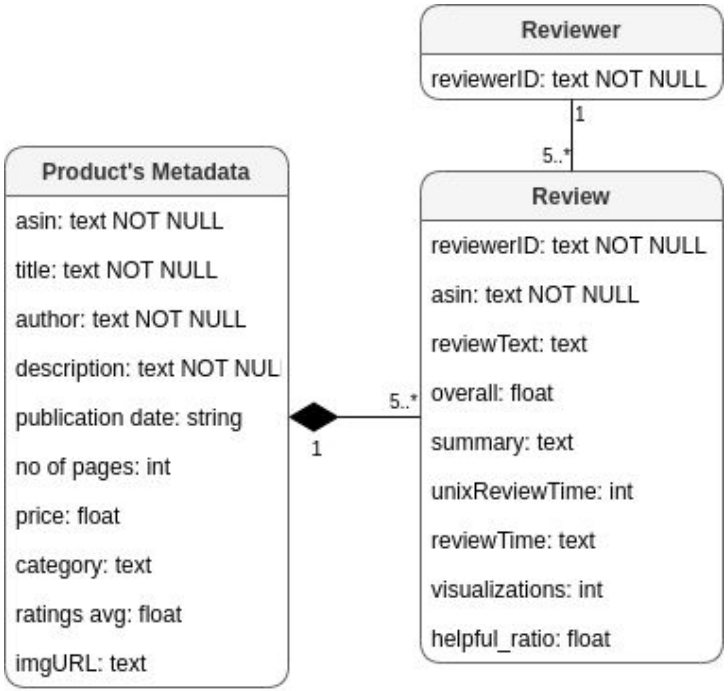
KindleMeSome: amazon Kindle Reviews

PRI - Information Processing and Retrieval 2022

M.EIC - Master's in Informatics and Computing Engineering

António Ribeiro - up201906761@edu.fe.up.pt
Diogo Maia - up201904974@edu.fe.up.pt
Luís Viegas - up201904979@edu.fe.up.pt

Conceptual Model UML



Documents

These documents represent, respectively, the books and the reviews. Both were added to the same core/collection, ‘kindle’.


```
{
  "id":"B0012W11D0",
  "title":"Hot Ticket (Serving Love) - Kindle edition",
  "brand":"K.A. Mitchell",
  "category":"Literature & Fiction",
  "publication_date":"2008-01-15T00:00:00Z",
  "no_pages":98,
  "description":"no description",
  "overall":4.3333335,
  "type":"book",
  "_version_":1749327849146286080},
{
```

```
{
  "id":"0",
  "reviewerID":"A3SZMGJMV0G16C",
  "asin":"B000FA64PK",
  "reviewerName":"Andrew Pruette \"Rancors Love to Read\"",
  "reviewText":"Troy Denning's novella Recovery was originally publis
  "overall":3.0,
  "summary":"Han and Leia reunited and Barabel Jedi introduced",
  "reviewTime":"2012-03-15T00:00:00Z",
  "helpful_ratio":0.0,
  "visualization":0,
  "type":"review",
  "_version_":1749327854948057088},
{
```

Indexing

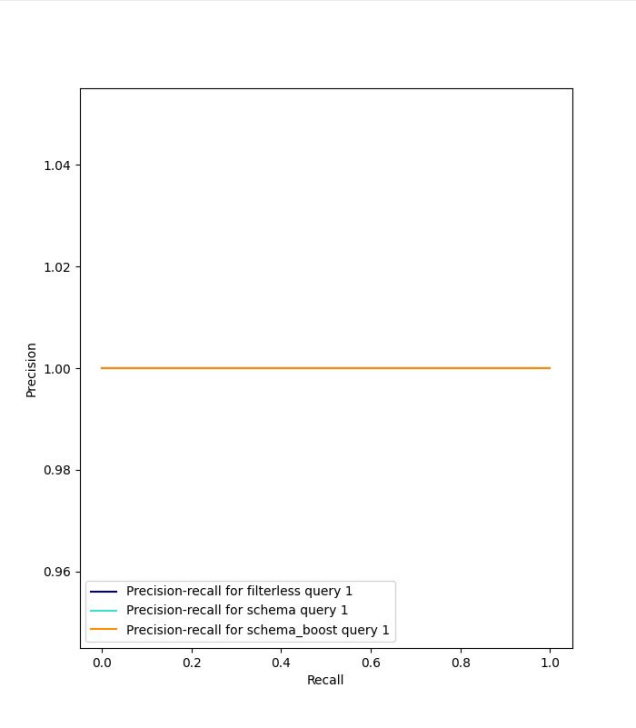
This is the schema of our database. As the two different type of documents were added to the same core, this schema is for both of them.

type	name	indexed
text	reviewerName	true
	brand	true
	category	true
	title	true
body	description	false
	reviewText	true
	summary_ratio	true
string	type	true
	imgUrl	false
	asin	true
	reviewerID	false
pdate	publication_date	true
	reviewTime	false
pfloat	price	false
	overall	true
	helpful_ratio	true
pint	no_pages	true
	visualization	false

Field Types Added

Field Type (tokenizer)	Filters
text StandardTokenizerFactory	ASCIIFoldingFilterFactory LowerCaseFilterFactory
body ClassicTokenizerFactory	ASCIIFoldingFilterFactory LowerCaseFilterFactory ClassicFilterFactory StopFilterFactory PorterStemFilterFactory SynonymGraphFilterFactory RemoveDuplicatesFilterFactory FlattenGraphFilterFactory

Top 10 books written by "Francis"

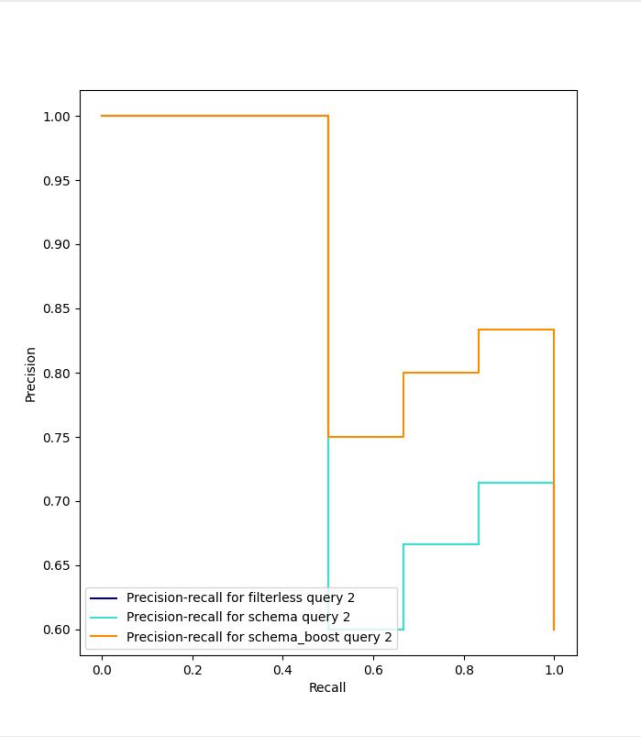


- lucene parser
- q: "brand: francis"
- q,op: OR
- sort : overall desc
- rows: 10

```
Boosts: No boosts
```

Scenario	AvP	P@10
1	1.0	1.0
2	1.0	1.0
3	1.0	1.0

Small soccer and football books

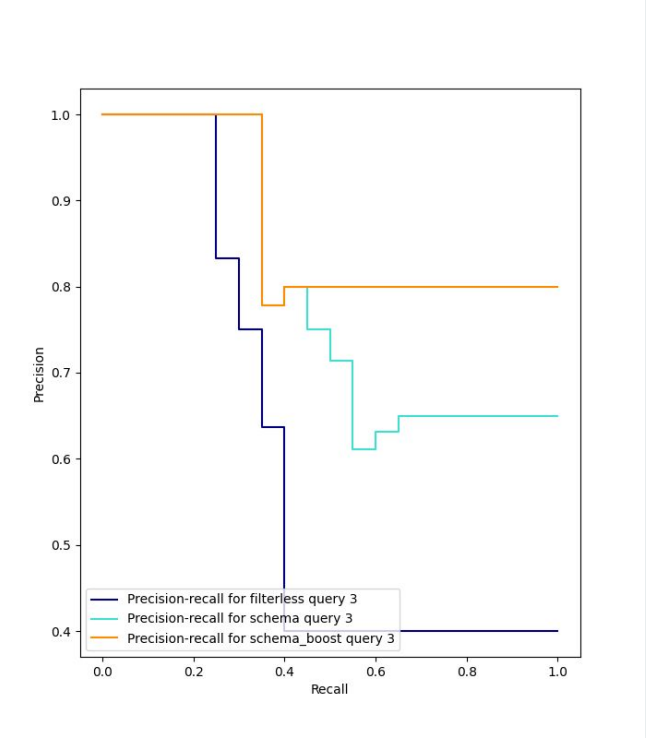


- eDisMax parser
- q: soccer football
- q,op: OR
- qf:: title brand description
- fq: no_pages[0 TO 150]

`Boosts: Field boost of 2x in 'title'`

Scenario	AvP	P@10
1	0.855159	0.6
2	0.915079	0.6
3	0.855159	0.6

Positive reviews from books of the author Dr.Leland Benton

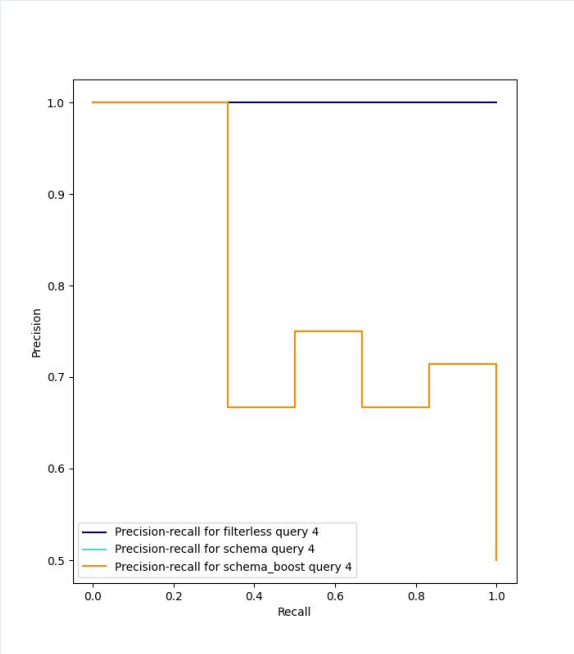


- eDisMax parser
- q: good
- q,op: OR
- fq: {!join from=id to=asin}brand:”Dr.Leland Benton” & overall:[4.0 TO *]
- sort: helpful_rate desc, visualization desc
- fl: reviewText summary

```
Boosts: Field boost of 2x in ‘summary’
```

Scenario	AvP	P@10
1	0.912698	0.7
2	0.877102	0.8
3	0.975	0.8

Reviews recommending books for vacations done in vacation period of 2013

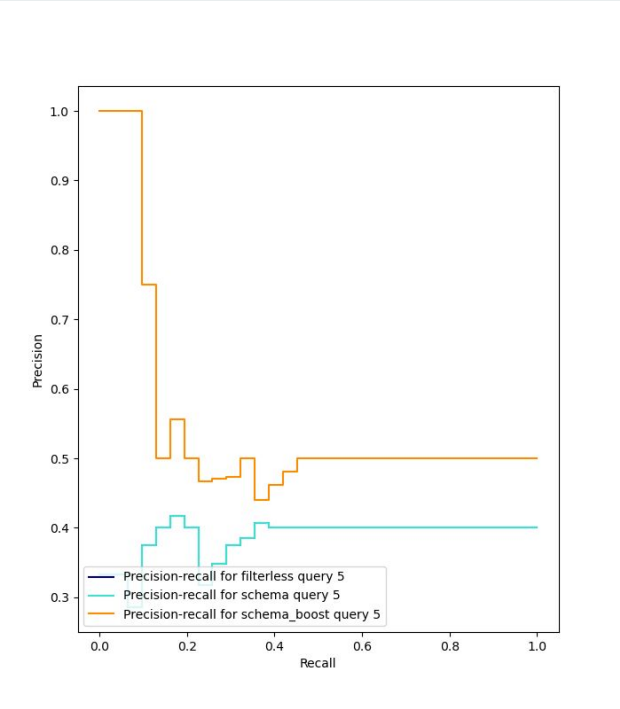


- eDisMax parser
- q: “vacation read”
- q,op: OR
- fq: type:”review” & reviewDate:[2013-06-01T00:00:00Z TO 2013-09-31T00:00:00Z]
- sort: helpful_rate desc, visualization desc
- fl: reviewText summary
- qs: 1

`Boosts: Field boost of 2x in ‘summary’`

Scenario	AvP	P@10
1	1.0	0.2
2	0.835714	0.6
3	0.835714	0.6

Reviews that depict the book as boring, difficult to read or long



- eDisMax parser
- q: dull difficult endless
- q.op: OR
- fq: type:review, overall:

[* TO 2.0]

- fl: id summary reviewText

asin overall

```
Boosts:
4x boost to dull and 2x to
endless.
Field boost of 1.5x on
'Summary'.
Boost function of term
frequency 'not' in
reviewText.
Phrase boost with 10 slop
in reviewText
```

Scenario	AvP	P@10
1	0.469319	0.4
2	0.469319	0.4
3	0.631873	0.6

Future Work

- Improve query performances and metric scores.
- Store reviews as a nested document list in Books.
- Research and implement more semantic analysis aspects in query 4 and 5.
- Create a frontend interface.
- Play around with other SolR tools, such as facets.