Overview on NLP techniques for content-based recommender systems for books

Melania Berbatova

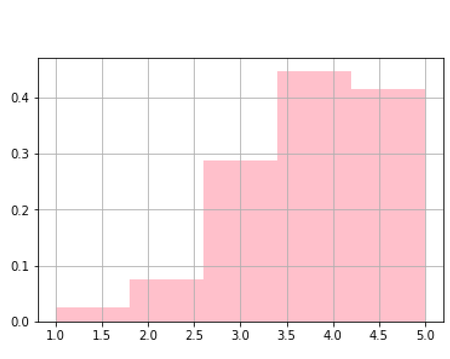
Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”

Introduction

- **Recommendation systems** provide users with recommendations for new content these users might be interested in (music, movies, books, etc). Recommendation systems can be divided into three main types: Collaborative Filtering (CF), Content-based Filtering (CBF) and Hybrid.
- **CF systems** analyze users' interactions with the items (ratings) to create recommendations.
- **CBF systems** use semantic information (metadata) about the items in the system.
- **Hybrid systems** are a combination of these two approaches. Their accuracy is usually higher.
- **Book recommenders** can take advantage of huge number of the metadata features, including title, author, year of publication, publisher, genre, summary, outline or even the whole text of the book.
- **User preferences**, in the form of custom tags, can be used to enrich the available metadata.

Dataset

- We are using the **goodreads-10k** dataset.
- The dataset consists of 5 files with the **10,000 most popular books** and their medatada, **5,976,479 ratings** and **34,251 tags**.



to-read	9983
favorites	9881
owned	9858
books-i-own	9799
currently-reading	9776
library	9415
owned-books	9221
fiction	9097
to-buy	8692
kindle	8316
default	8239
ebook	8054
my-books	7561
audiobook	7242
ebooks	7203
wish-list	7192
my-library	7000
audiobooks	6862
i-own	6670
adult	6604

Related work

Overview

- Previous research on book recommenders exists on the the Book-Crossing, LitRec, LibraryThing, INEX, Amazon reviews, and goodbooks-10k datasets.
- Currently there are 11 unique papers in English on recommender systems retrieved by Google Scholar when search is performed for goodbooks-10k.

Authors	Dataset	Features	Evaluation metrics	Algorithms	Dataset creation
Le (2019)	goodbooks-10k	ratings	MAP, CC, MPS, MNS, MDS	cosine similarity	test set - of 5-star ratings
Greenquist et al. (2019)	goodbooks-10k	tf-idf vectors	RMSE	cosine similarity	5+ ratings per user
(Alharthi and Inkpen, 2019)	Litrec	linguistic and stylometry features	precision@10, recall@10	kNN	10+ rating per user

Table 1: Overview of recent published papers

Feature engineering for the last approach included:

- **content and psychological word categories** using LIWC 2015 dictionary: percent of latinate words, function words, affect words, social words, perpetual processes etc.
- **six stylistic styles** using GutenTags tagger: informal vs. literary, concrete vs. abstract and subjective vs. objective
- **number of characters and number of locations** mentioned by LitNER text readability measurement - Flesch reading ease score.

Using **Extreme trees** (ET) algorithm, authors achieved 37% for precision@10 and 17% for recall@10.

Critiques and observations

- Lack of standardization in dataset preparation - definition of ”like”, data selection, test/train split
- Difference in evaluation metrics used leads to non-comparable results
- Lack of usage of deep learning for natural language processing
- Hierarchy, ambiguity and synonyms in tags not taken into account

Experiments

Experiments implemented simple CBF and CF systems. The following were used:

- Python library **TensorRec**
- 80%-20% train-test split
- Weighted margin-rank batch (WMRB) loss
- Evaluation metric **recall@k with k=10** as an evaluation metric.

$$recall\ at\ k = \frac{\#\ of\ recommended\ items\ at\ k\ that\ are\ liked}{\#\ of\ liked\ items}$$

Results

Algorithm	Parameters	Recall@10
CF	Like = 3+ rating	4.69%
CF	Like = 4+ rating	5.52%
CB	Like = 3+ rating	0.83%
CB	Like = 4+ rating	0.98%

Table 2: Results

Problems

For content-based approach, algorithms could not be trained with bigger feature sets, as there were memory errors both locally and in popular cloud service.

Suggested NLP improvements

Using tags information

- Tags of a book show its genres, readers intents, books features, awards, authors, etc., and many of the tags have similar or equal meanings.
- The set of tags per book is similar to textual data, except that there is no sequence between the tags.
- Linguistics resources and predefined dictionaries of book genres can be used for extracting features about genres and style.
- Alternatively, the bag-of-words representation of tags together with LSA, or the weighted average of embeddings of the tags‘ tokens can be used as a book representation.

Other suggestions

- **Data enrichment** - Scraping books descriptions from GoodReads website or from Amazon.com books pages.
- **Alternative approach** - to think of the recommendation task as a **classification problem** - ”like” versus ”dont like” set of books. In this setup, SVM or Naive Bayes, or deep learning algorithms such as RNN or LSTM can be used.

Conclusion

Many NLP techniques can be used in CBF for the recommendation of books, separately or in a hybrid with CF system. However, these techniques should be accompanied with standardized methods for dataset creation and results evaluation, so the results obtained are comparable to those from similar research.