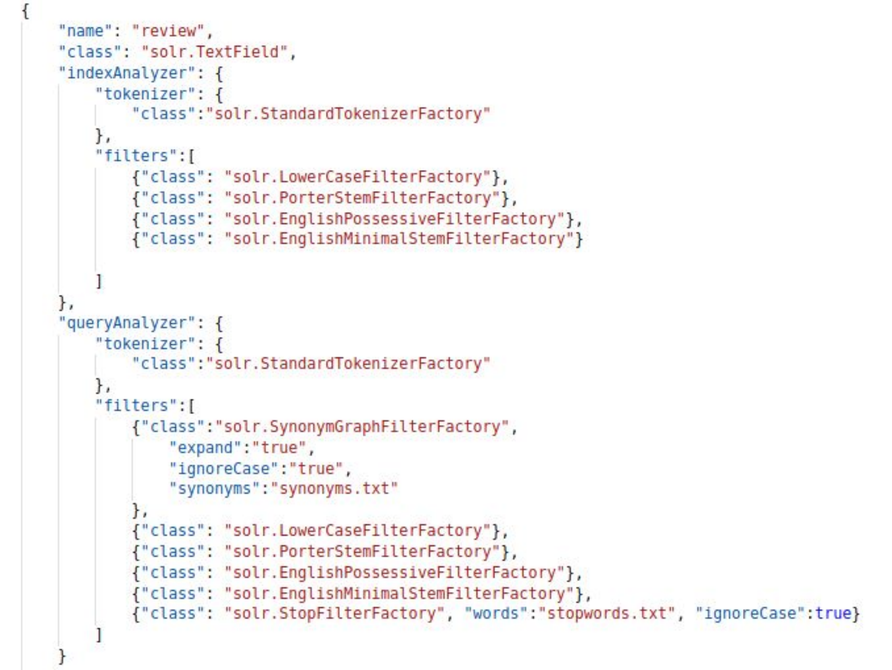
- ➔ *StandardTokenizerFactory*
- ➔ *LowerCaseFilterFactory*

Description

- ➔ *EnglishMinimalStemFilterFactory*

Review Text

- ➔ *PorterStemFilterFactory*
- ➔ *EnglishPossessiveFilterFactory*
- ➔ *EnglishMinimalStemFilterFactory*
- ➔ *StopFilterFactory* [Q]



```
{
    "name": "review",
    "class": "solr.TextField",
    "indexAnalyzer": {
        "tokenizer": {
            "class":"solr.StandardTokenizerFactory"
        },
        "filters":[
            {"class": "solr.LowerCaseFilterFactory"},
            {"class": "solr.PorterStemFilterFactory"},
            {"class": "solr.EnglishPossessiveFilterFactory"},
            {"class": "solr.EnglishMinimalStemFilterFactory"}

        ]
    },
    "queryAnalyzer": {
        "tokenizer": {
            "class":"solr.StandardTokenizerFactory"
        },
        "filters":[
            {"class":"solr.SynonymGraphFilterFactory",
                "expand":"true",
                "ignoreCase":"true",
                "synonyms":"synonyms.txt"
            },
            {"class": "solr.LowerCaseFilterFactory"},
            {"class": "solr.PorterStemFilterFactory"},
            {"class": "solr.EnglishPossessiveFilterFactory"},
            {"class": "solr.EnglishMinimalStemFilterFactory"},
            {"class": "solr.StopFilterFactory", "words":"stopwords.txt", "ignoreCase":true}
        ]
    }
```

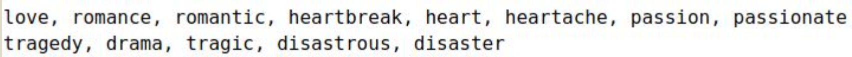
05

Retrieval and Evaluation

Q1 - Romantic Tragedy

The search seeks to find books based on the specified thematic, in this case books with romantic and tragedy would be retrieved.

Relevant Filters: *SynonymGraphFilterFactory*



```
love, romance, romantic, heartbreak, heart, heartache, passion, passionate
tragedy, drama, tragic, disastrous, disaster
```

Queries:

1) Simple query that searches for the words "romantic" and "tragedy" in the description field with a limit of 8 results.
2) Explores the proximity of the two words and applies a boost to the description.

System 1	System 2

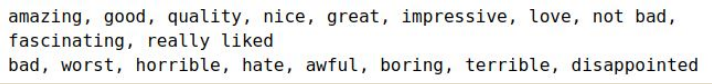
description:romantic tragedy
q..op = OR
rows=8
	q= {!q.op=OR df=edismax}romantic tragedy
	defType=edismax
	qf=description^2
	rows=8
	ident=true
	ps=4

	System 1		System 2	
Rank	**Title**	**R**	**Title**	**R**
1	Three Sides of a Heart: Stories about Love Triangles	Y	Wuthering Heights	Y
2	A Boy in Winter	Y	Three Sides of a Heart: Stories about Love Triangles	Y
3	The Medicine Man: Book 2	Y	A Boy in Winter	Y
4	Ride a Dark Horse	Y	The Medicine Man: Book 2	Y
5	Devil You Know (Lost Boys #1)	Y	Ride a Dark Horse	Y
6	Crown of Midnight (Throne of Glass, #2)	N	Devil You Know (Lost Boys #1)	Y
7	Wickett's Remedy	N	Crown of Midnight (Throne of Glass, #2)	N
8	Titus Andronicus	N	Wickett's Remedy	N
AP	0.8965		0.950893	
AP@5	1.0		1.0	
Recall	0.83		1.0	

Q2 - Negative/Positive Reviews

Retrieve negative and positive reviews from the dataset. A user could try to find bad reviews, for example, by adding words such as "hate" and "terrible" to the search field.

Relevant Filters: *SynonymGraphFilterFactory, StopFilterFactory*



```
amazing, good, quality, nice, great, impressive, love, not bad,
fascinating, really liked
bad, worst, horrible, hate, awful, boring, terrible, disappointed
```

Queries:

1) Uses Synonyms
2) Uses Synonyms, limits and sorts by rating

System 1	System 2

q= {!q.op=OR df=review_text}in love with this book
rows=10
	q= {!q.op=OR df=review_text}in love with this book AND rating:[3 TO *]
	sort=field(rating, min) desc
	rows=10

	System 1		System 2	
Rank	**Title**	**R**	**Title**	**R**
1	On Stranger Tides	N	The Collector (Dante Walker, #1)	Y
2	Taunting Destiny (The Fae Chronicles, #2)	Y	I Am the Messenger	Y
3	I Am the Messenger	Y	On Stranger Tides	Y
4	By Any Other Name (Forbidden, #1)	Y	The Six Rules of Maybe	Y
5	Taunting Destiny (The Fae Chronicles, #2)	Y		
6	Taunting Destiny (The Fae Chronicles, #2)	Y		
7	The Name of the Wind (The Kingkiller Chronicle, #1)	Y		
8	A Love Letter to Whiskey	N		
9	Solaris: The Definitive Edition	N		
10	Born a Crime: Stories From a South African Childhood	Y		
AP	0.435794		1.0	
AP@5	0.4		0.8	
Recall	0.285714		0.571429	

Q3 - World War

System 1 (no filter)	System 2	System 3

description: "world war"
rows=10
	description: "world war"
	rows=10
		description:"world war"~5^5
		defType:edismax
		qf=description^1
		rows=10

It's not uncommon that specific moments in history and named-entities are referenced with more than one word, such as: "World War II" and "Barack Obama". But frequence that these terms appears together can be high.

Relevant Filters: *LowerCaseFilterFactory, EnglishMinimalStemFilterFactory*

Queries:

1) Exact match, without filters.
2) Exact match, with filters.
3) Uses filters, tries to minimize the distance between words and uses boosts.

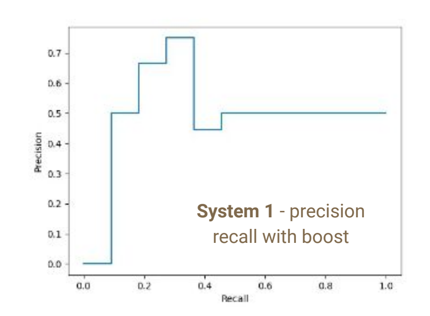
	System 1		System 2		System 3	
Rank	**Title**	**R**	**Title**	**R**	**Title**	**R**
1	The Omega Covenant	Y	The Winds of War (The Henry Family, #1)	Y	The Winds of War (The Henry Family, #1)	Y
2	The Pearl Harbor Murders (Disaster, #3)	Y	Ancestors in Arms	Y	Ancestors in Arms	Y
3	The Mandibles: A Family, 2029-2047	Y	Gated Grief: The Daughter of a GI Concentration Camp Liberator Discovers a Legacy of Trauma	Y	Gated Grief: The Daughter of a GI Concentration Camp Liberator Discovers a Legacy of Trauma	Y
4			Women Heroes of World War II: 26 Stories of Espionage, Sabotage, Resistance, and Rescue	Y	Women Heroes of World War II: 26 Stories of Espionage, Sabotage, Resistance, and Rescue	Y
5			Season In Purgatory	Y	Season In Purgatory	Y
6			History of Modern World	Y	History of Modern World	Y
7			Mein Kampf - My Struggle: Unabridged edition of Hitlers original book - Four and a Half Years of Struggle against Lies, Stupidity, and Cowardice	Y	Mein Kampf - My Struggle: Unabridged edition of Hitlers original book - Four and a Half Years of Struggle against Lies, Stupidity, and Cowardice	Y
8					Star Wars: The Ultimate Visual Guide: Updated and Expanded	N
9					Other People's Houses	Y
AP	1.0		1.0		0.973765	
AP@5	0.6		1.0		1.0	
Recall	0.3		0.7		0.8	

Q4 - Scientific Books
without schema

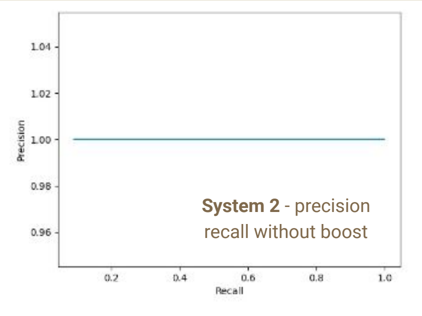
Science Fiction books were also being retrieved when searching for scientific books.
This query improves the relevance of the results when searching for words related to "science".

Queries:

1) Matches the term “science” in the description and title.
2) Improves the first query by excluding **fiction** and **historical-fiction** genres and boosts the **non-fiction** field.



System 1 - precision recall with boost



System 2 - precision recall without boost

System 1	System 2
description:science title:science q.op:OR	q:science -genres:fiction-genres:"historical-fiction" defType:edismax qf:description title q.op:AND bq:genres:"non-fiction"^4

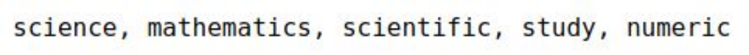
	System 1		System 2	
Rank	**Title**	**R**	**Title**	**R**
1	Science-Fiction 2007	N	Science Matters: Achieving Scientific Literacy	Y
2	Science Matters: Achieving Scientific Literacy	Y	The Science of Getting Rich	Y
3	A Reformed Approach to Science and Scripture	Y	A Reformed Approach to Science and Scripture	Y
4	The Science of Getting Rich	Y	Cure: A Journey Into the Science of Mind over Body	Y
5	Cure: A Journey Into the Science of Mind over Body	Y	Probability and Computing: Randomized Algorithms and Probabilistic Analysis	Y
6	Lightspeed Magazine, October 2012	N		
7	All Judgment Fled	N		
8	L. Ron Hubbard Presents Writers of the Future 27	N		
9	The Affair	N		
10	Probability and Computing: Randomized Algorithms and Probabilistic Analysis	Y		
AP	0.543		1.0	
AP@5	0.8		1.0	
Recall	0.4545		0.4545	

Q4 - Scientific Books
with schema

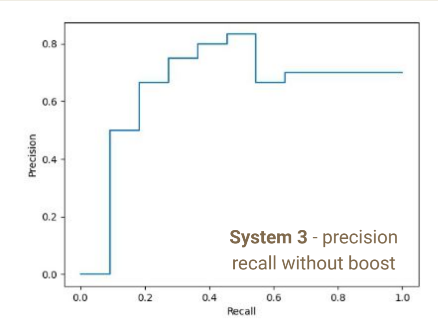
Queries:

1) Matches the term “science” in the description and title.
2) Improves the first query by excluding **fiction** and **historical-fiction** genres and boosts the **non-fiction** field.

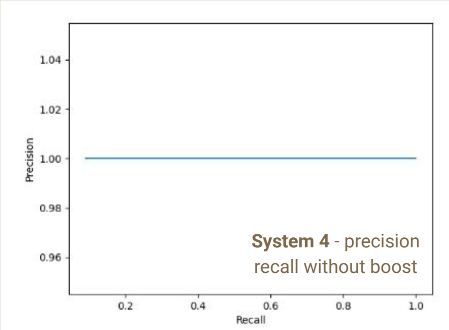
Relevant Filters: *SynonymGraphFilterFactory; LowerCaseFilterFactory; EnglishMinimalStemFilterFactory*



```
science, mathematics, scientific, study, numeric
```



System 3 - precision recall without boost



System 4 - precision recall without boost

	System 3		**System 4**	
Rank	**Title**	**R**	**Title**	**R**
1	Science-Fiction 2007	N	Science Matters: Achieving Scientific Literacy	Y
2	Science Matters: Achieving Scientific Literacy	Y	The Science of Getting Rich	Y
3	The Science of Getting Rich	Y	A Reformed Approach to Science and Scripture	Y
4	A Reformed Approach to Science and Scripture	Y	Cure: A Journey Into the Science of Mind over Body	Y
5	Cure: A Journey Into the Science of Mind over Body	Y	Encounters: Two Studies in the Sociology of Interaction	Y
6	Encounters: Two Studies in the Sociology of Interaction	Y	Probability and Computing: Randomized Algorithms and Probabilistic Analysis	Y
7	Probability and Computing: Randomized Algorithms and Probabilistic Analysis	Y	Real Change: Conversion	Y
8	The Affair	N	Literary Theory: The Basics	Y
9	Lightspeed Magazine, October 2012	N	Topoi: The Categorial Analysis of Logic	Y
10	Math for Smarty Pants	Y	Math for Smarty Pants	Y
AP	0.652381		1.0	
AP@5	0.8		1.0	
Recall	0.636364		0.91	

06

Future Work

Future Work

For the next and last milestone, we plan to improve our search system by:

- ➔ Creating a graphic user interface;
- ➔ Improving sentiment analysis in the reviews - *NLPK;*
- ➔ Generating synonyms automatically - *NLPK;*
- ➔ Using multi-word synonyms search;
- ➔ Creating new queries that use the date fields;
- ➔ Completing the dataset with recent books.