Final project

Telma Peura

Computational literacy

Helsinki university

Autumn 2021

Listing American reading habits

An analysis of The New York Times Best Sellers lists

1. Introduction

The New York Times Bestseller List is the most popular and internationally known list of bestseller books, published since 1931. According to their own description, the lists are "Authoritatively ranked lists of books sold in the United States, sorted by format and genre."[1] Today, the list is divided into several sub-lists focusing on different genres and categories (e.g., hardcover fiction, audio fiction, combined print and e-book nonfiction). The list is compiled of sales numbers from retailers nationwide, although the exact method is kept a secret. As the list reflects reading habits and works for many as the primary source of recommendations to look for new readings (Yucesoy et al., 2018), it is interesting to explore the list evolution as well as to analyze bestseller list characteristics to understand how the U.S. reads.

According to Berglund & Dahllöf (2021), the ways through which literature is distributed and sold affects the readers' experience of the text and how they interact with it. Moreover, the change to online audio subscriptions, such as Amazon Audible and Storytel, is suggested to change how some kinds of literature is appreciated and written. As a matter of fact, their stylistic analysis of the Swedish book market revealed that the stylistic features of books consumed as audiobooks indeed differ from read literature (Berglund & Dahllöf,

[1] https://www.nytimes.com/books/best-sellers/

2021). Similarly, Barakat (2018) found that popular audiobooks are often published as a series, and they write about topics such as human relationships.

While the study of bestseller lists does not allow me to analyze the stylistic features of the books themselves, the plot summaries provided in the lists might still reveal main topics of the books. The lists might also serve as a proxy for exploring diversity in terms of authors and books that are read or listened to in different formats.

2. The aim of this project

The change in reading habits is also reflected in the NYT lists. For instance, the audiobook list was published for the first time in March 2018. Although the analysis of lists does not allow going deep in the text itself, the nature of these lists can reveal interesting differences between characteristics of different lists, which I aim to explore in this project. My research questions are as follows:

- Are there any identifiable clusters of authors or titles that remain on the list longer than other books? (RQ1)
- What can plot summaries of each title on the list reveal about the books and topics appearing on the list? (RQ2)
- Can topic modeling help characterize differences in books that are read as text or listened to as audio books? (RQ3)

For RQ1, my hypothesis is that overall, there is less variation in the non-fiction lists, and on average, titles remain there for a longer time. To answer RQ2 and RQ3, I will use topic modeling to explore the book descriptions provided in the lists.

3. Data retrieval and preprocessing

I used the New York Times Books API[2] to retrieve the NYT Best Seller lists from a 5-year time period. Each retrieved list contains 15 bestsellers with metadata, such as title, rank, the number of weeks on the list, the isbn identification number, the author, and a short description,. The data was retrieved in a json format that I then converted to python dictionaries and finally to csv files[3].

[2] https://developer.nytimes.com/docs/books-product/1/overview

[3] All code is available on [my-github]

An example plot summary from the list "audio fiction" looks like this:

> *The 12th book in the Mercy Thompson series. The car mechanic who has the ability to turn into a coyote takes on a deadly foe. Read by Lorelei King. 10 hours, 9 minutes unabridged.*
>
> *(Smoke bitten, Patricia Briggs)*

The last sentences, about the audiobook reader and the length of the audiobook were not relevant for my topic models, so I decided to cut all descriptions after "Read by", and I added "hour", "minute" and "unabridged" as custom stopwords. I lemmatized all the descriptions. The mean length of a description obtained this way was 101.7 characters.

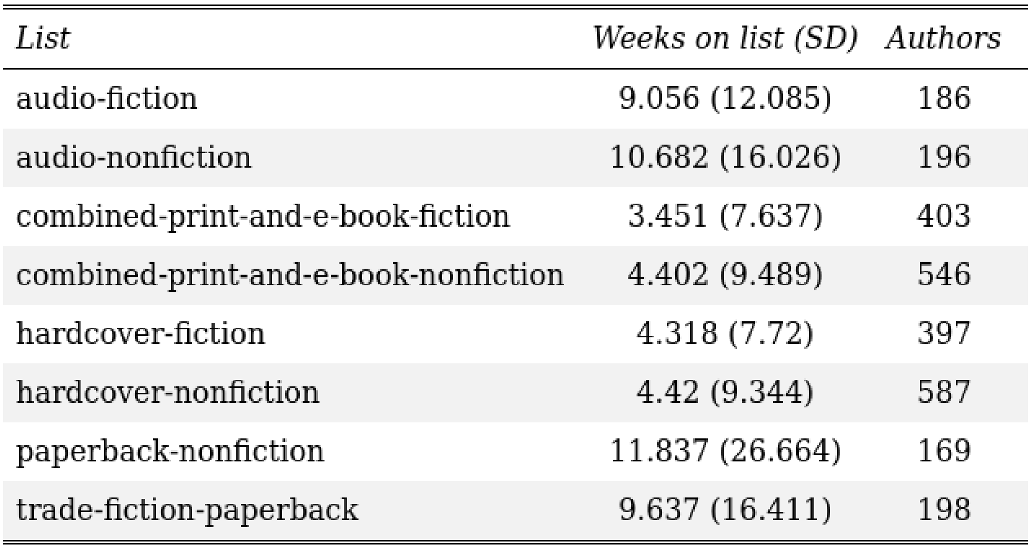
4. Methods and Analysis

The analysis was conducted in RStudio (RStudio team, 2021). First, descriptive statistics were computed to get an overview of the analyzed dataset. I used structural topic modeling (STM) to explore if there are any identifiable topics in the bestseller descriptions. STM is a generative topic modeling approach where a topic is defined as a mixture of words with corresponding probability distributions. Similarly, generated document-topic probability distributions associate each document in the dataset with the computed topics (Roberts et al., 2019). This way, I investigated if there are differences in the kind of topics readers prefer in audio/written format. To answer RQ3, I carried out an anova multiple mean comparison to see if there was a difference between topic distribution between the lists. Finally, I qualitatively examined my findings.

5. Results

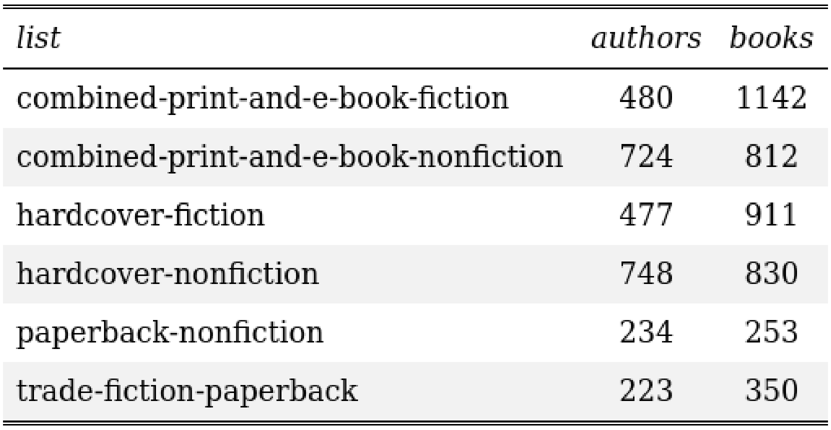
5.1. Statistical findings

As Table 1 shows, there are differences in how long a title remains on the list, as well as how many authors are represented. Across all lists, comparing fiction and non-fiction literature, the number of authors is bigger for non-fiction (Table 1). At the first sight, it seems that a smaller variety of authors is listened to as audio books, but this number might be biased by the fact that audio book lists are only updated monthly, whereas the other lists are updated twice a month. Appendix 1 shows the list of top 5 authors in each list, but analyzing them more in detail was out of the scope of this project.



List	*Weeks on list (SD)*	*Authors*
audio-fiction	9.056 (12.085)	186
audio-nonfiction	10.682 (16.026)	196
combined-print-and-e-book-fiction	3.451 (7.637)	403
combined-print-and-e-book-nonfiction	4.402 (9.489)	546
hardcover-fiction	4.318 (7.72)	397
hardcover-nonfiction	4.42 (9.344)	587
paperback-nonfiction	11.837 (26.664)	169
trade-fiction-paperback	9.637 (16.411)	198

Table 1. The table shows the average number of weeks a title remains on the list, with the corresponding standard deviation, and the total number of different authors appearing on the lists.



list	*authors*	*books*
combined-print-and-e-book-fiction	480	1142
combined-print-and-e-book-nonfiction	724	812
hardcover-fiction	477	911
hardcover-nonfiction	748	830
paperback-nonfiction	234	253
trade-fiction-paperback	223	350

Table 2. The total number of authors and books in the text format lists.

Since the audiobook lists are only updated monthly, I decided to focus on only written books for the comparison of fiction/nonfiction reading trends (RQ1). I expected there to be more variance in the fiction lists, which was not fully confirmed. In terms of how long one title remains on the list, a two-sample t-test showed that there was a statistically significant difference in mean between fiction (M=4.80, SD=10.27) and nonfiction (M=6.13 , SD=15.09); t(3209.8) = -3.29, *p*< .001., indicating that a nonfiction book remains for more weeks on the list. This supports my hypothesis. However, when I compared the number of weeks weeks one author appears on a list, the statistical analysis showed that one author stays on the lists longer in fiction, the number of weeks being M=9.78 SD=19.73, and on the nonfiction lists M=6.76, SD 16.46; t(2226.4)=4.31, *p*<.001. Thus, in terms of authors, there is more variance in nonfiction, contradicting my hypothesis. Similar observations can be done just looking at Table 2: the total number of books is higher in the fiction lists, whereas there is more diversity in authors in nonfiction.

I also did a visual inspection of the number of new titles over time. I filtered only the first date a book appeared on the list, counted the number of titles per date, and plotted the number of books grouped per month (Figure 1). In fiction literature, there is a slight decreasing trend in the number of titles, whereas in non-fiction, it is harder to draw any conclusions based on the visual inspection. This also suggests that variability in reading is decreasing in fiction literature, whereas it has remained more or less stable on the nonfiction side.

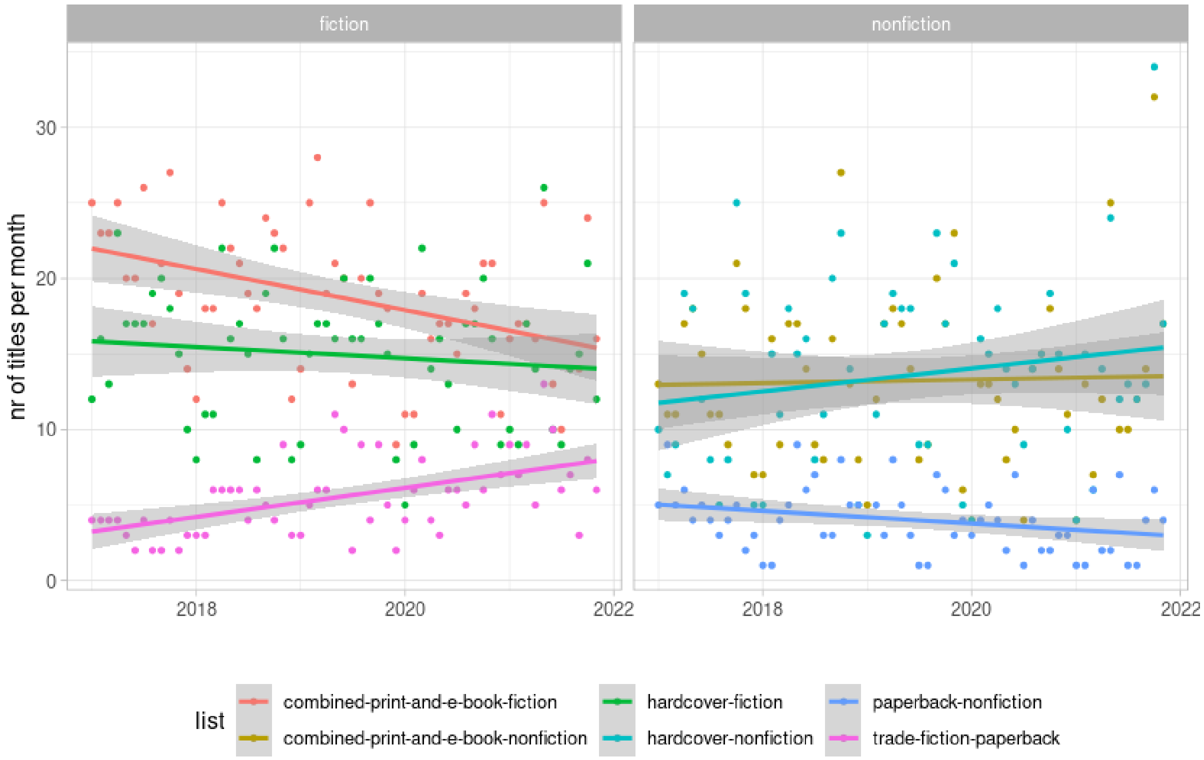


Figure 1. The number of new titles appearing on the lists over time.

5.2. Topic modeling fiction

For RQ2 and 3, I decided to focus on fiction literature. The number of topics was selected after a K search with k values 3, 5, 7, 10, 15 and 20. As Figure 2 shows, the held-out-likelihood was the best at 7 and the semantic coherence also seemed to decrease rapidly after that, so I decided to compute seven topics.

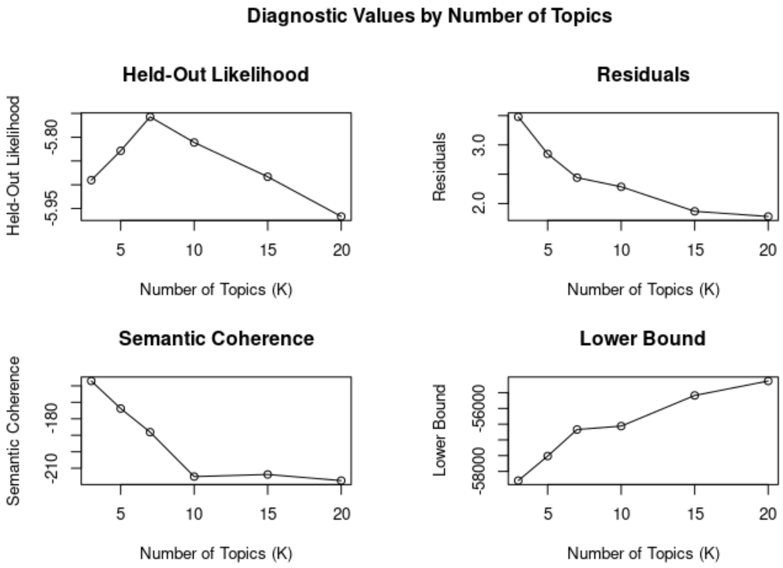


Figure 2. Diagnostic values for different numbers of topics.

Exploring the top words per every topic points at the same direction, from the first sight, it is hard to tell how the topics differ. The general direction seems to be that thriller/action novels tend to appear on the NYT bestseller lists.

Topic 1 Top Words:

Highest Prob: family, war, world, return, stone, jack, american

FREX: family, war, jack, american, four, change, state

Lift: british, crisis, double, family, half, jewish, mansion

Score: family, war, play, world, stone, jack, barrington

Topic 2 Top Words:

Highest Prob: young, secret, daughter, try, relationship, girl, save

FREX: daughter, girl, wed, russian, live, make, form

Lift: allon, assistant, entangle, intelligence, keep, live, race

Score: arctic, secret, daughter, young, girl, russian, try

Topic 3 Top Words:

Highest Prob: series, book, first, seek, face, trilogy, will

FREX: series, book, trilogy, final, hunter, danger, vampire

Lift: argeneau, castle, empire, red, book, series, agency

Score: book, series, invade, trilogy, hunter, saga, steel

Topic 4 Top Words:

Highest Prob: find, must, former, agent, sister, force, town

FREX: must, fall, black, small, stop, give, court

Lift: fall, adoption, area, arrive, assume, biologist, bookseller

Score: expeditionary, must, find, force, agent, town, team

Topic 5 Top Words:

Highest Prob: woman, new, life, man, three, become, take

FREX: life, become, love, york, city, mystery, new

Lift: affair, become, complication, director, expert, hospital, inherit

Score: life, new, woman, shifter, york, three, become

Topic 6 Top Words:

Highest Prob: murder, year, investigate, old, detective, may, miss

FREX: murder, investigate, old, death, case, kill, boy

Lift: affect, bosch, cause, lieutenant, lucas, maisie, nora

Score: investigate, murder, year, detective, old, death, addict

Topic 7 Top Words:

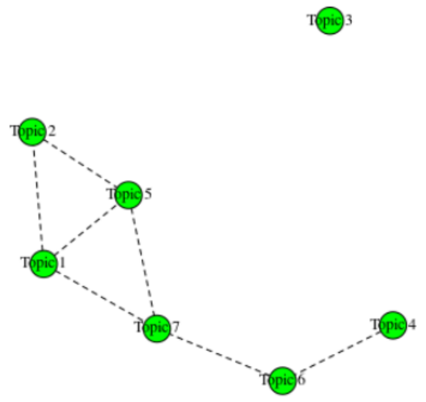
Highest Prob: two, one, story, novel, past, help, discover

FREX: novel, friend, childhood, private, struggle, one, carolina

Lift: carolina, clear, coast, francisco, malone, widow, friend

Score: boss, two, story, one, friend, novel, crime

The first visual inspection (Figure 3) of the topic model implies that Topic 3 is more different from other topics. The connectedness of many topics supports the initial observations based on the above key words.



Next, I grouped the titles per all the lists they appeared in. For example, if one title was found on the lists “audio-fiction” and “hardcover-fiction”, it was labeled as “audio-hardcover” to indicate both list types. The topic distributions are shown in Figure 4.

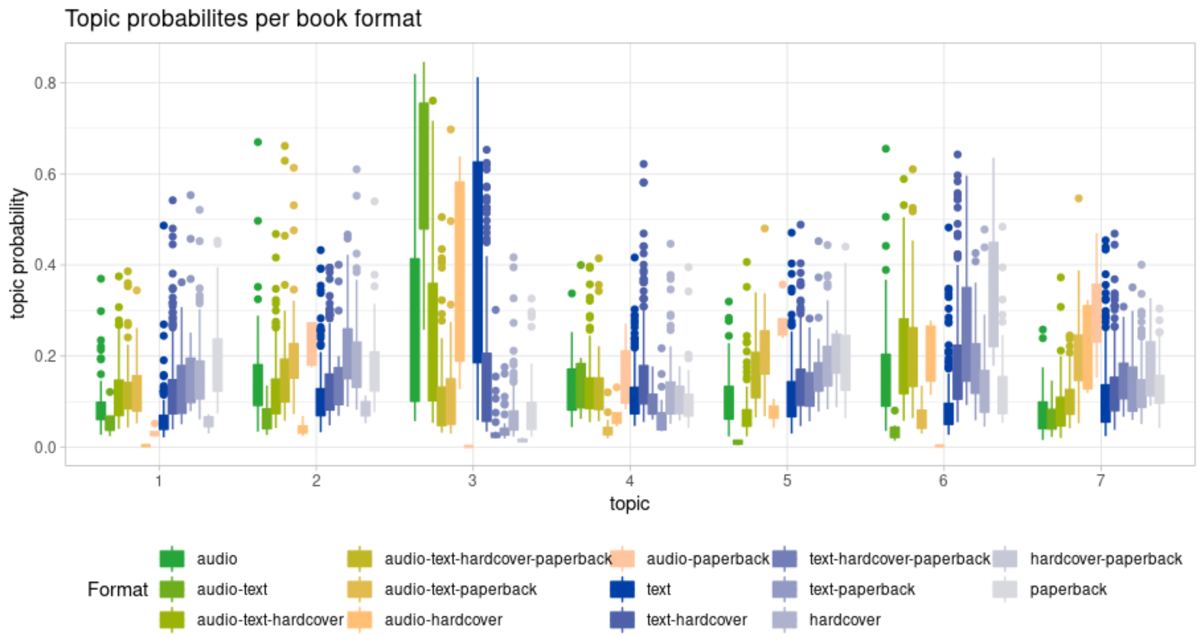


Figure 4. Topic probabilities per reading format. The plot represents the real distributions of topic probabilities. Titles that are found in audio format are colored in a green scale, whereas titles only found in written type lists are colored in blue. For clarity, the original list name “combined-printed-and-e-book” is labeled as “text”.

Based on the visual inspection, there does not seem to be an apparent difference in topic distribution that would be connected with a reading format. Books read as text or listened to as audio books seem to have rather similar topic probability distributions. However, the result for the multiple mean comparison in ANOVA (comparing formats audio only VS both audio and text VS text only) was statistically significant for the interaction between topic and format, F(12)=9.505, *p* < .001. To visualize the model, average predictions per topic are plotted in Figure 3.

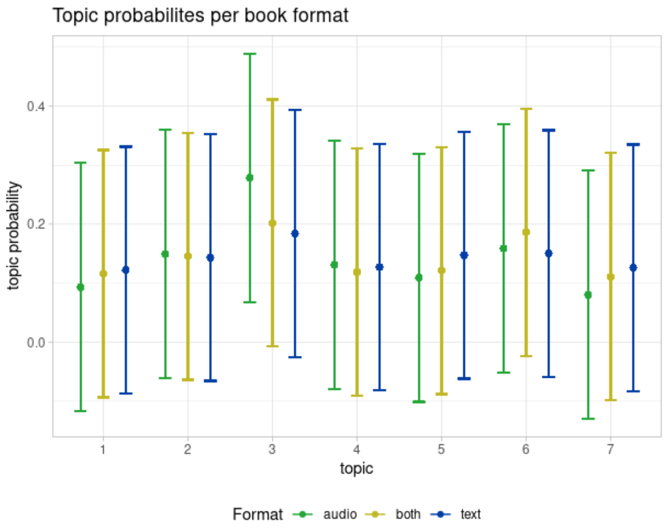
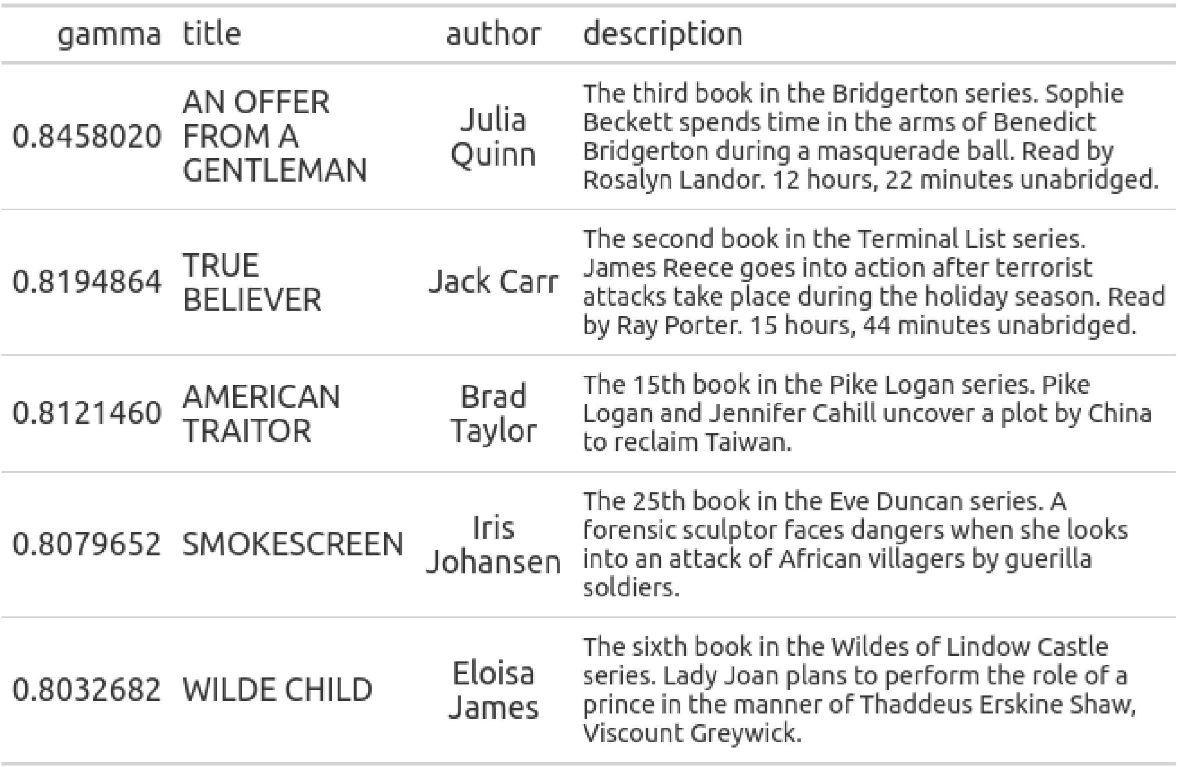


Figure 5. Predicted topic probabilities per reading format, with .95% confidence intervals.

Since topic 3 seemed to be associated with higher probabilities to appear across all formats, I decided to investigate it more closely. I retrieved the titles with the highest associated probabilites to this topic, shown in Table 4. (For the top 5 titles of each topic, see Appendix 2.)



gamma	title	author	description
0.8458020	AN OFFER FROM A GENTLEMAN	Julia Quinn	The third book in the Bridgerton series. Sophie Beckett spends time in the arms of Benedict Bridgerton during a masquerade ball. Read by Rosalyn Landor. 12 hours, 22 minutes unabridged.
0.8194864	TRUE BELIEVER	Jack Carr	The second book in the Terminal List series. James Reece goes into action after terrorist attacks take place during the holiday season. Read by Ray Porter. 15 hours, 44 minutes unabridged.
0.8121460	AMERICAN TRAITOR	Brad Taylor	The 15th book in the Pike Logan series. Pike Logan and Jennifer Cahill uncover a plot by China to reclaim Taiwan.
0.8079652	SMOKESCREEN	Iris Johansen	The 25th book in the Eve Duncan series. A forensic sculptor faces dangers when she looks into an attack of African villagers by guerilla soldiers.
0.8032682	WILDE CHILD	Eloisa James	The sixth book in the Wildes of Lindow Castle series. Lady Joan plans to perform the role of a prince in the manner of Thaddeus Erskine Shaw, Viscount Greywick.

Table 4. The titles with the highest probabilities for topic 3.

This indicates that book series often appear on the lists. Thus, rather than the story contents, topic 3illustrates how books are presented in a short description. Whereas the other titles seem to be thrillers or crime novels, the first one is a romance.

At this point, the results indicate quite clearly, that topic modeling based on short descriptions is not a very fruitful approach to explore the bestseller lists. It is hard to see any real differences between the distribution of topics related to audio or text format.

6. Discussion

The aim of this project was to explore reading habits through NYT Best Seller lists, and particularly compare the types and topics of books read in written or audio format. However, it turned out that the lists were a challenging dataset for this purpose, and not completely appropriate for answering my research questions. First of all, there exist several lists for different book formats and genres, which makes it difficult to gather the dataset. For example, I did not include the list "Advice, How-To & Miscellaneous" to my nonfiction dataset, and I also excluded the monthly lists of children's and young adult literature from my fiction dataset. This might have changed particularly the results of topic modeling. Moreover, as I mentioned in the introduction, it is not totally known how the lists are collected, and what kind of book market data is considered. According to the NYT methodology, "[s]ales are statistically weighted to represent and accurately reflect all outlets proportionally nationwide" ('About the Best Sellers - The New York Times', n.d.). Thus, editorial choices, such as what is considered young adult literature (and thus not included in the other lists) and how to proportionally balance the sales have supposedly a big impact on the obtained results.

In addition, as the listings are based on book sales reported by bookshops, it does not reveal anything about library loans or other ways of getting literature. The issue is even more prevalent with audiobooks, as "[f]ree-trial or low-cost audiobook sales are not eligible for inclusion" ('About the Best Sellers - The New York Times', n.d.). Since most audiobooks are listened to through audiobook subscription platforms that often have free trial offers, it is hard to tell how representative the current audiobook lists are.

The other big constraint of this analysis is methodological. It is generally found that topic modeling does not work well with short data, because of noisy and sparse data, insufficient word co-occurrence information, and non-meaningful or irrelevant words for retrieving meaningful topics (Albalawi et al., 2020). My topic model was computed from descriptions that were around 100 characters long, thus also resulting in a relatively small dataset to build a topic model. The topics might have been more informative if whole book reviews had been available, for example. More importantly, the descriptions answer rather to the question what is highlighted in a short book description, and not to my original question what

As to my results, they might reflect more the availability of different formats, rather than the readers' preferences. Audiobooks are a quite new increasing trend, so the number of authors and books in this format might actually reflect the availability of literature in audio. Naturally, there are still a lot more books published in written format only, which does not make my comparison fair.

7. Conclusion and further thoughts

The aim of this exam project was to explore NYT Best Seller lists and try to see what they can tell about American reading habits. The process made me realize how little a simple list can actually tell about reading habits. Rather, it tells about how reading is framed in the media. I still think the lists are a big factor influencing readers' reading choices - even in Finland, many books are advertised with the claim "New York Times Best Seller" appearing on the cover, but it was impossible to extrapolate this kind of information from the present analysis.

However, the project made me think of other applications of the lists. It could be fun to train a title or description generator model based on the best seller titles and descriptions. Another interesting analysis could be to combine the lists with the nationality of each author and to visualize this data on a map to see what kind of world is covered in literature, particularly in terms of what regions of the world are ignored. Thus, my conclusion is that reading habits and trends are hard to capture based on best seller lists only, but further combining the insights from this project with other data could lead to interesting applications.

References

About the Best Sellers—The New York Times. (n.d.). *The New York Times*. Retrieved 21 December 2021, from https://www.nytimes.com/books/best-sellers/methodology/https://www.nytimes.com/books/best-sellers/methodology/

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text

Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*, 42. https://doi.org/10.3389/frai.2020.00042

Barakat, A. (2018). *What makes an (audio)book popular?* https://www.diva-portal.org/smash/get/diva2:1265673/FULLTEXT01.pdf

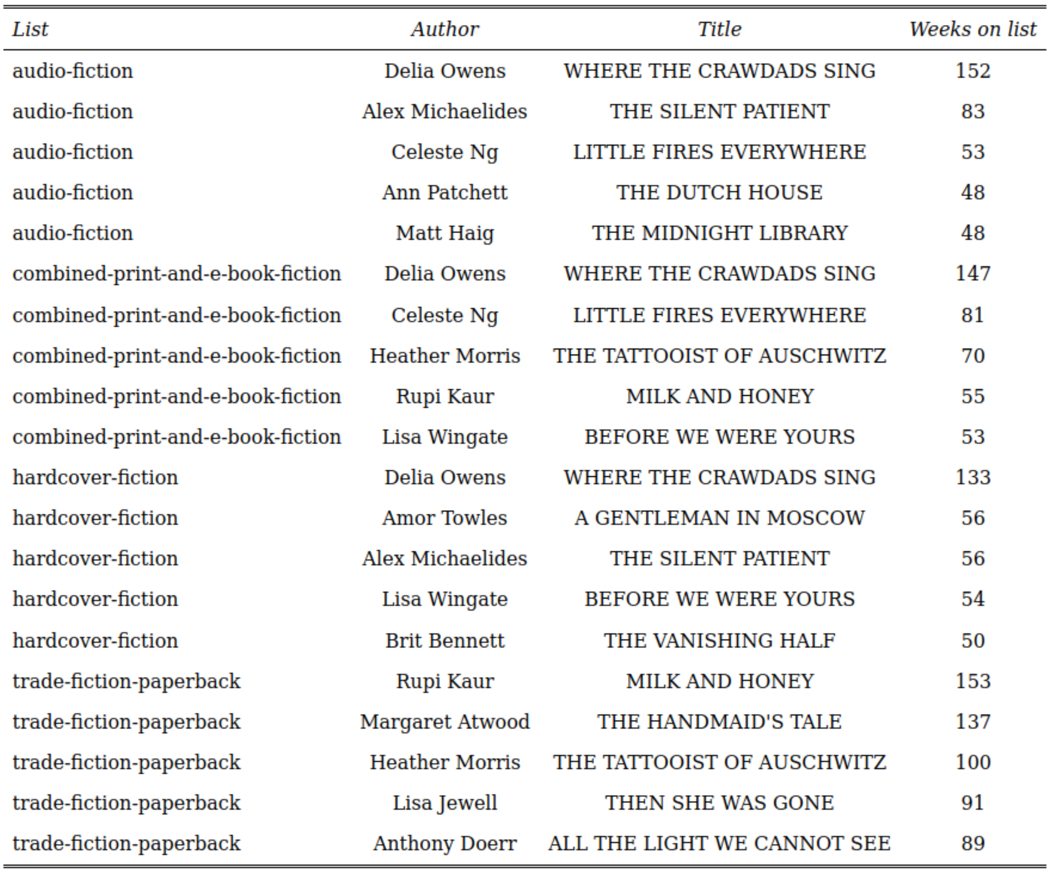
Berglund, K., & Dahllöf, M. (2021). Audiobook Stylistics: Comparing print and audio in the bestselling segment. *Journal of Cultural Analytics*, 29802. https://doi.org/10.22148/001c.29802

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(1), 1-40.

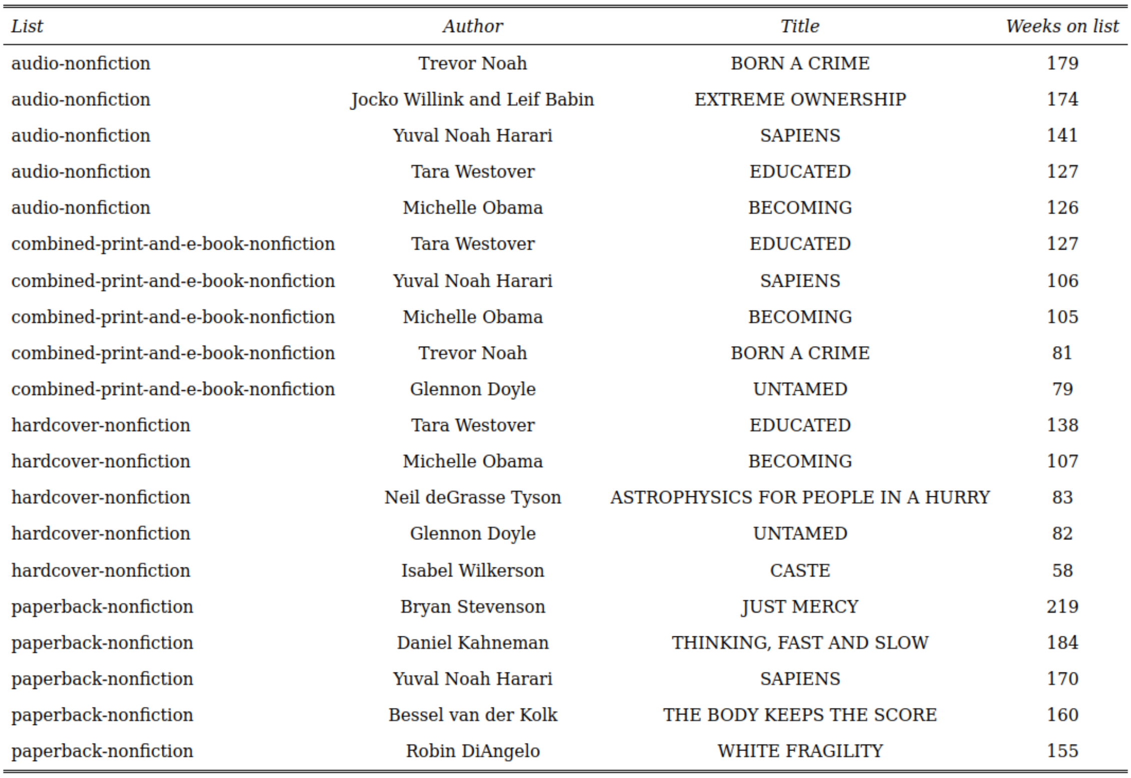
RStudio Team. (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Yucesoy, B., Wang, X., Huang, J., & Barabási, A.-L. (2018). Success in books: A big data approach to bestsellers. *EPJ Data Science*, *7*(1), 1–25. https://doi.org/10.1140/epjds/s13688-018-0135-y

Appendix 1. Top 5 titles per list, measured in number of weeks appearing on the list.



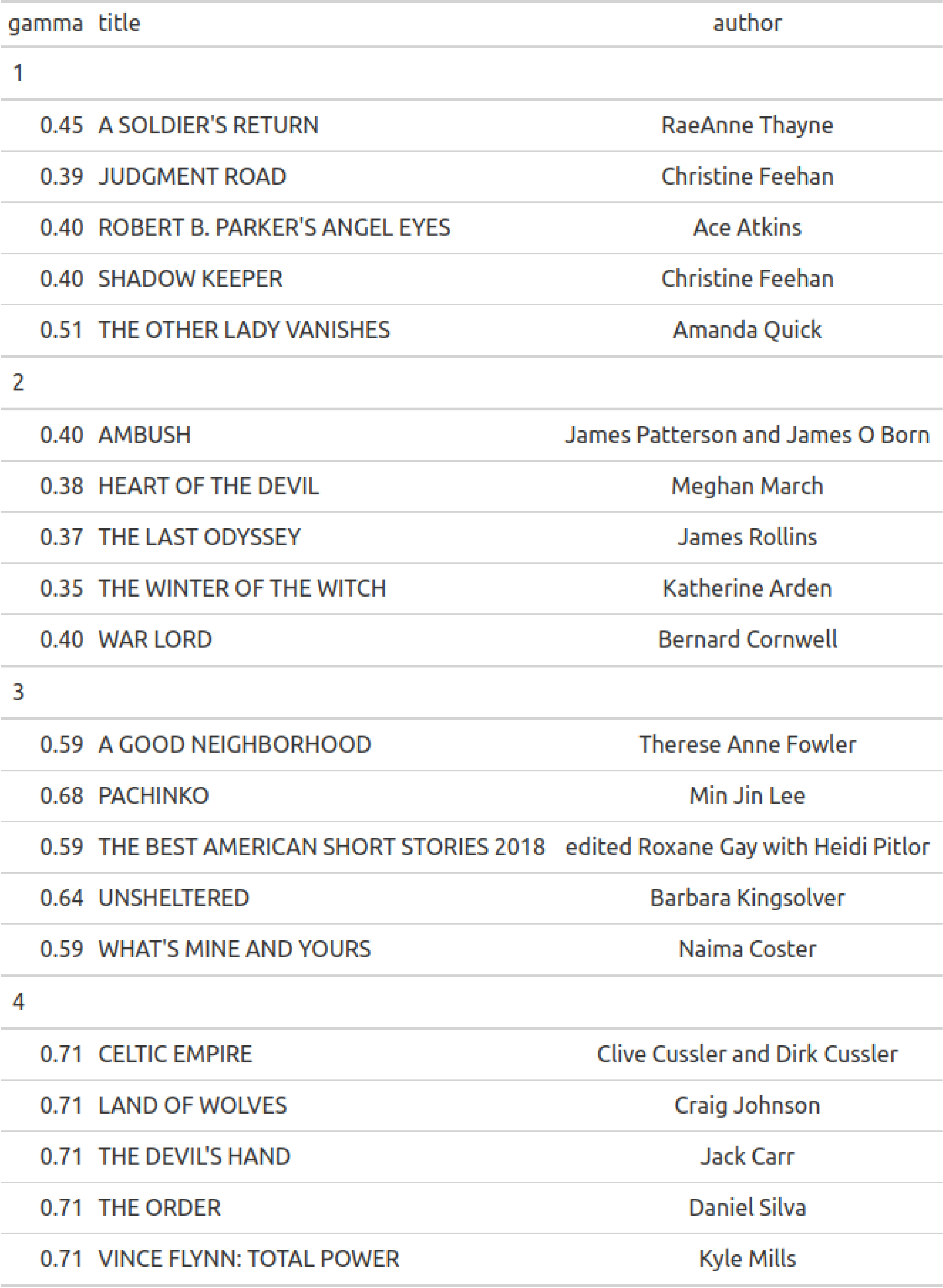
List	*Author*	*Title*	*Weeks on list*
audio-fiction	Delia Owens	WHERE THE CRAWDADS SING	152
audio-fiction	Alex Michaelides	THE SILENT PATIENT	83
audio-fiction	Celeste Ng	LITTLE FIRES EVERYWHERE	53
audio-fiction	Ann Patchett	THE DUTCH HOUSE	48
audio-fiction	Matt Haig	THE MIDNIGHT LIBRARY	48
combined-print-and-e-book-fiction	Delia Owens	WHERE THE CRAWDADS SING	147
combined-print-and-e-book-fiction	Celeste Ng	LITTLE FIRES EVERYWHERE	81
combined-print-and-e-book-fiction	Heather Morris	THE TATTOOIST OF AUSCHWITZ	70
combined-print-and-e-book-fiction	Rupi Kaur	MILK AND HONEY	55
combined-print-and-e-book-fiction	Lisa Wingate	BEFORE WE WERE YOURS	53
hardcover-fiction	Delia Owens	WHERE THE CRAWDADS SING	133
hardcover-fiction	Amor Towles	A GENTLEMAN IN MOSCOW	56
hardcover-fiction	Alex Michaelides	THE SILENT PATIENT	56
hardcover-fiction	Lisa Wingate	BEFORE WE WERE YOURS	54
hardcover-fiction	Brit Bennett	THE VANISHING HALF	50
trade-fiction-paperback	Rupi Kaur	MILK AND HONEY	153
trade-fiction-paperback	Margaret Atwood	THE HANDMAID'S TALE	137
trade-fiction-paperback	Heather Morris	THE TATTOOIST OF AUSCHWITZ	100
trade-fiction-paperback	Lisa Jewell	THEN SHE WAS GONE	91
trade-fiction-paperback	Anthony Doerr	ALL THE LIGHT WE CANNOT SEE	89



List	*Author*	*Title*	*Weeks on list*
audio-nonfiction	Trevor Noah	BORN A CRIME	179
audio-nonfiction	Jocko Willink and Leif Babin	EXTREME OWNERSHIP	174
audio-nonfiction	Yuval Noah Harari	SAPIENS	141
audio-nonfiction	Tara Westover	EDUCATED	127
audio-nonfiction	Michelle Obama	BECOMING	126
combined-print-and-e-book-nonfiction	Tara Westover	EDUCATED	127
combined-print-and-e-book-nonfiction	Yuval Noah Harari	SAPIENS	106
combined-print-and-e-book-nonfiction	Michelle Obama	BECOMING	105
combined-print-and-e-book-nonfiction	Trevor Noah	BORN A CRIME	81
combined-print-and-e-book-nonfiction	Glennon Doyle	UNTAMED	79
hardcover-nonfiction	Tara Westover	EDUCATED	138
hardcover-nonfiction	Michelle Obama	BECOMING	107
hardcover-nonfiction	Neil deGrasse Tyson	ASTROPHYSICS FOR PEOPLE IN A HURRY	83
hardcover-nonfiction	Glennon Doyle	UNTAMED	82
hardcover-nonfiction	Isabel Wilkerson	CASTE	58
paperback-nonfiction	Bryan Stevenson	JUST MERCY	219
paperback-nonfiction	Daniel Kahneman	THINKING, FAST AND SLOW	184
paperback-nonfiction	Yuval Noah Harari	SAPIENS	170
paperback-nonfiction	Bessel van der Kolk	THE BODY KEEPS THE SCORE	160
paperback-nonfiction	Robin DiAngelo	WHITE FRAGILITY	155

Appendix 2. Top 5 titles per topic.

These titles and topics were not analyzed, but I find it interesting that none of the top titles or authors appear on the top lists of Appendix 1.



gamma	title	author
1		
0.45	A SOLDIER'S RETURN	RaeAnne Thayne
0.39	JUDGMENT ROAD	Christine Feehan
0.40	ROBERT B. PARKER'S ANGEL EYES	Ace Atkins
0.40	SHADOW KEEPER	Christine Feehan
0.51	THE OTHER LADY VANISHES	Amanda Quick
2		
0.40	AMBUSH	James Patterson and James O Born
0.38	HEART OF THE DEVIL	Meghan March
0.37	THE LAST ODYSSEY	James Rollins
0.35	THE WINTER OF THE WITCH	Katherine Arden
0.40	WAR LORD	Bernard Cornwell
3		
0.59	A GOOD NEIGHBORHOOD	Therese Anne Fowler
0.68	PACHINKO	Min Jin Lee
0.59	THE BEST AMERICAN SHORT STORIES 2018	edited Roxane Gay with Heidi Pitlor
0.64	UNSHELTERED	Barbara Kingsolver
0.59	WHAT'S MINE AND YOURS	Naima Coster
4		
0.71	CELTIC EMPIRE	Clive Cussler and Dirk Cussler
0.71	LAND OF WOLVES	Craig Johnson
0.71	THE DEVIL'S HAND	Jack Carr
0.71	THE ORDER	Daniel Silva
0.71	VINCE FLYNN: TOTAL POWER	Kyle Mills

5			
	0.66	AFTER	Anna Todd
	0.54	PAYBACK'S A WITCH	Lana Harper
	0.55	ROYAL	Danielle Steel
	0.57	THE SUN DOWN MOTEL	Simone St James
	0.54	TOM CLANCY: CODE OF HONOR	Marc Cameron
6			
	0.53	MEANT TO BE YOURS	Susan Mallery
	0.50	PLAYING FOR KEEPS	Jill Shalvis
	0.59	STEALTH	Stuart Woods
	0.55	THE KAISER'S WEB	Steve Berry
	0.53	WILDE CHILD	Eloisa James
7			
	0.59	LADY IN THE LAKE	Laura Lippman
	0.67	NEVER TELL	Lisa Gardner
	0.66	THE BOY	Tami Hoag
	0.71	WE BEGIN AT THE END	Chris Whitaker
	0.66	YOU DON'T OWN ME	Mary Higgins Clark and Alafair Burke