Book Characterization using Project Gutenberg's Open Library with Goodreads Reviews

Information Processing and Retrieval

Group 68

Motivation

Massive bodies of text, such as books, have become readily available in large online e-book libraries.

Hard to:

- Look for context around well-known quotes
- Know a book's general rating
- Browse public domain texts

→ Information Retrieval System

Conceptual Model and Document Definition

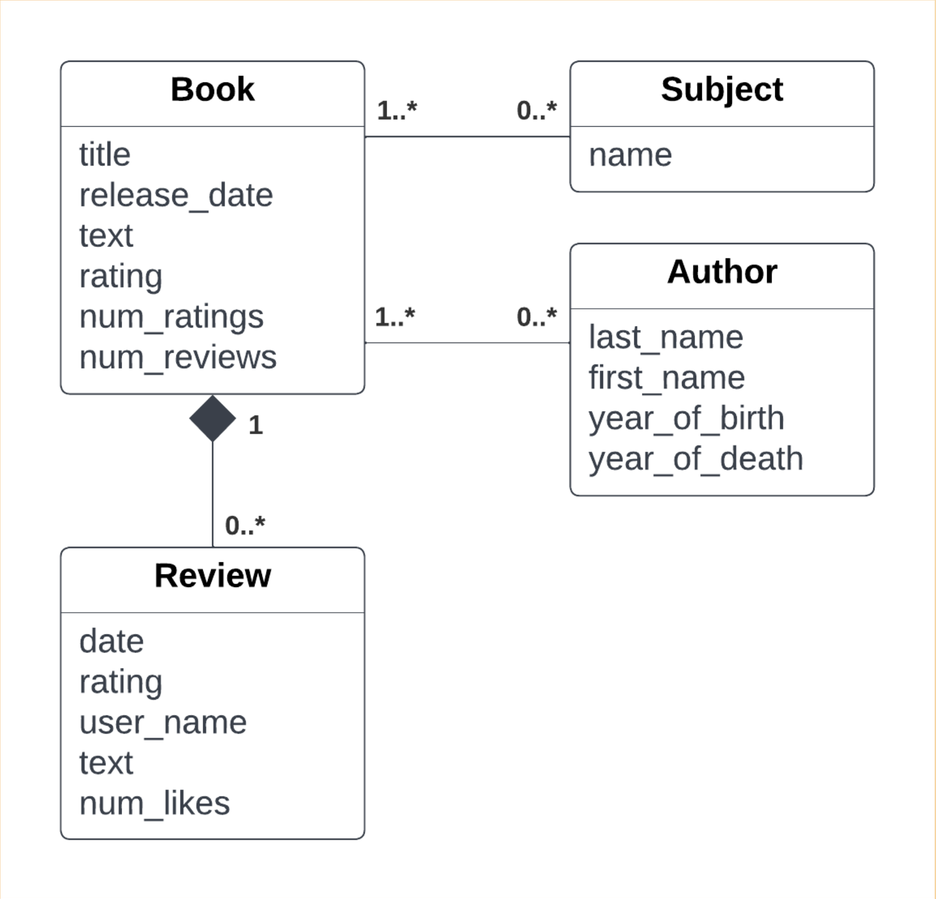

Fig. 1: Dataset Conceptual Model

Three types of documents:

- Book Document
  - Contains array of subjects
- Author Document
  - Nested in Book Documents
- Review Document
  - Nested in Book Documents

Two fields common to all documents:

- content_type (*BOOK, AUTHOR, REVIEW*)
- id

Search Scenarios

There are 5 main search scenarios:

> I want to find excerpts about a specific concept or event.
> I want to find context on a book quote.
> I want to read a book from a certain time period.
> I want to understand what other people think of a book.
> I want to browse an extensive library of copyright-free literature.

Indexing — Indexed Fields

Table 1: Fields defined in the schemas.

field name	field type	indexed
content_type	ContentType	yes
title	GeneralTextField	yes
release_date	pdate	yes
subjects	GeneralTextField	yes
rating	pfloat	yes
num_ratings	plong	no
num_reviews	plong	no
text	GeneralTextField	yes
last_name	NameTextField	yes
first_name	NameTextField	yes
year_of_birth	DateRangeFieldType	yes
year_of_death	DateRangeFieldType	yes
date	pdate	no
user_name	string	no
num_likes	plong	no

Indexing — Index Analyzers (Schema 1)

GeneralTextField

Tokenizer: Standard Tokenizer
Filters:
- ASCII Folding (preserve original)
- Lower Case

NameTextField

Tokenizer: Standard Tokenizer
Filters:
- ASCII Folding (preserve original)
- Lower Case

Indexing — Index Analyzers (Schema 2)

GeneralTextField

Tokenizer: Standard Tokenizer
Filters:

- ASCII Folding (preserve original)
- Lower Case
- Synonym Graph
- Flatten Graph (not in the query analyser)
- English Possessive
- English minimal Stem

NameTextField

Tokenizer: Standard Tokenizer
Filters:

- ASCII Folding (preserve original)
- Lower Case

Systems

sys1

Schema 1 (base schema)
Basic Field Querying

sys1_syn

Schema 2 (synonym schema)
Basic Field Querying

sys2

Schema 1 (base schema)
Complex Field Querying

sys2_syn

Schema 2 (synonym schema)
Complex Field Querying

Information Need 1

Search Scenario

I want to find excerpts about a specific concept or event

Information Need

I want to find excerpts about a specific concept or event

Base Query

```
q: (subjects:economics OR text:economics
ORtitle:economics) (subjects:home OR
text:homeOR title:home)
q.op: AND
```

Improved Query

```
q: ((subjects:econ*) ^ 1.5 OR text:econ* OR
title:econ*) ((subjects:home~) ^ 1.5
ORtext:home~ OR title:home~)
q.op: AND
```


```
▼ responseHeader:
    status:          0
    QTime:           1
  ▼ params:
    ▼ q:             "(subjects:economics OR text:economics OR title:economics) (subjects:home OR text:home OR title:home)"
      indent:        "true"
      q.op:          "AND"
      rows:          "1"
▼ response:
    numFound:        452
    start:           0
    numFoundExact:   true
  ▼ docs:
    ▼ 0:
        id:          "36498"
      ▼ title:       "The Teaching of Art Related to the Home\nSuggestions for content and method in related art instructio
        release_date: "2011-06-23T00:00:00Z"
      ▼ subjects:
          0:         "Art -- Study and teaching"
          1:         "Home economics -- Study and teaching"
        rating:      0
        num_ratings: 0
        num_reviews: 0
        content_type: "BOOK"
      ▶ text:        "Produced by Juliet Suthe…for Vocational Education"
        _version_:   1749399962770210800
```


Fig. 2: Query Result (sys1, basic querying)

Information Need 1 - Evaluation

Table 2: Information Need 1 Metrics

Rank	sys1	sys2	sys1_syn	sys2_syn
1	I	R	I	R
2	R	R	I	I
3	R	I	R	R
4	I	R	R	R
5	R	R	R	R
6	R	R	R	R
7	R	R	R	R
8	I	R	R	I
9	I	R	I	I
10	R	R	I	I
P@10	0.6	0.9	0.6	0.6
AP	0.536464	0.852337	0.470106	0.758201

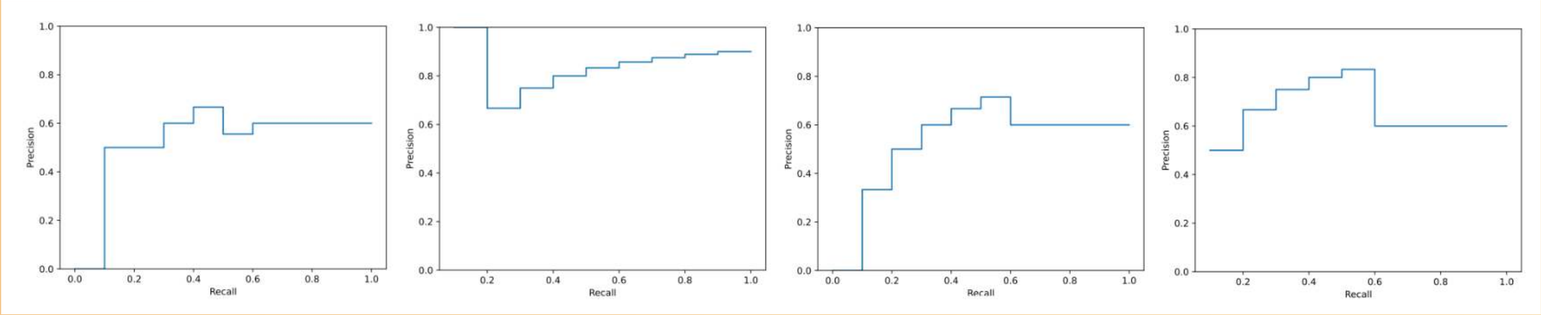

Fig. 3: P–R Curve

Information Need 2

Search Scenario

I want to read a book from a certain time period

Information Need

Recent romantic novels releasedpreferably around 2018

Base Query

```
q: subjects:fiction (subjects:love OR text:love)
q.op: AND
```

Improved Query

```
q: (subjects:lov* ^ 2 OR text:lov*)
q.op: AND
bq: release_date:[NOW/DAY-6YEAR TONOW/DAY-2YEAR]
```


```
▼ responseHeader:
    status:          0
    QTime:           1
  ▼ params:
    ▼ q:             "(subjects:lov*^2 OR text:lov*) (text:home OR title:home)"
      defType:       "edismax"
      indent:        "true"
      q.op:          "AND"
      rows:          "1"
      bq:            "release_date:[NOW/DAY-6YEAR TO NOW/DAY-2YEAR]"
▼ response:
    numFound:        7121
    start:           0
    numFoundExact:   true
  ▼ docs:
    ▼ 0:
        id:          "67621"
        title:       "Hilda's Home: A Story of Woman's Emancipation"
        release_date: "2022-03-13T00:00:00Z"
      ▼ subjects:
          0:         "Women -- Fiction"
          1:         "Women's rights -- Fiction"
          2:         "Free love -- Fiction"
        content_type: "BOOK"
      ▶ text:        "HILDAâ\u0080\u0099S HOME… font in _underscores_."
        _version_:   1749400112010887200
```


Fig. 4: Query Result (sys1, improved querying)

Information Need 2 - Evaluation

Table 3: Information Need 2 Metrics

Rank	sys1	sys2	sys1_syn	sys2_syn
1	I	R	I	R
2	I	I	I	I
3	I	R	I	I
4	R	I	I	R
5	R	R	I	I
6	R	R	I	I
7	I	I	I	I
8	I	I	I	R
9	I	R	I	R
10	I	I	I	R
P@10	0.3	0.5	0.0	0.5
AP	0.254101	0.617813	0.0	0.463536

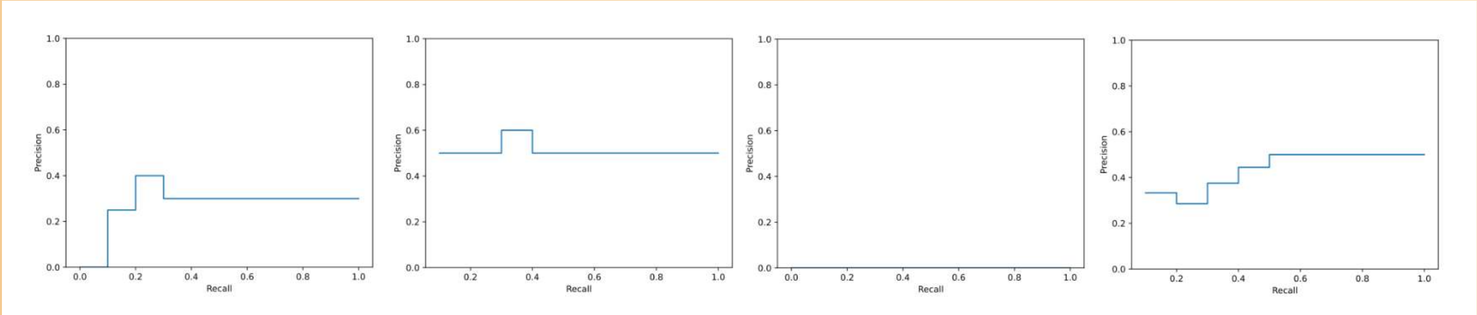

Fig. 5: P–R Curve

Information Need 3

Search Scenario
I want to find context on a book quote.

Information Need
Books containing the quote "waste not, want not", even if the user doesn't quite know the quote (inputting "want not, waste not")

Base Query
`q: text:"want not, waste not"`
`q.op: AND`

Improved Query
`q: text:"want not, waste not"~5`
`q.op: AND`


```
responseHeader:
  status:          0
  QTime:           56
  params:
    q:             "text:\"want not, waste not\""
    hl:            "true"
    indent:        "true"
    q.op:          "OR"
    hl.fragSize:   "10000000"
    hl.fl:         "text"
response:
  numFound:        1
  start:           0
  numFoundExact:   true
  docs:
    0:
      id:          "38390"
      title:       "A Dictionary of English … Variety of Phraseology"
      release_date: "2011-12-23T00:00:00Z"
      subjects:
        0:         "English language -- Synonyms and antonyms"
      rating:      0
      num_ratings: 0
      num_reviews: 0
      content_type: "BOOK"
      text:        "Produced by Betsie Bush,…ssions, by Richard Soule"
      _version_:   1749401112730927000
```


Fig. 6: Query Result (sys1_syn, basic querying)

Information Need 3 - Evaluation

Table 4: Information Need 3 Metrics

Rank	sys1	sys2	sys1_syn	sys2_syn
1	N/A	R	I	I
2	N/A	R	N/A	I
3	N/A	R	N/A	R
4	N/A	I	N/A	R
5	N/A	I	N/A	I
6	N/A	R	N/A	I
7	N/A	I	N/A	I
8	N/A	R	N/A	I
9	N/A	R	N/A	I
10	N/A	R	N/A	I
P@10	0.0	0.7	0.0	0.7
AP	0.0	0.764418	0.0	0.764418

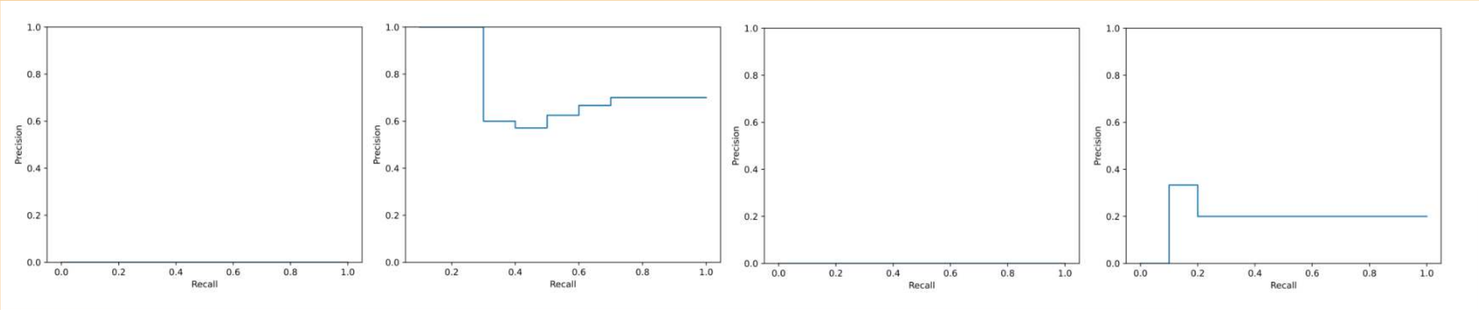

Fig. 7: P-R Curve

Information Need 4

Search Scenario
I want to find context on a book quote.

Information Need
Books containing the quote "wastenot, want not", even if the user doesn't quite know the quote and has typos (inputting "want nto, waste not")

Base Query
`q: text:"want nto, waste not"`
`q.op: AND`

Improved Query
`q: text:"want nto, waste not"~5`
`q.op: AND`


```
responseHeader:
  status:          0
  QTime:           15
  params:
    q:             "text:\"want nto, waste not\"~5"
    hl:            "true"
    indent:        "true"
    q.op:          "OR"
    hl.fl:         "text"
    rows:          "1"
response:
  numFound:        676
  start:           0
  numFoundExact:   true
  docs:
    0:
      id:          "3655"
      title:       "The Parent's Assistant; Or, Stories for Children"
      release_date: "2003-01-01T00:00:00Z"
      subjects:    […]
      rating:      3.07
      num_ratings: 41
      num_reviews: 2
      content_type: "BOOK"
      text:        "This eBook was produced …se of the 18th century."
      _version_:   1749401060218241000
```


Fig. 8: Query Result (sys2_syn, improved querying)


```
highlighting:
  3655:
    text:
      0:   "                                   135 FORGIVE AND FORGET                                   173 <em>WASTE</em> <em>NOT</em>, <em>WANT</em> NOT; OR, TWO STRINGS"
```


Fig. 9: Highlighted terms

Information Need 4 - Evaluation

Table 5: Information Need 4 Metrics

Rank	sys1	sys2	sys1_syn	sys2_syn
1	N/A	N/A	I	I
2	N/A	N/A	N/A	I
3	N/A	N/A	N/A	R
4	N/A	N/A	N/A	R
5	N/A	N/A	N/A	I
6	N/A	N/A	N/A	I
7	N/A	N/A	N/A	I
8	N/A	N/A	N/A	I
9	N/A	N/A	N/A	I
10	N/A	N/A	N/A	I
P@10	0.0	0.0	0.0	0.2
AP	0.0	0.0	0.0	0.258289

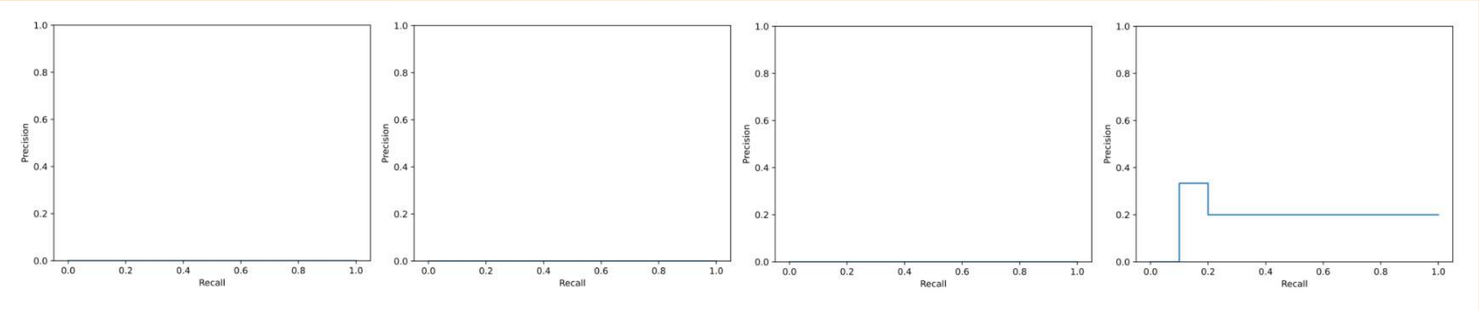

Fig. 10: P-R Curve

Mean Average Precision

Table 6: Mean Average Precision per system

System	mAP
sys1	0.197641
sys1_syn	0.117526
sys2	0.558641
sys2_syn	0.561111

Other Search Scenarios

Two search scenarios were not tackled as its results were not subject to evaluation — that is, its results are either binary or too broad.

I want to understand what other people think of a book.

> Only one relevant result
> Queries such as “retrieve books where the user reviews claim it is ‘easy to read’ proven to be too arduous: defining correct relevant documents was not possible

I want to browse an extensive library of copyright-free literature.

> Too broad
> Tackled by the creation of a user-friendly interface

Other Search Scenarios

Base Query

```
q: {!parent
which="content_type:BOOK"
}last_name:"Le Fanu"
q.op: AND
```


```
responseHeader:
  status:          0
  QTime:           0
  params:
    q:             "{!parent which=\"content_type:BOOK\"}\nlast_name:\"Le Fanu\""
    indent:        "true"
    fl:            "id, title, authors, first_name, last_name,[child fl=\"first_name,last_name\"]"
    q.op:          "OR"
response:
  numFound:        7
  start:           0
  numFoundExact:   true
  docs:
    0:
      id:          "10007"
      title:       "Carmilla"
      authors:
        last_name:   "Le Fanu"
        first_name:  "Joseph Sheridan"
    1:
      id:          "11610"
      title:       "Madam Crowl's Ghost and the Dead Sexton"
      authors:
        last_name:   "Le Fanu"
        first_name:  "Joseph Sheridan"
    2:
      id:          "11635"
      title:       "Green Tea;  Mr. Justice Harbottle"
      authors:
        last_name:   "Le Fanu"
        first_name:  "Joseph Sheridan"
```


Fig. 11: Books by ‘Le Fanu’

Conclusions and Future Work

Conclusions:

> Information retrieval system covered search scenarios
> Good schema definition and query optimization tools are essential for good performance

Future work:

> Improvements on the information retrieval system
> Experiment OpenNLP Tokenizers and Filters
> Develop a frontend that facilitates the retrieval of information for the end-user