Book Retrieval System

Maheen Asghar, Shi Lu
[1] School of Information, University of Michigan

OBJECTIVE

Finding a good book to read used to be a troublesome prospect that meant a trip to the library. However, in the age of information and the explosion of digitized data, it's easier than ever to discover new books. However, sites such as GoodReads only retrieve books with titles similar to a query. We wanted to create a book information retrieval system that would allow users to retrieve information on books relevant to their query--whether through the title or the description of a book. Our books were found using TF-IDF and BM25.

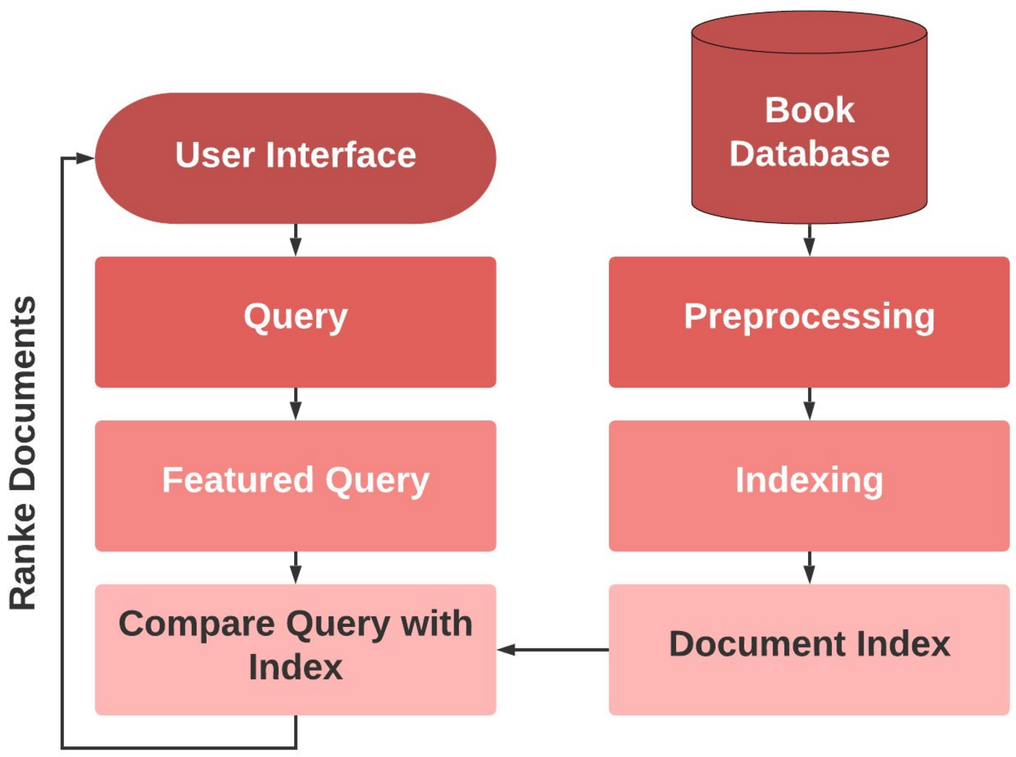
METHODOLOGY

DATASET

- 1,271 books: Google Books API[1]
  - Book titles, authors, ISBNs, publishers, publish dates, descriptions
- List of genres: Wikipedia[2]

PREPROCESSING

- Tokenize
- Convert numbers to words
- Remove stopwords
- Remove punctuation
- Remove non-alphabetic words



RESULTS

The search result reached the following benchmark:

1. When “beautiful joe” is entered as a query, the book *Beautiful Joe* should be the first in the result list.
   a. Compared to the same search in GoodReads, we are returned the same top two results in the same order.
2. The TI/IDF index on the page is ranked from largest returned.
3. Looking up unique names returns one result.

Goodreads	Google	Our Book Retrieval
Beautiful Joe	Beautiful Joe - Wikipedia	Beautiful Joe
Beautiful Joe's Paradise	Beautiful Joe: Marshall Saunders: 9781420944587: Amazon	Beautiful Joe's Paradise
Beautiful Joe: The True Story of a Brave Dog	Beautiful Joe by Marshall Saunders - Goodreads	Wheat that Springeth Green
Republic	Beautiful Joe - Broadview Press	Documentary Graphic Novels and Social Realism
A Beautiful Funeral (The Maddox Brothers, #5)	Beautiful Joe. - Digital Library	Alien Empires
Beautiful Joe: An Autobiography	Beautiful Joe	Nightlife
Beautiful Joe an Autobiography	The Book - Beautiful Joe Heritage Society	Vicious Circle

DISCUSSION

The current project focus on the implementation of TF-iDF using BM25 formula, which is a very classical way of constructing a retrieval function. No experiment was conducted to test different variations of the BM25 formula or a different retrieval model.

Due to the lack of labeled data, we could not use any of the most often used metrics for search ranking such as mean average precision or normalized discounted cumulative gain. We could only use one of two cases known in the dataset to see if they will show up in the search result. And the sudo-feedback system is not set up to improve the relevance of the results.

For future tasks, we are looking forward to optimizing the retrieval function and implement a sudo-feedback system when labeled data is not available.

REFERENCES

1. https://www.googleapis.com/auth/books
2. https://en.wikipedia.org/wiki/Genre

Contact: masghar@umich.edu, lushi@umich.edu