TU Dortmund

Data Mining Cup 2021

LDA-based Book Recommender

Lecturers:
Prof. Dr. Emmanuel Müller
Dr. Daniel Horn
Dr. Michel Lang
M.Sc. Jonas Rieger

Author: Hassan Ali

September 30, 2021

Contents

1. Introduction

Recommender systems enable companies with voluminous data to provide personalized item recommendations to end users (Ricci et al. 2011, p. 1-2). Many online platforms implement such sophisticated algorithms. For instance, Pandora recommends audio based on someone's music consumption; Facebook suggests friends depending on a person's circle of acquaintances; and Netflix comes up with movies based on an individual's streaming history.

Service providers like those mentioned above make use of recommenders for a variety of reasons. Besides increasing the quantity of items sold, these techniques also stimulate diversity in sales. Additionally, recommenders help companies in better understanding the preferences of their customers. This can be leveraged to improve user experience and inculcate a sense of brand loyalty.

There are several approaches to building recommender systems from a methodical point of view. Some systems are content-based. These suggest items purely on the basis of the degree of similarity between the features of products bought by a user in the past, and the attributes of other items in the database. Collaborative filtering techniques on the other hand, recommend items that other users with matching tastes have previously preferred. Besides these two popular methods, hybrid approaches also exist which are an amalgam of both strategies (Ricci et al. 2011, p. 11-13).

This report is closely related to the above-mentioned examples. Here the items under consideration are books and the primary concern is training a content-based recommender to suggest similar products. This evaluates the degree of resemblance between items by processing their descriptions via Latent Dirichlet Allocation (LDA). These suggestions of matching books are compared with those generated by two naive alternatives. The first of these is a random sampler of books written by the same author or publisher while the other suggests top selling books of the same main topic.

The statistical analysis of the data comprises of four parts. Firstly, the data is preprocessed to deal with items that have the same book title and author. In the second part, the variables are analysed to investigate the underlying structure of the available data sets. Part three involves a description of the three recommender systems and their fallback strategies. Thereafter, eight handpicked item IDs are introduced and their recommendations are listed. Finally in part four, the recommendations generated by each recommender for 3 of these 8 item IDs are evaluated.

The goal of the report is to train and assess the aforementioned LDA-based recommender which consists of four LDA models trained with different hyperparameter settings. The results of these base learners are averaged to produce mean similarity scores between books. On the other hand, the naive recommenders do not require any training. Their results are obtained by simple scans, random samples or trivial manipulations of the dataframes.

The resulting LDA ensemble provides decent recommendations for books: the quality depending on the content of the descriptions. In general, short descriptions or those containing many irrelevant keywords lead to poor results. In such cases, the two simpler approaches outperform the LDA-based recommender.

Besides this Introduction, the report consists of four more sections. The Problem Statement provides thorough descriptions of the variables in the datasets. For instance, around three-fourths of the products have missing descriptions: the variable used to compute the results for the LDA recommender. The Methods portion of the report provides formulas and assumptions for LDA, ensemble learning and box plots. The Statistical Analysis section describes the preprocessing of the dataset; provides an analysis of the variables; and assesses the quality of the LDA-based recommender by comparing its results to its relatively naive counterparts . Finally, the Summary concisely rehearses the most important results. It discusses potential improvements to the experiment, such as collecting more data, and suggests ways in which it can be extended.

2. Problem Statement

Three data sets have been used in this report. Two of these, i.e. the "items" and the "transactions" dataframes are taken from the prudsys AG Data Mining Cup 2021 (AG 2021). These can be found online at `https://www.data-mining-cup.com/reviews/dmc-2021/`.

The items data contains 78334 unique instances of the variable "itemID" which is a positive integer ranging from 1 to 5 digits in length. Including the item ID, there are 6 variables in this data set: all qualitative. The variables "title", "author" and "publisher" indicate the title of a book, its writer and publishing company respectively.

The variable "main topic" is a cryptic string denoting a hierarchical classification of the main theme of each book. While its leftmost character indicates the highest level, the

rightmost points to the lowest stage in the hierarchy. These subject categories for the books are defined by EDItEUR Limited and version 1.4 of the thema has been followed in this report (Limited 2020). The main topics vary in length and the hierarchy is not necessarily followed down to its lowest level in the dataset.

If the main topic string has no "-" symbol, then each character denotes a category. So, if the main topic is "NKP" for instance, then "N" indicates the broad category of History and Archaeology; "K" signifies the subcategory of Archaeology; and "P" stands for the bottom-most category of Environmental Archaeology. On the other hand, if the main topic string does have a "-" symbol, then each set of characters separated by the hyphen indicates a category. For instance, in the main topic "3MN-DK-G", "3MN" indicates the time period of the 19th century while "DK-G" indicates the era of the Danish Golden Age from 1800 to 1850.

The variable "subtopics" is a list of strings that indicates the subtopics of the associated book. The list of subtopics is given inside square brackets "[]" and the subtopics are separated by commas "," in the data set. This variable pinpoints themes besides the main topic that occur in the book.

The transactions data set contains information on user interaction with products in separate online sessions. It consists of 5 variables out of which "sessionID" and "itemID" are qualitative. These indicate the unique ID of each session and the item IDs of products that featured in it. The other three variables are quantitative: more specifically positive integers. The variable "click" indicate the number of times a product was clicked on during a session; "basket" denotes the quantity of the product that was added into the basket during a session; and "order" denotes the number of orders for a given item ID in a session.

The third data set has been obtained via web scraping from the Hugendubel website (*Buchhandlung Hugendubel: Online Medien- und Buchversand* 2021). The variable "description" in this dataframe is a string that outlines the details of a given item ID. Other features of the data set include: subtitle, recommended age, language, book type, price, availability, the number of pages and ISBN, etc.

The data quality suggests that preprocessing is required before training any recommender system. For example, the items data set contains multiple item IDs that correspond to the same book title. The book title "Peter Pan" for instance, has 16 different item IDs in the data set. Moreover, the variables "author", "publisher", "main topic" and "subtopics" have missing values. Furthermore, there are also missing descriptions

for books in the data obtained via web scraping. So, while using any of these variables for recommendations, a fall back strategy has to be specified.

The objectives of the report are, first of all, to preprocess the data and investigate the variables. Thereafter, a recommender system based on Latent Dirichlet Allocation is trained and its results compared with two other naive alternatives. Fall back strategies for the recommenders are also laid out, if needed. The top suggestions generated by the three recommenders for 8 chosen item IDs are specified and 3 of them are analysed in detail.

3. Statistical Methods

The following statistical methods and mathematical formulas are used. The software Python (Foundation 2021) version 3.9.7 has been used for analysis.

3.1. Box-and-whisker-plot

The box-and-whisker-plot is a graphical method of depicting the five number summary (minimum, first quantile, median, third quantile and maximum) of a set of observations. The "box" spans the interquartile range: the two lines at its edges indicating the lower and upper quantiles. A line through the box denotes the median. The "whiskers" extend to 1.5 times the interquartile range on both sides of the box (Everitt and Skrondal 2010, p. 61). The plot is also referred to as a box plot.

3.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic model for discrete data. While LDA can be used in domains such as collaborative filtering, content-based image retrieval and bioinformatics, it is primarily used for text corpora whereby each text is modeled as a distribution of latent topics.

The fundamental unit of discrete text data is a word or token. If a corpus consists of M documents, then for any document $m = 1, ..., M$ of the corpus, $W_n^{(m)}$ denotes a single token at position n. The set of all possible word tokens is the vocabulary indexed by $1, ..., V$. The vocabulary $\boldsymbol{W} = W_1, ..., W_V$ has size $V = |\boldsymbol{W}|$. A document consists of

$N^{(m)}$ word tokens where each $W_n^{(m)} \in \boldsymbol{W}$. Each document $m = 1, ..., M$ can thus be represented as a vector of words (Blei et al. 2003),

$$\boldsymbol{D}^{(m)} = (W_1^{(m)}, ..., W_{N^{(m)}}^{(m)}).$$

If K is the number of topics being modelled, then the set of topics is given by $\boldsymbol{T} = T_1, ..., T_K$. Each token $W_n^{(m)}$ in a document is assigned to a topic $T_n^{(m)} \in \boldsymbol{T}$. Consequently, for any document m, the vector of topic assignments is given by (Blei et al. 2003),

$$\boldsymbol{T}^{(m)} = (T_1^{(m)}, ..., T_{N^{(m)}}^{(m)}).$$

The number of times a word v in text m has been assigned to topic k is given by $n_k^{(mv)}$, where $k = 1, ..., K$, and $v = 1, ..., V$. Building on this, $n_k^{(\cdot v)}$ then denotes the total number of times a word v occurs in topic k in all documents. Similarly, $n_k^{(m\cdot)}$ signifies the total count of topic k over all words in document m. Finally, $n_k^{(\cdot\cdot)}$ denotes the total number of words assigned to topic k in all documents.

The underlying probability model (Griffiths and Steyvers 2004) can thus be written as follows,

$$\begin{gathered} W_n^{(m)} | T_n^{(m)}, \boldsymbol{\phi}_k \sim \text{Discrete}(\boldsymbol{\phi}_k), \\ \boldsymbol{\phi}_k \sim \text{Dirichlet}(\eta), \\ T_n^{(m)} | \boldsymbol{\theta}_m \sim \text{Discrete}(\boldsymbol{\theta}_m), \\ \boldsymbol{\theta}_m \sim \text{Dirichlet}(\alpha). \end{gathered}$$

Here α and η are parameters of the Dirichlet distributions of the topics per document and the words per topic, respectively. The set of parameters for LDA is given by K, α, η. Given these parameters, the algorithm assigns each token per document to one of the K topics.

The topic distribution of each text m is given by the vector,

$$\boldsymbol{\theta}_m = (\theta_{m,1}, ..., \theta_{m,K})^T \in (0,1)^K.$$

On the other hand, the vector of word distributions per topic is given by,

$$\boldsymbol{\phi}_m = (\phi_{k,1}, ..., \phi_{k,V})^T \in (0,1)^V.$$

These distributions can be estimated using the Collapsed Gibbs Sampler method (Griffiths and Steyvers 2004).

$$\hat{\theta}_{m,k} = \frac{n_k^{m\cdot} + \alpha}{N^{(m)} + K\alpha},$$

$$\hat{\phi}_{k,v} = \frac{n_k^{\cdot v} + \eta}{n_k^{\cdot\cdot} + V\eta}.$$

3.3. Jensen-Shannon distance

After model training, LDA produces vectors that describe the distribution of topics in each document. The similarity of these topic vectors, which represent probability distributions, can be found using measures like the Manhattan distance, χ^2 distance, Hellinger distance or the Jensen-Shannon divergence. The last of these is used this report. The Jensen-Shannon divergence is based on the Kullback–Leibler divergence and is symmetric. The square root of the Jensen-Shannon divergence gives the Jensen-Shannon distance. For two probability vectors $\mathbf{p}$ and $\mathbf{q}$, this distance is given by (Lin 1991),

$$\sqrt{\frac{D(\mathbf{p}||\mathbf{m}) + D(\mathbf{q}||\mathbf{m})}{2}},$$

where D is the Kullback–Leibler divergence and $\mathbf{m}$ is the point-wise mean of $\mathbf{p}$ and $\mathbf{q}$. The Kullback–Leibler divergence for two vectors $\mathbf{p}$ and $\mathbf{q}$ is in turn calculated as follows (Lin 1991),

$$D(\mathbf{p}||\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}.$$

3.4. Ensemble Learning

Instead of finding a single model that best fits the data, ensemble methods construct a set of base learners (called an ensemble) and then aggregate its results to predict the labels of new data points. Such a multi-model approach can reduce the risk posed by high variance. This is the case when, for example, a Random Forest is trained instead of a single regression or classification tree. In addition, ensemble methods can also reduce the bias of models (Polikar 2012). In this report, LDA models trained with different hyperparameter settings are combined to form an ensemble to improve the predictive

quality of the primary recommender system. The results of different LDA models are simply averaged by taking the mean to obtain the final similarity scores.

4. Statistical Analysis

4.1. Data Preprocessing

While there are more than 78000 unique item IDs in the items data set, the number of unique book titles is 72404. Hence, there are many instances of multiple item IDs corresponding to the same book title. In fact, there are more than 4000 book titles that are associated with more than one item ID.

Figure 1 shows a box plot of the number of item IDs per book title in the dataframe. The book titled "The Secret Garden" has the maximum number of matching item IDs: 25. The distribution is very right skewed. The box of the box plot and even its whiskers are all squashed onto the number 1 due to the enormous number of titles that have just a single corresponding item ID. Multiple item IDs per title is an exception, not the norm.

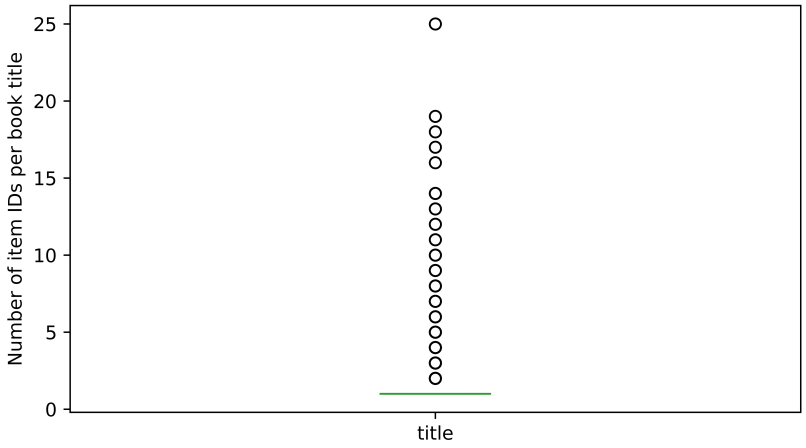

Figure 1: Box plot of the number of item IDs per book title in the items data set

Table 14 in the appendix shows one particular example of a recurring book title. There are 7 instances of the title "The Jungle Book". Since each one of these is also written by the same author, it has the same contents. A unique item ID is necessary to distinguish these seven products because they have been published by separate companies. Another noteworthy point is that despite being the same book, the main topic and subtopics of the books do not agree with each other in table 14. As for the main topics, none of

these classifications is strictly wrong. One might be better than the other. In the case of subtopics however, some instances of the same book title have no associated subtopics, yet others have one to three. And even when the number of subtopics is the same, there is a difference in the exact strings of those subtopics.

Another example given in table 15 in the appendix shows that some identical book titles do not have the same author. Different writers have authored books with the same titles. Obviously, these books vary content-wise. Hence, in the preprocessing step, it is important to retain these kinds of repetitions.

For item IDs with the same title and author, one instance is sampled at random. A new column named "oldID" is then added to the dataframe. This contains other item IDs pertaining to the same title and author combination. Old item IDs are also replaced in the transactions data set. Furthermore, another column containing book descriptions obtained via web scraping is also merged with the items data set. For all further analyses of the items and transactions data set in this report, the preprocessed versions of the dataframes are used.

4.2. Analysis of the Variables

After preprocessing, the items data set now has 74039 unique item IDs and 72404 unique book titles, with no missing values for any of these variables. 4.3% of the products have missing values for the author; 0.01% for the publisher; 0.3% for the main topic; 49.2% for the subtopic and 23.6% for the description. The languages of the books that were obtained via web scraping do not have any missing values. The pie chart in figure 2 shows that more than half of the products (56.6%) are in English while 36.9% of them are written in German, the second most frequent language.

Table 1 summarizes the distribution of the number of books per author, per publisher and per main topic in the preprocessed items dataframe. All three distributions are highly skewed. This makes the median a better indicator of central tendency than the mean, and the interquartile range the statistic of choice for gauging dispersion (instead of the standard deviation).

Figure 3 shows a box plot of the number of books per author in the preprocessed items dataframe. Note that the y-axis of the plot is logarithmic. The distribution is heavily right skewed with a median and interquartile range of 1. Out of a total of 35723 unique authors, just 9267 (or 25.9%) have more than one book title associated with them. The

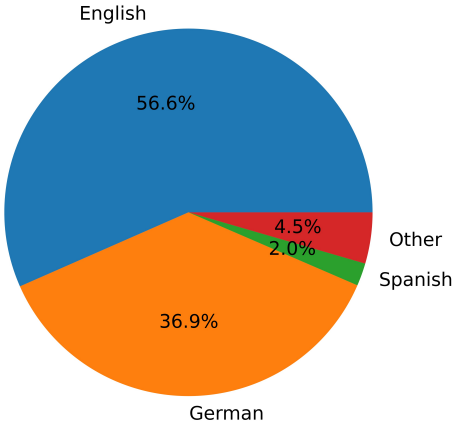

Figure 2: Percentage of books pertaining to each language after preprocessing

	Books per author	Books per publisher	Books per main topic
Mean	1.98	10.54	106.64
Standard deviation	8.37	73.63	483.60
Minimum	1	1	1
25th percentile	1.00	1.00	1.00
50th percentile	1.00	1.00	4.00
75th percentile	2.00	4.00	19.25
Maximum	1370	3397	6154

Table 1: The univariate distribution of the number of books per author, publisher and main topic in the preprocessed items data set

most frequently occurring authors in the preprocessed items dataframe are given in table 16 in the appendix. The author "Garcia Santiago" has the highest number of associated books: 1370. These are in Spanish and account for the overwhelming majority of the total 1495 books written in this language.

Figure 4 shows a box plot of the number of books per publisher in the preprocessed items dataframe. Once again, this is a highly right skewed graph, with the presence of outliers causing the mean to be more than ten times the median according to table 1. The median number of books per publisher is 1 while the interquartile range is 3. Table 17 in the appendix shows the five most frequently occurring publishers in the data set. The most oft-repeated of these is "Books on Demand" with a count that is just shy of 3400. The other four publishers also have a count that is above one thousand. There are 7025 unique publishers, out of which approximately 400 occur just once in the data set.

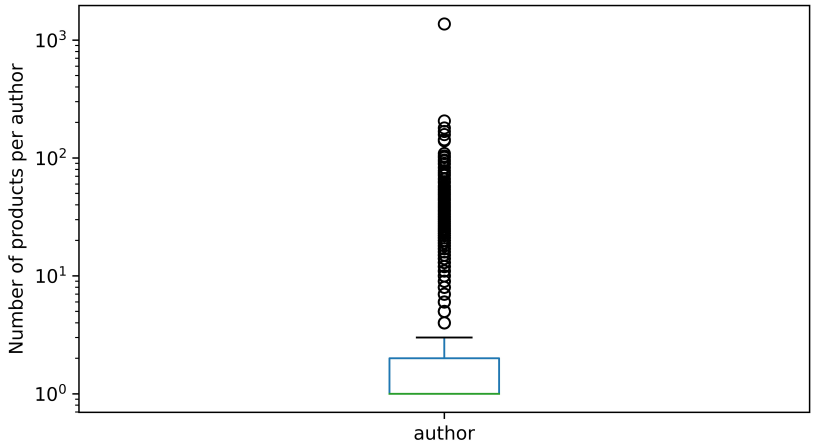

Figure 3: Box plot of the number of products per author in the preprocessed items data set

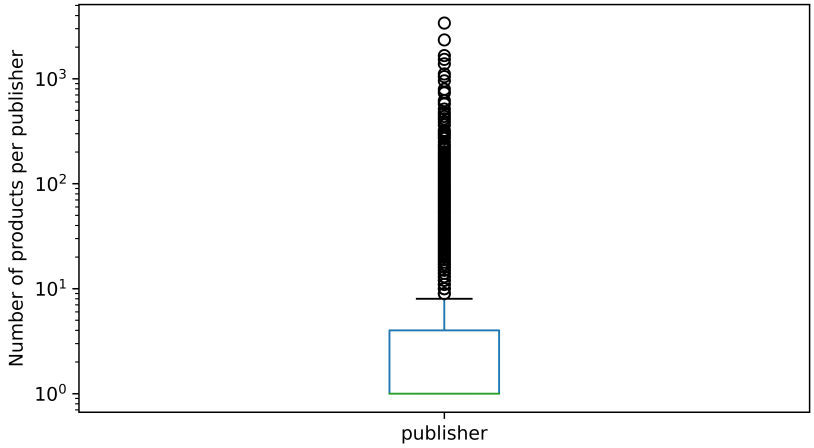

Figure 4: Box plot of the number of books per publisher in the preprocessed items data set

Figure 5 shows the distribution of the number of books per main topic after preprocessing. The distribution is once again right skewed; though not to the degree that the previous box plots for books per author or publisher are. The median number of books associated with each main topic is 4 while the interquartile range is just above 18. The number of books pertaining to each main topic varies more than the books per publisher, which in turn has more variation than the number of books per author (c.f. table 1).

Table 2 shows the 10 most frequently occurring main topics and subtopics in the preprocessed items data set. The most frequent main topic "FM", which occurs more than 6000 times, and the most oft-repeated subtopic "YF" which occurs more than 1300

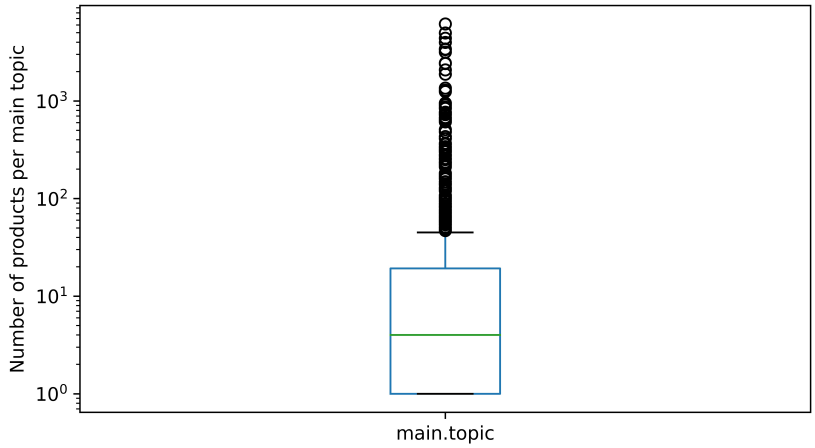

Figure 5: Box plot of the number of books per main topic in the preprocessed items data set

times, both do not follow the category hierarchy to its bottom-most level. "FM" stands for "Fantasy" and is also the sixth most frequent subtopic. "YF" which stands for "Children's / Teenage fiction and true stories" is also the fifth most oft-repeated main topic. Moreover, six of the top ten main topics and five of the top ten subtopics by count start with this particular letter combination (i.e. "YF").

Main topic	Count
FM	6154
YFB	4979
FL	4391
YFH	3980
YF	3413
YFC	3364
YBG	3167
YFCF	2447
FMB	2419
YFJ	2082

(a) Most frequent main topics

Subtopics	Count
YF	1386
FL	955
YFQ	821
YFH	724
YFP	691
FM	680
1KBB	580
YFB	512
5AK	421
YBL	377

(b) Most frequent subtopics

Table 2: The ten most frequently occurring main topics and subtopics in the preprocessed items data set

The first letters of the main topics indicate the broad categories to which the books belong. Their frequencies of occurrence are shown in figure 6. The three most frequently repeated first letters "Y", "F" and "X" which refer to the categories "Children's, Teenage and Educational", "Fiction and Related items" and "Graphic novels, Comic books, Car-

toons", occur 46588, 24469 and 912 times, respectively. Together, these three make up for a whopping 97.2% of the products. Since there is a lot more training data that is fictional and aimed at children, we would expect the recommender to perform better on average on books featuring these themes, compared to non-fictional and adult-oriented products.

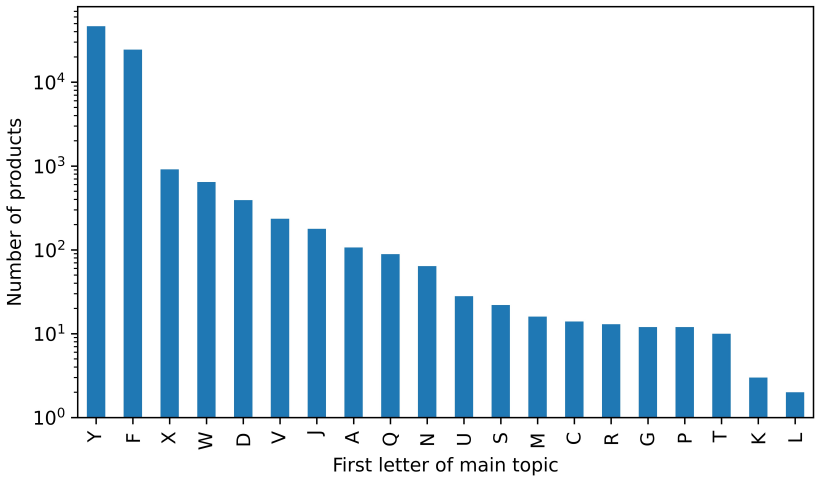

Figure 6: Frequency of incidence of the first letters of the main topics in the preprocessed items data set

Figure 7 shows the number of products pertaining to to each number of subtopics. A generally decreasing trend is visible: the higher the number of subtopics, the fewer the number of books associated with that many subtopics. 16810 items have one subtopic, while 36451 or around 50% of the products have none. In contrast, only 246 or just 0.33% of the books have no main topic. This greatly reduces the predictive power of subtopics in finding recommendations for books. We cannot even make comparisons with about half of the items data set when looking at subtopics alone.

Table 3 outlines the distribution of the number of clicks, baskets and orders per item ID in each session in the preprocessed transactions dataset. There are no missing rows in this dataframe. The mean value of the number of clicks per product per session, is 1.23 which is more than 8 times higher than the number of baskets; a value of 0.14. The mean number of orders per product in each session are even lower; a meagre 0.05. The distributions of all three variables are very right skewed. The minimum value is zero for each. The 25th, 50th and 75th percentiles for the clicks are all one and the maximum is a whopping 118. Similarly for the baskets and orders, all percentiles are zero up until the 75th percentile while the maximum is 293 and 28 respectively. This shows that most

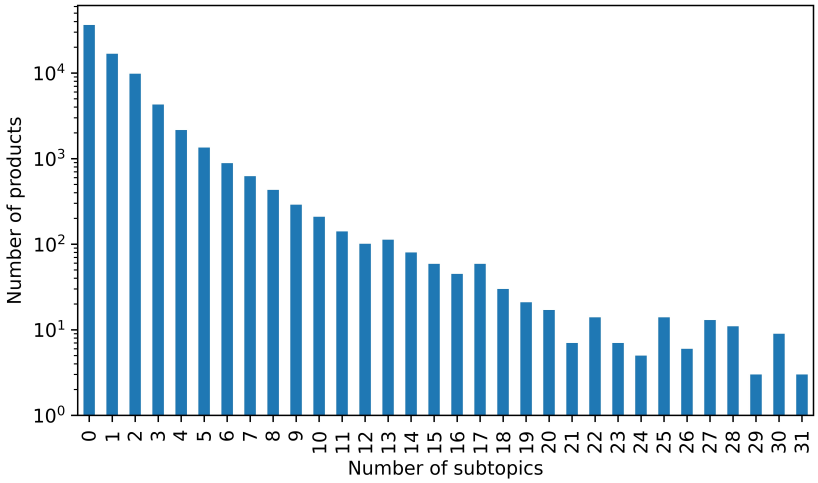

Figure 7: Line graph of the number of subtopics with their frequencies of occurrence in the preprocessed items data set

of the items have just 1 click and no basket or order in each session. So in most of the sessions, customers browsed through products and clicked on them with the intent of perhaps informing themselves about details such as price and availability, instead of actually buying books there and then.

	Clicks	Baskets	Orders
Mean	1.23	0.14	0.05
Standard deviation	1.07	1.11	0.27
Minimum	0	0	0
25th percentile	1.00	0.00	0.00
50th percentile	1.00	0.00	0.00
75th percentile	1.00	0.00	0.00
Maximum	118	293	28

Table 3: The univariate distribution of clicks, baskets and orders per product per session in the processed transactions data set

93.55% of the item IDs in each session in the transactions dataframe have one or more clicks. For baskets, this percentage is just 12.32% and for orders, an even lower 4.63%. So, the likelihood of an order is lower than that of a basket, whose chances are in turn smaller than those of a click. The total number of clicks in all sessions is 450287; for baskets this sums up to 51559; while there are 17674 orders placed in all sessions.

The most ordered products are shown in table 18 in the appendix. Four out of the 5 top selling books have German titles. However, according to the data obtained by web

scraping, the remaining item with an English title "The Hate U Give" is also in the German language. This suggests that the transactions data may have been collected from buyers who are primarily German-speaking.

Figures 9, 10 and 11 in the appendix show scatter plots of the number of baskets against clicks; orders against clicks; and orders against baskets per product per session in the preprocessed transactions data set. These figures show that when comparing any two variables with each other, most of the data points occur near zero. The sparsity of the scatter plots increases as we move away from the origin. Moreover, in all three plots, when the value of one of the variables is high, that of the other is approximately zero. For example, in figure 9, for points where the clicks are more than 20, the corresponding number of baskets is around zero; and vice versa. In figure 11, when the number of orders for some product in a given session are 5 or more, the corresponding number of baskets is always higher than zero.

Figure 8 is a box plot showing the number of item IDs per session in the preprocessed transactions data set. There are 271983 sessions out of which more than 200,000 have just one product. There are no sessions that do not have any items. The maximum number of items looked at in any given session is 213 and the mean and median are 1.34 and 1 respectively. This distribution is also right skewed.

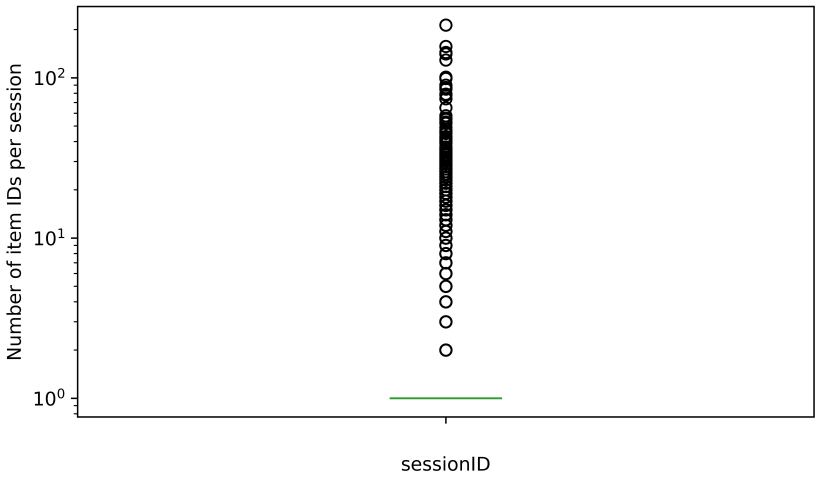

Figure 8: Box plot of the number of item IDs per session in the preprocessed transactions data set

4.3. Recommender Systems Used

The primary recommender system used in this report is based on Latent Dirichlet Allocation and utilizes item descriptions. As already mentioned, descriptions have been successfully scraped for more than 75% of the books. For LDA, Python's Natural Language Processing library "gensim" (Rehurek and Sojka 2011) and the Natural Language Toolkit "nltk" (Bird et al. 2009) have been used. Before training the LDA model, the descriptions of the books have to be processed by removing their stopwords. These are the most commonly occurring words of a language (e.g. "that", "and", "or", "your", etc. in English) and do not add much value to the meaning of a piece of text. Stopwords are removed for the 6 most frequently occurring languages in the dataset: English, German, Spanish, French, Italian and Portuguese.

Four LDA models are trained. These differ only in the value of the hyperparameter "number of topics". This is set to 20, 30, 40 and 50 for the four models respectively. The other hyperparameters of α and η which denote the topic density per document and the word density per topic respectively, are set to "auto" for all four models. When books are to be recommended, similarity scores are found using the Jensen-Shannon distance for each of the four models. The most similar books have the smallest distance. Similarity scores are calculated by subtracting the Jensen-Shannon distance from 1. These scores from each of the four models are then averaged by taking their mean. In case a description is not available for the book whose recommendations are sought, a fallback strategy has to be employed. Since the LDA recommender is purely content-based, it makes sense that the fallback strategy is also of the same kind. In this report, we randomly select recommendations with the same main topic from the preprocessed items data frame.

The other two recommenders are relatively naive. Their recommendations are compared with the LDA-based recommender. The first of these extracts the author and publisher of the book whose recommendations need to be found. If the book's author has also written other books, those are randomly selected. Otherwise, if the author does not have any other product attributed to him/her, then books from the same publisher are randomly recommended. This recommender shall be referred to as the author/publisher recommender from hence forth. The other naive recommender is based on both the transactions and the items data. It extracts the main topic of the book for which recommendations are sought and then finds the most ordered books of that main topic from

the transactions data set. This recommender is simply referred to as the transactions recommender from this point onward.

We take the following 8 item IDs and compare the recommendations of the three recommenders for each of them: 12152, 13382, 24603, 47675, 56180, 60644, 61261 and 72826. These item IDs and their corresponding titles and main topics are given in table 4. Item IDs 13382 and 60644 in the list are associated with the same book title and have the same author. Hence they are the same book and have identical recommendations.

Index	Item ID	Title	Main topic
1	12152	White Rose	YFB
2	13382, 60644	Die Macht der Drei	FL
3	24603	The Hobbit	FM
4	47675	Cirkus Cannelloni i traditionens snara: Swedish Edition of Circus Cannelloni Invades Britain	YFQ
5	56180	There it is! Da ist es!: A search and find book in English and German	YBLD
6	61261	Die drei ??? Kids 45 - Ein Fall für Superhelden (drei Fragezeichen)	YFCF
7	72826	Mandalas Weihnachten	YFB

Table 4: The 8 item IDs whose recommendations are compared in this report, along with their title and main topic

Recommendations for the 3rd, 6th and 7th item IDs from the list in table 4 are discussed in the next subsection. The remaining item IDs are not discussed but their recommendations are given in the appendix.

Out of the non-discussed item IDs, a description is only available for the 2nd item in the list in table 4. So, LDA-based recommendations are only possible for this book (associated with item IDs 13382 and 60644). For items number 1, 4 and 5 in the table, the fallback strategy of the LDA recommender (randomly sampling books with the same main topic) is employed. The LDA or its fallback strategy's recommendations for items 1, 2, 4 and 5 are given in tables 19, 20, 21 and 22 respectively in the appendix.

For book number 5 in table 4, there are no other items by the same author or publisher, so the author/publisher recommender does not output any recommendations in this case. Items number 2 and 4 in the table have other books written by the same author. For item 1, there are no other books written by the same author, but books written by the same publisher can be found and these are recommended. The author/publisher recommender's suggestions for items 1, 2 and 4 are given in tables 23, 24 and 25 respectively in the appendix.

The transactions-based recommendations for items 1, 2, 4 and 5 in table 4 are given in tables 26, 27, 28 and 29 respectively in the appendix.

4.4. Analysis of the Recommendations for Some Items

4.4.1. Item ID 24603

The item ID 24603 corresponds to the title "The Hobbit" written by John Ronald Reuel Tolkien. The top 5 recommendations for this book by the LDA recommender are shown in table 5. The similarity scores and main topics for these suggestions are also given. LDA appears to have filtered out books from other languages as these recommendations are also, like the book itself, in English.

Title	Score	Main topic
The Pocket Roverandom	0.734	FM
Little Women and Good Wives	0.717	YFA
Alice in Wonderland - Illustrated by Dudley Jarrett	0.712	YFA
After the Sundial	0.695	FL
Metaphorosis 2018	0.695	FM

Table 5: Top 5 recommendations for item ID 24603 by the LDA-based recommender, their similarity scores and main topics

Part of the description for "The Hobbit" is as follows:

"A new edition of the classic work of fantasy, with 150 new illustrations specifically designed for children, but which can be also enjoyed by adults..."

"The Hobbit" has the main topic "FM" which refers to "Fantasy". According to the description, this is a classic fantasy book for both children and adults, but mainly targeted at the former. Moreover, the book is illustrated with drawings.

A glimpse at the first LDA-based recommendation in table 5 shows that it has the same main topic as "The Hobbit". Moreover, the word "fantasy" also appears in its description and it is also written by the same author. According to its description, it also features "Tolkien's own paintings and line drawings". So this seems to be a very good first recommendation. None of the other four recommendations are authored by Tolkien.

The second and third books in table 5 have almost equal scores, which are 0.02 points below that of the best recommendation. Their main topic "YFA" refers to the theme "Children's / Teenage fiction: Classic fiction". Fiction can be deemed as a broad category that includes fantasy related themes but is not limited to them. While both are based on non-truth, fantasy revolves specifically around the supernatural and mystical such as dragons or wizards. So these two are also very good recommendations. They are aimed at children and are classic works of literature. The word "classic" even appears in their descriptions like it does in the description of "The Hobbit". According to its description, "Alice in Wonderland" is also illustrated with paintings and is aimed at both children and adults.

As mentioned in table 5, the fourth book's main topic is "FL" or science fiction. Science fiction revolves more around technologies like cyborgs, aliens and spaceships. So it is slightly divergent from fantasy. However, this book might have been recommended because words associated with themes of fantasy are found in its description. An excerpt from the description reads: "...that focuses specifically on science fiction works, and can be viewed as a companion volume to her earlier collection, Salt of the Air which focused on fable, myth, and fantasy...".

It appears that LDA is not aware of the context in which the words "fable, myth, and fantasy" are used in the description above. These words have been used for a previous literary work named "Salt of the Air" by the author of "After the Sundial". Nonetheless, the recommender assumes that they pertain to this book and therefore produces a higher similarity score than it should have.

The fifth recommendation in table 5 is of the same main topic as "The Hobbit" and is also a work of fantasy. So this is also a good recommendation. An excerpt from its description reads: "Fifty-two great science fiction and fantasy stories...". It may have been the case that LDA learned to put word tokens related to science fiction and fantasy into the same latent topic, instead of different ones. This may also explain the fact that

the similarity scores of the last two recommendations are identical (up to 3 decimal places) even though one book is related to science fiction and the other to fantasy.

The recommendations by the same author/publisher recommender are given in table 6. All these recommendations are by the same author. Two of them are also The Hobbit books: one simply being a color illustrated version and the other a pocket edition. The second and fifth recommendations, "Die Kinder Húrins" and "The Lord of the Rings" are also fantasy books; although the former is in German. According to its main topic, the fourth recommendation "Letters from Father Christmas" is a children's fantasy book. So, this recommender has also done a respectable job. It not only recommends some books of the same genre, but also recommends two different editions of "The Hobbit".

Title	Main topic
The Colour Illustrated Hobbit	FM
Die Kinder Húrins	FMB
The Pocket Hobbit. 75th Anniversary Edition	FBC
Letters from Father Christmas	YFH
The Lord of the Rings	FM

Table 6: Top 5 recommendations for item ID 24603 by the same author/publisher recommender and the main topics of the books

Title	Main topic	Orders
City of Glass	FM	27
City of Ashes	FM	21
Das Lied von Eis und Feuer 01. Die Herren von Winterfell	FM	17
The Invisible Life of Addie LaRue	FM	12
Schattenkämpfer	FM	9

Table 7: Top 5 recommendations for item ID 24603 by the transactions-based recommender, the main topic and the number of orders of the books

The recommendations for the transactions-based recommender are shown in table 7. All of them are also fantasy books since the algorithm picks the most sold books with the same main topic. Two of these are in German while none is identical to the ones generated by the previous two recommenders. Furthermore, none of them shares the same author with "The Hobbit". The most ordered book "City of Glass" lacks a description. The descriptions for the English book titles lack keywords like "fantasy" or "classic" and these keywords are obviously not present in the two German books (since these are

English words and their descriptions are in German). This partially explains why LDA fails to recommend any of these five books.

4.4.2. Item ID 61261

The item ID 61261 is associated with the title "Die drei ??? Kids 45 - Ein Fall für Superhelden (drei Fragezeichen)" and is a German book. It is authored by Ulf Blanck and its main topic "YFCF" points to the category of "Children's / Teenage fiction: Crime and mystery fiction". The top 5 recommendations for this product generated by the LDA recommender are shown in table 8. Once again, LDA appears to have filtered out books from other languages. All recommendations are also in German.

The author and main topic of the first recommendation are the same as that of "Die drei ??? Kids 45". However, none of the other four recommendations are associated with a matching author or main topic.

The score for the best recommendation is an impressive 1.0. Upon further examination, it turns out that both this book and "Die drei ??? Kids 45" have the same description whose entirety comprises of just 4 words: "Erzählt von Ulf Blanck". So it is not surprising that the similarity score is perfect. Such a short description which consists of only 3 word tokens after omitting the stopword "von" means that the recommender has very little information to find matching products. The descriptions of none of the other four recommendations contain any of these words. In these cases, the recommender must have found other words from the same latent topics as this word triple.

Title	Score	Main topic
Die drei ??? Kids 46 - Jagd auf das Dino-Ei (drei Fragezeichen)	1.000	YFCF
Die Ameisenkolonie	0.902	YFB
Der unheilige Gral	0.798	FM
Der kleine Hobbit Veredelte Mini-Ausgabe	0.798	FBC
Der kleine Hobbit Normalformat	0.798	FBC

Table 8: Top 5 recommendations for item ID 61261 by the LDA-based recommender, their similarity scores and main topic

The second recommendation has the main topic "Children's / Teenage fiction: General fiction" and its score is considerably higher than the last three suggestions. Both of the top recommendations are fiction books for children. The similarity scores of the last three recommendations are equal up to 3 decimal places. The main topics "FM"

and “FBC” stand for “Fantasy” and “Classic Fiction”, respectively. So these books are fantasy or fiction but unlike the two best recommendations, are not necessarily aimed at children alone. It can be said that the recommender still does a decent job; though a longer description of the original book would have improved the results. Note that the last two books are different editions of the Hobbit in the German language. The last three books all contain the word token “Klassiker”, indicating that they are classic books of their respective genres.

The recommendations produced by the same author/publisher recommender are given in table 9. All books here have the same author (Ulf Blank), are in the same language (German) and even have the same main topic: all being crime and mystery fiction books. Four of these books are even from the same regular series of “Die drei ??? Kids”. The third book “Die drei out Kids und du: Labyrinth der Piraten” is a special volume published in addition to the regular series. All in all, these recommendations appear to be superior to the ones generated by LDA.

Title	Main topic
Die drei ??? Kids 24: Im Bann des Zauberers (drei Fragezeichen)	YFCF
Die drei ??? Kids 05: Flucht in die Zukunft (drei Fragezeichen)	YFCF
Die drei out Kids und du: Labyrinth der Piraten	YFCF
Die drei ??? Kids 51 - Tatort Kletterpark	YFCF
Die drei ??? Kids 18: Mission Maulwurf	YFCF

Table 9: Top 5 recommendations for item ID 61261 by the same author/publisher recommender and the main topics of the books

Title	Main topic	Orders
Rico, Oskar und die Tieferschatten (Rico und Oskar 1)	YFCF	170
Level 4. Die Stadt der Kinder	YFCF	37
One of Us is Lying	YFCF	27
Beschützer der Diebe	YFCF	25
Oskar und das Geheimnis der verschwundenen Kinder	YFCF	14

Table 10: Top 5 recommendations for item ID 61261 by the transactions-based recommender, the main topic and the number of orders of the books

Table 10 shows the recommendations produced by the transactions-based recommender. Four of these are in German while one is in English. Since Ulf Blank’s books are not the best-sellers in the crime and mystery fiction category, they do not appear in this list. These recommendations are still better than those of the LDA recommender because

they are of the same genre and they can be said to be based on empirical data as people actually purchased these books. The first book in these recommendations is the most sold book in the preprocessed transactions data set. Overall however, the recommendations generated by the same author/publisher recommender are the best here since they are all in the same language and because almost all of them are books of the same series as "Die drei ??? Kids 45".

4.4.3. Item ID 72826

The item ID 72826 corresponds to the title "Mandalas Weihnachten" and author Andreas Abato. This German book contains Christmas motifs for children to color in and features mandalas which are circular geometric patterns. Its description is rather detailed- an excerpt from which is given below. Some of the keywords here are "Mandalas", "Kindergärten" and "Ausmalen" or coloring.

"...Kinder können sich stundenlang mit Mandalas beschäftigen. Deshalb gibt es viele Erzieherinnen, die Mandalas in Kindergärten einsetzen. Das Ausmalen fordert Geduld, Farbgefühl und Konzentration...."

The description of the book also includes names of other books in its series:

"...Alle MANDALA-Bände dieser Reihe: Mandalas Rund um den Bauernhof · Mandalas Pferde · Mandalas Autos · Mandalas Weihnachten · Mandalas Elfen Drachen Zauberer · Mandalas Alphabet · Mandalas der Kelten · Mandalas Ornamente · Mandalas Liebe Rosen Herzen."

The LDA-based recommendations for the product are shown in table 11. Note that the first three of these are titles from the excerpt of the description above. So, the recommender has done a decent job in extracting books from the same series.

All five recommendations in table 11 are mandalas coloring books for children from the same author and in the same language: German. This makes them very good recommendations. The similarity scores for the first three of these are very high; the reason being that their descriptions are the same as for "Mandalas Weihnachten" except for the first sentence. This first sentence, which is the only unique component of the description for these three books, is slightly longer in the case of the third book "Mandalas Liebe Rosen Herzen". Hence, this book has more dissimilar word tokens compared to the previous two items which explains the slightly lower similarity score. Nonetheless,

Title	Score	Main topic
Mandalas rund um den Bauernhof	0.992	YFB
Mandalas Autos	0.986	YNPH
Mandalas Liebe Rosen Herzen	0.921	YFB
Meine Mandalas - Mein Ausmalbuch - Wunderschöne Mandalas zum Ausmalen	0.755	YNPH
Meine Mandalas - Mit Freude Ausmalen - Wunderschöne Mandalas zum Ausmalen	0.755	YNPH

Table 11: Top 5 recommendations for item ID 78286 by the LDA-based recommender, their similarity scores and main topic

the similarity scores of these three products are more similar to each other than to the last two recommendations.

The fourth and fifth books also have a few sentences that match exactly with some of those present in the description of "Mandalas Weihnachten". However, since there are fewer matching word tokens, the score is much lower than the top three recommendations.

"Mandalas Weihnachten" pertains to the main topic "YFB" or "Children's / Teenage fiction: General fiction". Two of the LDA-based recommendations also lie in the exact same category. The other three have the main topic "YNPH" or "Children's / Teenage general interest: Handicrafts". As all these products are mandalas coloring books for children, it can be argued that placing them in the category of handicrafts for children makes more sense than classifying them as general fiction for kids. Nonetheless, despite the slightly less accurate classification of "Mandalas Weihnachten" as a fiction book, the LDA recommender has found very good matches.

The recommendations by the same author/publisher recommender are shown in table 12. All these books are written by the same author. The preprocessed items data set contains a total of 14 book titles associated with the author Andreas Abato. All of these are mandalas coloring books in German. Therefore, this recommender luckily outputs very good recommendations. However, unlike LDA, only one item in the book series

containing “Mandalas Weihnachten” is present in the 5 suggested books. So for this particular random sample, LDA has done a better job than this recommender.

Title	Main topic
Meine Mandalas - Das macht mir Spass! - Wunderschöne Mandalas zum Ausmalen	YNPH
Meine Mandalas - Cool und kreativ - Wunderschöne Mandalas zum Ausmalen	YNPH
Mandalas rund um den Bauernhof	YFB
Meine Mandalas - Meine schönsten Muster - Wunderschöne Mandalas zum Ausmalen	YNPH
Meine Mandalas - Für coole Kids - Wunderschöne Mandalas zum Ausmalen	YNPH

Table 12: Top 5 recommendations for item ID 72826 by the same author/publisher recommender and the main topics of the books

The five recommendations by the transactions-based recommender are shown in table 13. These recommendations have the same main topic as “Mandalas Weihnachten”, which is less precisely labelled as fiction for children instead of handicrafts for them. Nevertheless, none of these five is a coloring book or is related to mandalas. So these predictions are definitely not nearly as good as those generated by the other two recommenders.

Title	Main topic	Orders
Die Welle	YFB	125
The Hate U Give	YFB	95
Harry Potter 3 und der Gefangene von Askaban	YFB	46
Gangsta-Oma	YFB	44
Der Sprachabschneider	YFB	43

Table 13: Top 5 recommendations for item ID 72826 by the transactions-based recommender, the main topic and the number of orders of the books

5. Summary

This report concerns a content-based recommender for books that employs Latent Dirichlet Allocation. The data used includes the items and transactions datasets taken from the prudsys AG Data Mining Cup 2021. The items dataset features 78334 unique item IDs. Besides this variable, the title, author, publisher, main topic and subtopics of the products are also provided. The transactions data summarizes how users have interacted with the books in distinct online sessions. For each session, its ID, the item IDs of books browsed in it and the number of clicks, baskets and orders pertaining to them can be found. Additionally, web scraping is employed to extract, besides other attributes, product descriptions and their languages from the Hugendubel website.

The items dataset contains identical book titles by the same author which are associated with multiple item IDs. In the preprocessing phase, only one instance out of such repeated occurrences is retained to make each combination of title and author in the dataframe unique. The scraped book descriptions are also incorporated into the items dataframe and obsolete item IDs (which have been dropped in the items dataset) are replaced in the transactions data.

Thereafter, the variables in both datasets are analysed. The analysis reveals that more than half of the books are in English and above one-thirds are written in German. The distributions for the number of books per author, per publisher, and per main topic are very right skewed, with median values of 1, 1 and 4 respectively. Around 97% of the products belong to the broad categories of children's books, fiction or comic books. About fifty percent of the books lack information about their subtopics. The median number of clicks, baskets and orders per item in each session are 1, 0 and 0 respectively in the transactions dataset. Additionally, most sessions contain just one item ID. The top 5 most ordered books are in German.

After preprocessing and analysis, an ensemble of four LDA models is trained. This recommender suggests books with matching descriptions: a variable that has been successfully scraped for just above three-quarters of the products. These recommendations are compared with those generated by two naive alternatives: a random sampler of books written by the same author or publisher; and a selector of the most sold books pertaining to the same main topic. Eight item IDs are chosen and their recommendations are listed; three of them are also analysed in detail.

The quality of the recommendations produced by the LDA ensemble depends largely on the length and content of the description of each book. The longer the description and the more informative keywords it contains, the better the suggestions generated. The LDA recommender filters out books from other languages. Given a German book for instance, its top suggestions are also products in German: primarily because their descriptions are also in the same language and therefore contain many identical word tokens. The naive recommenders outperform the LDA ensemble when descriptions have too few word tokens for the latter to compare with those of other books. Even then, the recommendations are not horribly wrong.

To enhance the quality of the results, the proportion of available scraped descriptions has to be increased. Moreover, a certain minimum length for the descriptions can be specified before web scraping to ensure good performance of the recommender. For further improving the recommendations, a larger LDA ensemble comprising of more base learners could be trained. More values for the number of latent topics can be used and changing the values for the other two hyperparameters, i.e. α and η can also be experimented with. A recommender based on the LDA ensemble plus some other naive or sophisticated recommendation algorithms can be constructed, whereby the suggestions generated by each of these can be weighted and summed to produce final scores. Lastly, the size of the dataset can also be further increased to improve recommendations; especially by incorporating more non-fictional books for adults as the number of such items is relatively small.

References

[1] prudsys AG. *Data Mining Cup 2021*. URL: https://www.data-mining-cup.com/reviews/dmc-2021/ (visited on 09/04/2021).

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." In: *Journal of Machine Learning Research, 3* (2003), pp. 993–1022.

[4] *Buchhandlung Hugendubel: Online Medien- und Buchversand*. URL: https://www.hugendubel.de/de/ (visited on 09/29/2021).

[5] B. S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics, Fourth Edition*. Cambridge University Press, 2010.

[6] Python Software Foundation. *Python Language Reference, version 3.9.7*. URL: https://www.python.org/ (visited on 09/04/2021).

[7] T. L. Griffiths and M. Steyvers. "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.Supplement 1 (2004), pp. 5228–5235. DOI: 10.1073/pnas.0307752101.

[8] EDItEUR Limited. *Thema Subject Categories*. 2020. URL: https://ns.editeur.org/thema/en (visited on 09/29/2021).

[9] J. Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151. DOI: 10.1109/18.61115.

[10] Robi Polikar. "Ensemble Learning". In: *Ensemble Machine Learning: Methods and Applications*. Ed. by Cha Zhang and Yunqian Ma. Boston, MA: Springer US, 2012, pp. 1–34. ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7_1. URL: https://doi.org/10.1007/978-1-4419-9326-7_1.

[11] Radim Rehurek and Petr Sojka. "Gensim–python framework for vector space modelling". In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

[12] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to Recommender Systems Handbook*. Springer, 2011.

A. Appendix

Item ID	Title	Author	Publisher	Main topic	Subtopics
51121	The Jungle Book	Rudyard Kipling	Penguin Books Ltd (UK)	YFA	[5AH, 5AK, YFP]
71897	The Jungle Book	Rudyard Kipling	Classic Comic Store Ltd	XADC	[YF]
48475	The Jungle Book	Rudyard Kipling	OXFORD UNIV PR	DCA	[FB]
15576	The Jungle Book	Rudyard Kipling	Lulu.com	FM	[]
40074	The Jungle Book	Rudyard Kipling	Les prairies numériques	YNNB2	[YFP]
1408	The Jungle Book	Rudyard Kipling	Rupa Publications India	YFB	[1F, YNM]
59660	The Jungle Book	Rudyard Kipling	DOVER PUBN INC	YFA	[YFC, YFP]

Table 14: Multiple occurrences of the book title “The Jungle Book” in the items data set

Item ID	Title	Author	Publisher	Main topic	Subtopics
49722	The Odyssey	Gillian Cross, Homer	Walker Books Ltd	YFA	[5AJ]
29439	The Odyssey	Homer	Lulu.com	FM	[]
72406	The Odyssey	Diego Agrimbau	STONE ARCH BOOKS	YFA	[]

Table 15: Multiple occurrences of the book title “The Odyssey” in the items data set

Author	Count
Garcia Santiago	1370
Shelley Admont, Kidkiddos Books	206
James Manning	180
Idries Shah	168
Erin Hunter	158

Table 16: The five most frequently occurring authors in the preprocessed items data set

Publisher	Count
Books on Demand	3397
LIGHTNING SOURCE INC	2344
Lulu.com	1665
Xlibris	1533
Fichas de preescolar	1389

Table 17: The five most frequently occurring publishers in the preprocessed items data set

Item ID	Title	Orders
53695	Rico, Oskar und die Tieferschatten (Rico und Oskar 1)	170
69803	Die Welle	125
47120	Wir Kinder vom Bahnhof Zoo	122
31591	The Hate U Give	95
45799	Die Chroniken von Alice - Finsternis im Wunderland	93

Table 18: The most ordered items in the preprocessed transactions data set

Title	Main topic
The City of Guardian Stones	YFB
The Green Bicycle	YFB
Boy Overboard	YFB
Kendia's Abc's and Things I Can Be	YFB
Mr Gum und der Mürbekeksmilliardär	YFB

Table 19: Top 5 recommendations for item ID 12152 by the fallback strategy of the LDA-based recommender and the main topic of the books

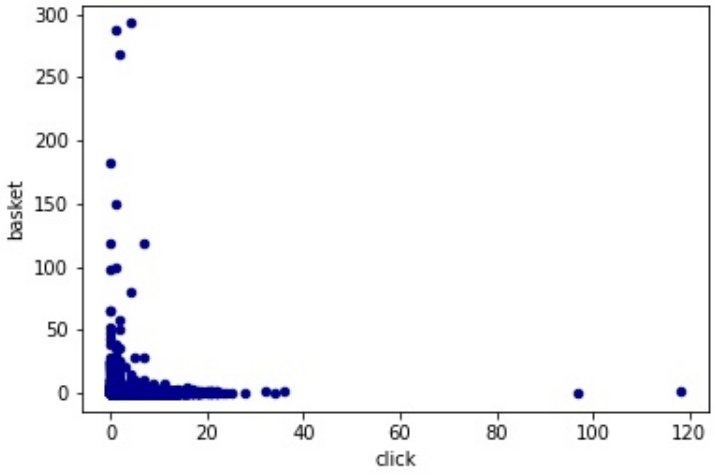

Figure 9: Scatter plot of the number of clicks and the corresponding number of baskets per item in each session in the preprocessed transactions data set

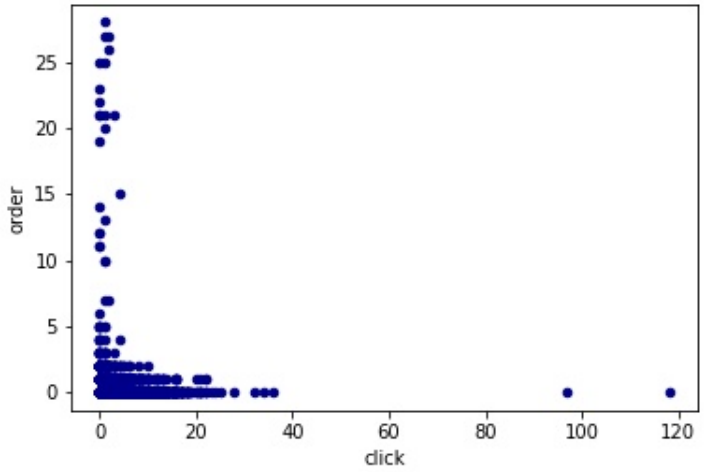

Figure 10: Scatter plot of the number of clicks and the corresponding number of orders per item in each session in the preprocessed transactions data set

Title	Score	Main topic
Freiheit ohne Schranken	0.610	FLC
Ich bin V wie Vincent	0.609	YFM
Schavna	0.607	FB
Geheimkurier A	0.603	YFCW
Weltentod	0.603	FMH

Table 20: Top 5 recommendations for item ID 13382 and 60644 by the LDA-based recommender, their similarity scores and main topics

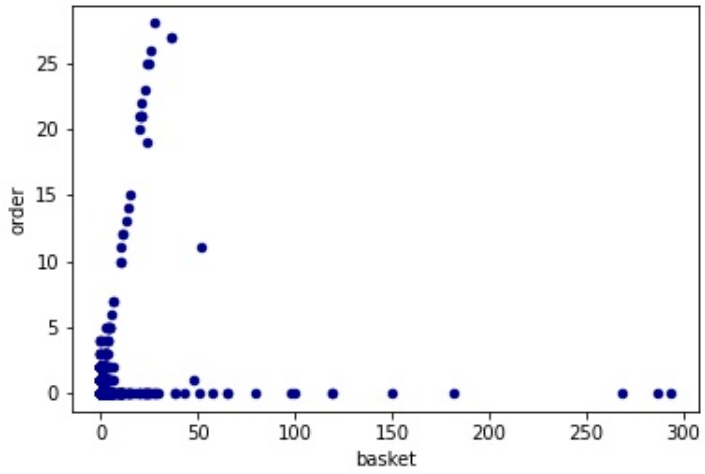

Figure 11: Scatter plot of the number of baskets and the corresponding number of orders per item in each session in the preprocessed transactions data set

Title	Main topic
The Gossip File	YFQ
Hiram Ulysses Aloysius Higgenbotham's Special Day	YFQ
Diary of a Tennis Prodigy	YFQ
In My Attic	YFQ
Super Organized	YFQ

Table 21: Top 5 recommendations for item ID 47675 by the fallback strategy of the LDA-based recommender and the main topic of the books

Title	Main topic
Blue	YBLD
Crayola Ramadan and Eid Al-Fitr Colors	YBLD
Colors On Our Papers/Rangi Za Makaratasi Yetu	YBLD
The Crayola (R) Patterns Book	YBLD
Das Farbenmonster	YBLD

Table 22: Top 5 recommendations for item ID 56180 by the fallback strategy of the LDA-based recommender and the main topic of the books

Title	Main topic
Me I Meant to Be	YFM
How to Make a Wish	YFB
Girl Named Digit	YFCF
Last Last-Day-Of-Summer	YBLJ
Mechanica	YFH

Table 23: Top 5 recommendations for item ID 12152 by the same author/publisher recommender and the main topics of the books

Title	Main topic
Land aus Feuer und Wasser	FL
Himmelskraft	FL
Das Erbe der Uraniden	FL
Befehl aus dem Dunkel	FL
Das stählerne Geheimnis	FL

Table 24: Top 5 recommendations for item IDs 13382 and 60644 by the same author/publisher recommender and the main topics of the books

Title	Main topic
Between the Walls	YFC
Leo, the Little Wanderer	YFB
Sanni ja taikakivi	YFN
Den helbredende kat	YNNJ22
Timo Taskuravun aarre	YNNS

Table 25: Top 5 recommendations for item ID 47675 by the same author/publisher recommender and the main topics of the books

Title	Main topic	Orders
Die Welle	YFB	125
The Hate U Give	YFB	95
Harry Potter 3 und der Gefangene von Askaban	YFB	46
Gangsta-Oma	YFB	44
Der Sprachabschneider	YFB	43

Table 26: Top 5 recommendations for item ID 12152 by the transactions-based recommender, the main topic and the number of orders of the books

Title	Main topic	Orders
1984	FL	14
Perry Rhodan 150. Stalker	FL	6
Eines Menschen Flügel	FL	5
Perry Rhodan 152. Die Raum-Zeit-Ingenieure	FL	4
Paradox - Am Abgrund der Ewigkeit	FL	4

Table 27: Top 5 recommendations for item IDs 13382 and 60644 by the transactions-based recommender, the main topic and the number of orders of the books

Title	Main topic	Orders
Gregs Tagebuch 05 - Geht's noch?	YFQ	17
Gregs Tagebuch 8 - Echt übel!	YFQ	17
Gregs Tagebuch 11 - Alles Käse!	YFQ	17
Gregs Tagebuch 06 - Keine Panik!	YFQ	16
Gregs Tagebuch 07 - Dumm gelaufen!	YFQ	15

Table 28: Top 5 recommendations for item ID 47675 by the transactions-based recommender, the main topic and the number of orders of the books

Title	Main topic	Orders
Das Farbenmonster	YBLD	36
Fingerstempel-Spaß Kunterbunt	YBLD	24
Finger-Malspaß: Unter Wasser	YBLD	2
MILLAS KRITZEL MALBUCH - Mach es Fertig!	YBLD	2
Peppa Pig Stickerspaß	YBLD	2

Table 29: Top 5 recommendations for item ID 56180 by the transactions-based recommender, the main topic and the number of orders of the books