Books versus eBooks : The customer's choice

Pierre-Alexandre Lee
pierre-alexandre.lee@epfl.ch

Marc Jollès
marc.jolles@epfl.ch

Yves Lamonato
yves.lamonato@epfl.ch

Abstract

This document is the report for the final project in ADA 2017. We choose to compare the reviews of books and ebooks with ratings obtained from Amazon between 1996 and 2014. Our main goal is to find out if for the same author, there is an user preference between the paper and electronic version of the book. We will explain how we scraped more info on each books and ebooks from Amazon, and what kind of analysis we performed on this data.

1 Introduction

There has been a lot of discussion about the difference between books and ebooks, some arguing that the former is better than the latter (Kumar, 2015), as it may be easier to focus on it, and the feeling when reading a physical copy is better. However, there are some additions like interactivity, hyperlinks and functionalities, inherent to the ebook format, that may have a positive impact on readers. Thus, it's legitimate to ask ourselves whether people, tend to prefer the virtual or physical format, if we base our reasoning on real data, and not on subjective reasons. For this study, we will use the Amazon database, in which we do have reviews and grades for different books and ebooks.

2 First look into Data

For this study, multiple datasets were used, coming from the website of Amazon product data (McAuley, 2015). The 5-core files of books and kindle store were obtained from there, and the metadata file was retrieved from the cluster of the course, as it was not available directly from the website. As the first approach for the study was to compare book and ebook reviews for the same content, the metadata was mandatory to make a link between a book asin (Amazon Standard Identification Number) and an ebook asin. The way chosen to do so was to compare title, as it is the only matching information that could exist in the metadata.

As the files are really big, a reduction to keep only necessary information has been done : for the 5-core files, files containing only the unique ASIN for available books and ebooks were created. For the metadata file, as it contains information for a lot of amazon articles that are not books or ebooks, a file containing only needed data has been created.

After that, the team realized that for most of the ebooks (kindle store), the metadata file does not contain the title data. After having searched if there was some other discriminatory criterion and concluded that there were not, the team has created a bot to retrieve this data from Amazon directly. (Datahut, 2016)

3 Data retrieving

3.1 Requests through Amazon website

A first attempt has been done by forging a link for every ebook asin available, as for every Amazon article with asin $*id*$, the corresponding web page is https://www.amazon.com/dp/$*id*$/ref=rdr_kindle_ext_tmb. Using the requests and BeautifulSoup libraries, it was easy to get and scrape the pages.

However, after some execution time, the bot was scraping less entries, and when trying to figure out why it was doing so, the team noticed that Amazon has a bot detection system, and prevents people to obtain data this way, saying that the bot has to go through their API. The bot was then modified to obtain information despite the detection system, by using user-agent rotation and random delay between requests (ScrapeHero, 2014), which was

still detected by Amazon after some time.
The team could have tried the Amazon API, but after little research it was found that billing information had to be given when registering, so an alternative was preferred and has been tried.

3.2 Requests through ASINDB

Instead of going through Amazon, a possible alternative was to go through http://www.asindb.com/ to obtain the title information. After having obtained some information about the title (with some data missing, as it's not Amazon, it's a website that has scrapped some data from Amazon), the data obtained and the book metadatas were matched on the title. As the data retrieval was pretty long, after having obtained 2741 entries from the 61934 unique ebook entries, 1506 entries were matched. However, those 1506 entries were not always on the same content : there are multiple books with the same title on the entries we obtained. If we drop all elements that are present multiple times, we have 148 elements. And in those 148 elements, there was no guarantee that the content is the same (in fact, after some manual analysis, most of those are not on the same content, but it is not possible in this this context to automatically check if the content is the same). To be almost sure that the book and ebook are on the same entry, there is a possibility : check also the author. But, it's not an information available from ASINDB, so the group chose to finally try Amazon AWS and Amazon Associates API.

3.3 Requests through Amazon API

After having obtained the accounts to use the API, the author information has been retrieved in addition to the title information for a given ASIN. When matching ebook entries with book entries for some couple (title, author), very few entries were obtained : 94 entries. As it's clearly not sufficient, we had to modify the project subject to work with book and ebook matching given the same author instead of given the same content.
As in the research questions we have also discussed about the price and other attributes that can be compared between the book and ebook version, some additional information has been scrapped from Amazon : the price, the salesRank, the number of pages, the release date and the language.

4 About the reviews

As said before, we work with the 5-core files for the books and for the kindle store (where we find the ebooks). The 5-core version is a reduced version of the whole ratings data, where each user has given at least 5 reviews, and each item has at least 5 reviews given to it. We choose to take this to have a denser dataset to work with. More reviews mean that we have more information on an user or product. For example, an item with only one bad review could have been the victim of a malicious or uninformed user.

We also noticed that in the data we were given, there is a huge difference between the number of reviews for the books and for the ebooks: in our 5-core files, there are 8,898,041 reviews for books, and only 982,619 reviews for ebooks. This tendency also applies for the unfiltered data. We think that this is the result of the period over which the data was collected (from 1996 to 2014). In comparison, the "Amazon kindle" was released only in November 2007[1]. Thus, reviews about books were made long before ebooks were even popular.

5 Data Analysis

5.1 Page and Rank analysis

One aspect of the analysis that has been done is to look if there is some link between the number of pages for a book or a ebook and its reviews. Some could say that a smaller book might be more appreciated as it's quicker to read, or conversely people could dislike it as there is not enough content in it. If we do correlate the number of stars with the number of pages, the Pearson correlation obtained is of -0.034 (-0.005 for ebooks). If we do the Pearson correlation between the average sentiment and the number of pages, the score obtained is -0.038 (-0.007 for ebooks). Both are pretty low scores, and one might conclude that a bigger or smaller book does not affect the client review, other parameters as the content are more relevant. However, the Pearson correlation shows only linear correlation, so the conclusion can only be that there is no linear correlation between those two parameters.

Other correlations tried are - between the sales rank and the score given : The Pearson correlation obtained is of -0.056 (-0.001 for ebooks) between

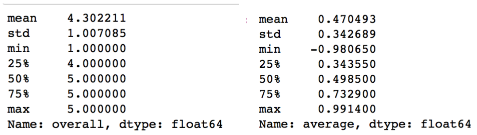
[1] https://en.wikipedia.org/wiki/E-book#2000s

the number of stars and the sales rank, and -0.010 (-0.014 for ebooks) between the sentiment score and the sales rank. - between the sales rank and the number of pages : The Pearson correlation obtained is of 0.020 (-0.158 for ebooks). We can see that these scores are small, so there is no real linear correlation between those parameters.

5.2 Sentiment analysis and stars

To perform the analysis of a review, we used two metrics. The first one being the number of star which is given on the review and the second one being a score based on the sentiment of the text and summary of the review. The sentiment score is computed using VADER sentiment analyzer from the nltk package. VADER is based on lexicons of sentiment-related words and each words is rated as whether it is positive and negative, and how negative or positive it is. For example, the 'excellent' would be treated as more positive than 'good'. The score Vader returns is between -1 and 1, 1 for a very positive review, -1 for a very negative review, and 0 if it is neutral. (Vader, 2014)

If we group all the reviews by the number of stars and then average the sentiment score for each star, we observe that the sentiment score is consistent with the number of stars given in general. In average, the sentiment will be positive for a five stars review and negative for a single star review.



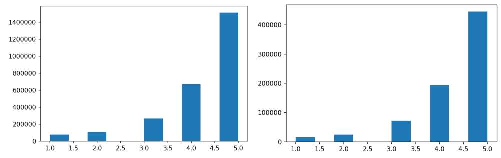
```
mean     4.302211              mean     0.470493
std      1.007085              std      0.342689
min      1.000000              min     -0.980650
25%      4.000000              25%      0.343550
50%      5.000000              50%      0.498500
75%      5.000000              75%      0.732900
max      5.000000              max      0.991400
Name: overall, dtype: float64  Name: average, dtype: float64
```

(a) Sentiment (b) Rating

Figure 1: Books: statistics on metrics

We can observe from figure 1 that the majority of the reviewers give the highest rating possible with the median being at five stars, but when looking at the sentiment of the reviews the median corresponds to a semi-strongly positive sentiment. Our first intuition would be to say that reviewers tend to give the highest rating even when they don't really feel it. But we will discuss about it more in details in the analysis of the ratings through time.

We can see on the bar plot that the distribution of the ratings is exactly the same for the kindle than for the books, even though the number of ratings is one order of magnitude higher for the books.



(a) Books (b) Kindle

Figure 2: Distribution of ratings

5.3 Evolution of reviews over time

We plotted the reviews and their sentiment over time, for the books and the ebooks. The sentiment values are displayed between 1 and 5 for an easier comparison with the ratings.

We noticed that before 2012, the number of reviews for both the books and ebooks is too low to make any conclusion (below 200), thus leading to a lot of noise :

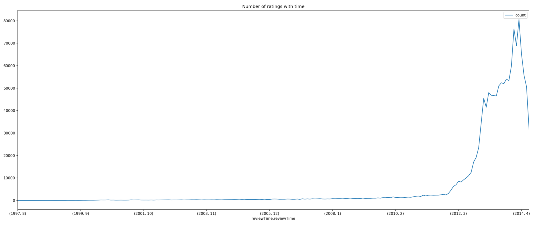


Figure 3: Number of reviews over time for the books.

We observe the same trend for the ebooks.

Because of that, we choose to only display the evolution of the reviews starting from 1st January 2011 :

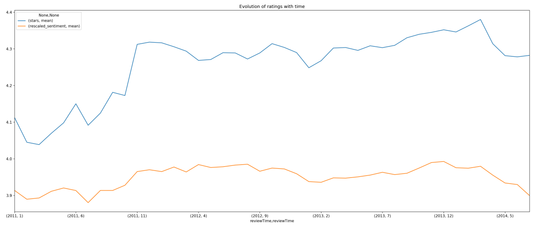


Figure 4: Evolution of books reviews over time.

we can observe that the reviews are quite constant and almost the same between the books and ebooks, with a small advantage for the latter. Interesting thing, the average reviews follow the trends of the average sentiment, which is obvious. Happier comment means a better rating.

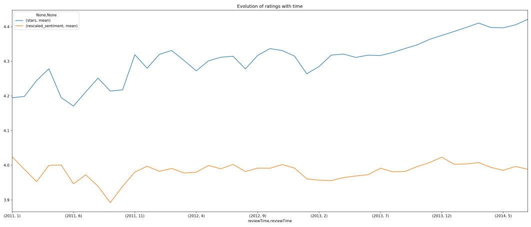


Figure 5: Evolution of ebooks reviews over time.

It can be noted that the sentiment score is slightly shifted downward compared to the ratings score, suggesting that readers give more generous ratings than what is reflected in the reviews they wrote. This confirms what we have seen earlier and is illustrated by the majority of five-stars ratings in the reviews.

5.4 Reviews per authors

Since matching the books and ebooks on the title wasn't possible, we tried to match on the authors, in order to see if for a given authors, books and ebooks reviews were different.

To compute the score of a book, we use two approaches:

1) A weighted average of the stars taking into account the helpfulness of the review as described below.

Let $s_{i,j}$ be the jth rating of book i and n the number of ratings for this book. Let $k_{i,j}$ be the number of the person who reviewed the review $s_{i,j}$ and let $u_{i,j}$ be the number of reviewer who found the review helpful among the $k_{i,j}$ reviewers.

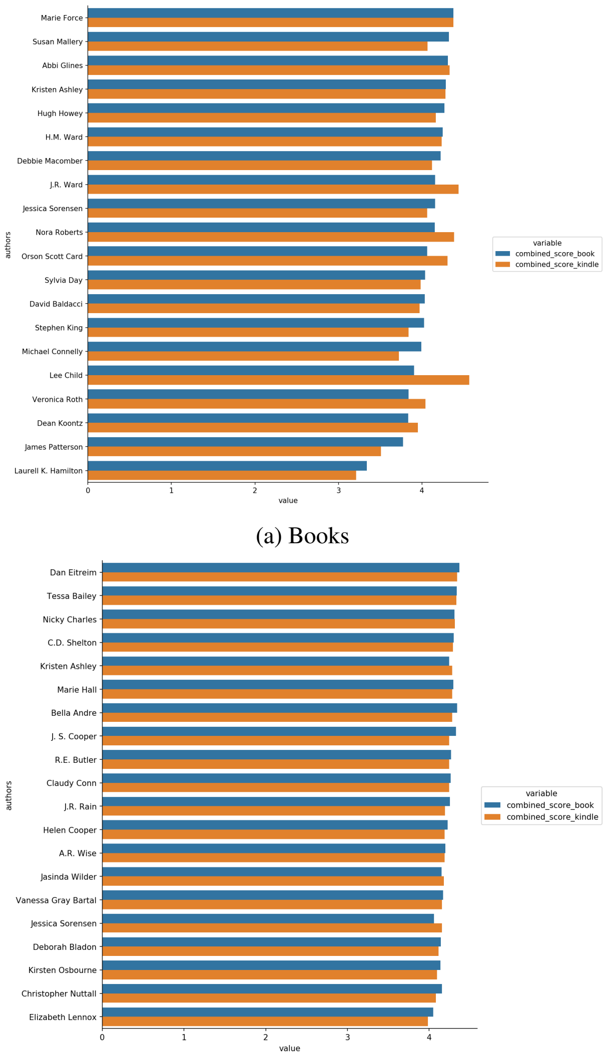
$$w_{i,j} = \begin{cases} \frac{u_{i,j}}{k_{i,j}}, \text{ if } k_{i,j} \neq 0 \\ 0.5 \text{ if } k_{i,j} = 0 \end{cases} \tag{1}$$

The weighted average is then:

$$S_i = \frac{\sum_{j=1}^{n} w_{i,j} s_{i,j}}{\sum_{j=1}^{n} w_{i,j}}$$

2) A weighted average of the sentiment's intensity using VADER in the reviews taking into account the helpfulness with the weight being derived similarly as above.

Then, for each support, we took the authors with the most reviews in it and compared to their reviews from the second support. Then for each authors, we took the average of the score of their books and obtained the following results:



(a) Books

(b) Kindle

Figure 6: Rating average per authors

As we can see, in both case there is a slight advantage to the ebooks, even if given the difference in reviews count between the two support it is difficult to draw meaningful conclusion.

6 Conclusion

We have seen that people in their reviews give generous ratings. This could be the result of the problem of incentivized reviews that Amazon has been fighting against for years (Independent, 2016). To answer our initial question, it seems that people tend to give better ratings for ebooks, but the difference is very thin and could just be noise in our data.

References

Deepak Kumar 2015. *eBooks vs Books (Pros and Cons): The Never Ending Debate.* http://device.is/1Jr8a9x

J. McAuley, C. Targett, J. Shi, A. van den Hengel 2015. *Image-based recommendations on styles and substitutes.* SIGIR, http://jmcauley.ucsd.edu/data/amazon/, http://cseweb.ucsd.edu/ jmcauley/pdfs/sigir15.pdf

ScrapeHero 2014. *How to prevent getting blacklisted while scraping.* https://www.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping/

Datahut 2016. *Tutorial: How To Scrape Amazon Using Python Scrapy.* http://blog.datahut.co/tutorial-how-to-scrape-amazon-using-python-scrapy/

Hutto, C.J. & Gilbert, E.E. 2014. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI.

Independent 2016. *Amazon bans biased reviews that have been influenced by brands* http://www.independent.co.uk/news/business/news/amazon-bans-biased-reviews-incentivised-a7344761.html/