Bookseller Recommendation
Using Amazon Book Review Dataset

Sumeet Bhalla, Nikhil Kumar Singh, Shivam Agarwal, Kishore Madhava Muruganandan

Abstract - The advancements in the field of Information Technology has led to the increase in the size of data exponentially. Rise of recommendation systems have greatly addressed this problem, but most of the recommendation systems present a list of recommendations to the end users which is visually unappealing and unintuitive for the common user. In this paper, we demonstrate the use of data visualization and data analytics in the recommendation systems for improving business model and providing an improved user experience and a much more intuitive story for users to easily derive meaningful inferences.

---

1. INTRODUCTION

'Data is the new Bacon' the tagline of IBM research [12] truly justifies the importance of data in the current scenario. The advancement in the field of computer and information technology has made it easier to collect and store large amount of data from various sources easier, and digitalization allows this information to be shared over various channels. But as the information size is increasing exponentially we face a major problem 'Curse of Dimensionality', There is an information load in every field and this has become a major problem. It makes the decision-making process more complex when too much information is available.

In this context, the recommendation systems have proved to be a useful tool to address this 'curse of dimensionality'. They allow us to make decisions easily by limiting the data for the users based on their interests. The major challenge in recommendation is choose the correct filtering approach based on correct knowledge of domain, user preference and market statistics.

Rise of recommendation systems have greatly addressed this problem, but most of the recommendation systems use complex algorithms and present a list of recommendations to the end users.

Analysis of these raw results are often difficult to understand and do not motivate user to use them.

Moreover, flow of information is unidirectional and there are not many tools where user can interact and choose the information he was to see.

To address this issue, we combine the recommendation system with Data Visualization.

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information"

In this paper, we demonstrate the use of data visualization and data analytics for improving business model and providing a recommendation system with improved user experience. This is done by involving the abilities of recommender system and displaying the results using scatter plots, line chart and word cloud instead of simple stacked results list. The data we have used is for Amazon Book Reviews. This project creates recommendation system that can be used by bookstores, book sellers, public libraries and online book stores. Using this tool these booksellers can decide on important questions like which books to stock during which time of the year. They will also be able to find out which books not to stock during what time of the year to avoid losses on unsold books.

It provides a visually interactive view of most popular and least popular books of the year, month, quarter and weeks. Along with a trend analysis of each book

2. MOTIVATION

Firstly, the major motivation of this solution is to solve the dimensionality problem in the domain of books. According to google data[11], currently there are 129,864,880 books in the world. This a huge number for any organization be it bookstore, library or even online platforms to stock these books.

This is a typical problem in any field of study and solving this problem has challenged researchers from a long time. Over several years recommendation systems have proved to be quite useful in proposing the popular books based on demographics, language, context, etc.

But the world is becoming smaller each day and barriers of geographical locations, language and context are diminishing. Interest of readers and crossed these barriers and they like to read books irrespective of their origin language and other physical factors.

User reviews have gained a lot of popularity over several years and use of user reviews in recommendation systems have given a new aspect to the readers. Most recommender systems that are generally used in practice provide a list of recommended items as the only output. The results are in raw format and un appealing to the user. The outcomes are not very intuitive and do not tell a story clearly to keep the viewer engaged.

Currently, Bookstores, Libraries and Online Libraries face a major problem of which books to stock and which not to stock , during which time of the Year and the quantity to inventory to keep. Due to this uncertainty, the libraries and online bookstores have to waste a lot of money and resources on books that might not sell or remain unread when they could have stocked up more on the bestsellers.

According to National Center for Education Statistics [14], average number of books in a public library range between 100,000 to 130,000 and the number of electronic books range between 100,000 to 145,000. Almost 14% of these books will go unread or unsold. This leads to huge losses which can be curbed using our solution.

Our solutions address these problems and offer an interactive view of most popular and least popular books for each year, month, quarter and weeks. Along with a trend analysis of each book.

Another major problem bookstores face is regarding the strategy to store books such that it is appealing to its customers. They have to make decisions like which books should be displayed near the front and which books should be displayed in the back. On which books should they offer promotions, and on which books they should remove. Our solution solves this problem. Stores can clear their stocks of least popular books by giving more discounts while keeping higher profit margins on the bestseller.

3. Visualization Design

We have created our tool as a Visual story that contains multiple Dashboards where we display most popular and least popular books. We present a simple design with the use of scatter plots, line charts and Word clouds.

We have multiple dashboards on our home page showing most popular books and other shows least popular books.

We use a custom interactive Timeline Graph to show the top three most popular books and top three least popular books in a 5-year timeframe. User can Drill down on the timeline and see the most/least popular books in a quarter, month and day.

Popularity can be selected based on different attributes.

- Ratings - these are the average ratings of each book as per amazon
- No. of Reviews - these are the total number of reviews that the book has received.

This gives this solution a unique feature of not just finding the most popular books, but also finding the most talked about books using the number of reviews. A book might be rated very low but if it's a controversial book it might be reviewed by a lot of

people, a book store should be keeping is this as more number of users can be interested in reading this book.

There are three types of Interactive Visualizations in our project:

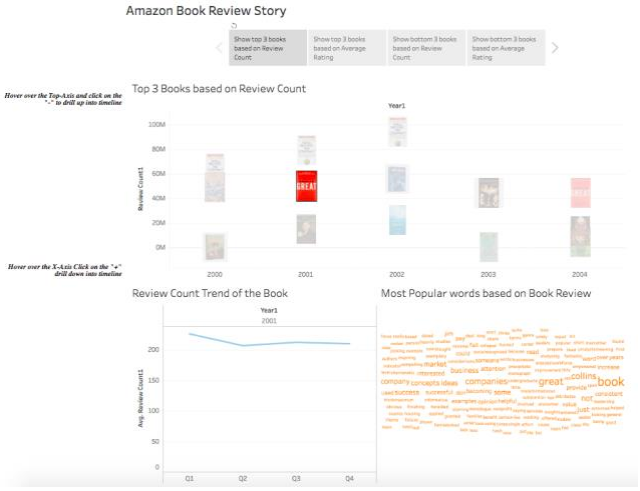
Firstly, a custom timeline scatter plot, which is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. We have created custom scatter plots where we have time on X axis and Popularity on Y axis. Each point of the scatter plot refers to a book and instead of a point it shows a book image.

This is an interactive graph we can drill up and drill down functionality, so we can fine tune the time to a year, month or week.

If we click on a particular book it shows popularity trends of the book for these ten years in a line chart. A line chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically.

In our design we have line chart to show the popularity trends of a book. for example, if popularity of a book is decreasing from 2000 to 2004 a bookseller may choose to reduce its stock even if currently it has an average rating.

On clicking on any book, we also display a word cloud.



A word cloud is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free-form text. Tags are usually single words, and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

Word Cloud is one of the popular visualization technique to depict the content of textual data in one picture. Textual reviews are helpful alternative to the ratings as ratings do not give much information. Text can cover much broader aspects of reviews.

We show a word cloud of frequent words in different review of each book. This can provide an overall sentiment of the reviews in a single visualization.

Using these visualization, a bookseller can get a good idea about the popular books he needs to stock and which books to avoid

4. METHODOLOGY

4.1 Data Gathering

We downloaded 10 years of Amazon Book Review data [13]. Data provided us following information:

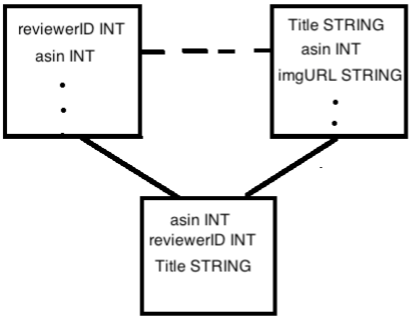
- **reviewerID** - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- **asin** - ID of the product, e.g. 0000013714
- **reviewerName** - name of the reviewer
- **helpful** - helpfulness rating of the review, e.g. 2/3
- **reviewText** - text of the review
- **overall** - rating of the product
- **summary** - summary of the review
- **unixReviewTime** - time of the review (unix time)
- **reviewTime** - time of the review (raw)

We also used metadata set for the product details which provided us following details [13]:

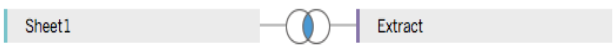
- **asin** - ID of the product, e.g. 0000031852
- **title** - name of the product
- **price** - price in US dollars (at time of crawl)
- **imUrl** - url of the product image
- **related** - related products (also bought, also viewed, bought together, buy after viewing)
- **salesRank** - sales rank information
- **brand** - brand name
- **categories** - list of categories the product belongs to.

We extracted the details of the top 50 books from Amazon Review Data Set for our story.

4.2 Data Modelling



The amazon review data set [5] does not have name of the book. The book name is present in metadata [6]. We merged the two files to get the name of the book along with the reviewer ID and other details of the book, in our final table, Sheet1, which is used as a data source for tableau [3] for plotting the charts for our visualization.



The Alteryx [4] tool was used to generate the word cloud. That word cloud is stored in .tde format in a file named Extract, which is recognized by Tableau. This file was merged with our sheet, Sheet 1, to generate the final data source and produce the same word-cloud in our Tableau. Then, this word cloud was added in all of our dashboards to generate our final story.

4.3 Implementation

We used python to first split the 10Gb review data file into smaller splits since none of the editors were able to process such a huge amount of data. We did the same for the Metadata file as well.

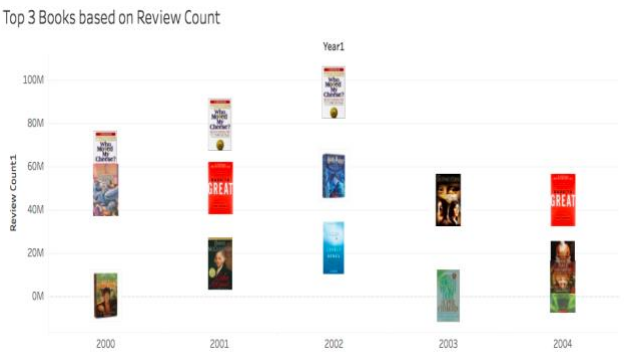
After that we cleaned the JSON and removed the unwanted fields and mapped the book data in one dataset to its corresponding name in the metadata file to get the names of books instead of their ID's.

After that we transformed this code from a JSON to a csv for size reduction and better management. We used Tableau 10.5 [3] for creating our visualizations and also used Alteryx [4] for creating a word cloud for each book based on the review text data.

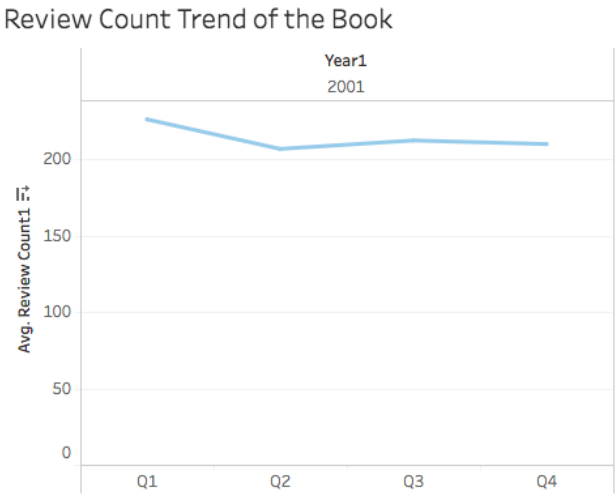
Based on each year, we created ranking on Review Count of each book and Ratings of the book. We then used this rating to plot timeline scatter plot chart showing Top 3 and Bottom 3 books for each year. Since the timeline is interactive, the user can drill down to quarters, months and even days to see how the values of Top/Bottom 3 books changed over the course of time.

We show the top 3 and bottom 3 books based on 2 different parameters namely rating given by the users and also the review count of that book. This helps us find not just the popular books based on the ratings given by the user but also the most talked about books even though they might be rated low e.g. some controversial books.

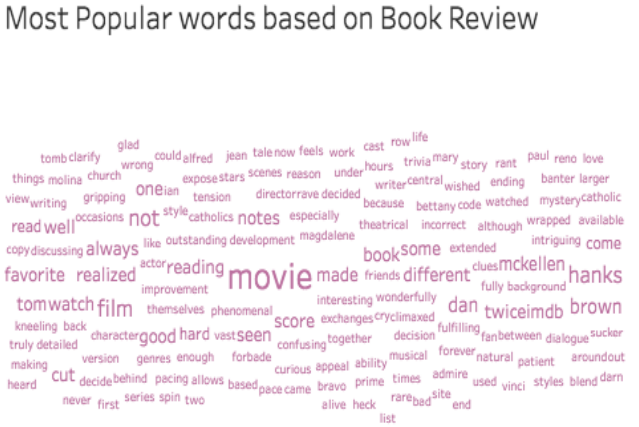
We also use the image URL provided in the data to change the shape of the scatter plot dots to custom images of the books so that it's easier and more intuitive for user to see the results without the need of a legend.



We also created a line chart and linked it to the above main chart in the dashboard so that the user could click on any of the books shown in the above graph and see that books specific trend based on the number of reviews or ratings and see how the book has been performing over a course of time. This can help the customer decide if he wants to stock up a book or not e.g. if he sees the book whose ratings have constantly been declining, he can decide not to stock up that book.



Also, we used Alteryx [4] to create a word cloud for each book. First, we loaded our dataset into the tool and applied certain cleaning operations to remove the unwanted characters like random spaces, exclamation marks and apostrophe marks. Secondly, we split the data based on space as a delimiter and then saved it as a twb file for tableau. Then we loaded this file in tableau and removed the unwanted preposition and joining words to make the word cloud more effective. We linked this word cloud to the main chart above so that the user could click on any book and see the most popular words used to describe that book based on the review data.



Finally, we stitched all the dashboards and create a story for the user to tell the entire story at one place. The user can see the 3 most/Least popular books, the trend of that book over the course of time and also the most popular words used to describe that book at the same place.

5. EVALUATION PLAN

The project gave us many insights regarding the readers, bookstores, and impact of user ratings on businesses and readers. User rating dominate the business strategies.

During our statistical analysis of data, we found that reviews and ratings have an huge impact of the business model of a bookstore. The bookstores can save a lot of money if they base their business strategies on the reviews and ratings.

We also realized that textual data cover a greater range of reviews and more helpful to the users, Hence we also generated a word cloud which shows most frequent words in different review of each book. This can further justify the ratings and can provide an overall sentiment of the reviews in a single visualization.

Our dashboard is amalgamation of these techniques and broadly cover the concepts taught in the Data Visualization class like Text Analysis, Maps & Cartography ,Visual Analytics & Dashboard .

6. FUTURE WORK

Although our Dashboard covers a vast set of book collection, but our views are based on only Amazon Reviews.

Some of the other platforms like Goodreads, New York Times and Huffington post provides most accurate and helpful reviews.

In Future we aim at building a platform that aggregates the results from all these sources and provide popular books based on different Sources.

Also, the dataset we used for this project is static. It does not incorporate the changes for subsequent years. We propose to connect our visualization engine to a database, so that if the database is updated in the consecutive years, the visualization engine pulls up the most recent data from the database instead of a static data source.

To be more user friendly, we can introduce the option of receiving mails about the book trends for every month, quarter and year. This avoids the stakeholder running the visualization engine every time he needs an insight. The user can register for weekly, bi-weekly or monthly book trend emails.

7. ACKNOWLEDGMENTS

We would first like to thank our Professor Dr. Sharon Hsiao for the constant encouragement and motivation for the project.

We would also like to thank my fellow classmates for the amazing reviews and feedback which helped us to improve our work.

We would also like to thank Julian McAuley for providing the Amazon Book Review dataset.

Finally, our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

8. REFERENCES

[1] https://en.wikipedia.org/wiki/Data_visualization

[2] https://en.wikipedia.org/wiki/Scatter_plot

[3] https://www.tableau.com/trial/tableau-prep?utm_campaign_id=2018114&utm_campaign=Prospecting-CORE-ALL-ALL-ALL-ALL&utm_medium=Paid+Search&utm_source=Google+Search&utm_language=EN&utm_country=USCA&kw=tableau&adgroup=CTX-Brand-Core-E&adused=265857380818&matchtype=e&placement=&gclid=CjwKCAjwt5DXBRAtEiwAa3vyEkIkg2E1WUyHJBFecws82QwgnHbUA-zZj8iiyRLLS5wAcnvn_YMTFxoCnpsQAvD_BwE&gclsrc=aw.ds&dclid=CPGR6p-n3toCFULWZAod-I8HaA

[4] https://www.alteryx.com/

[5] R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016

[6] J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015

[7] Belgin Mutlu, Eduardo Veas, and Christoph Trattner 2015. VizRec: Recommending Personalized Visualizations.

[8] Christian Richthammer,Johannes Sanger,Gunther Pernul. Interactive Visualization of Recommender Systems Data

[9] Design principles for visual communication. (2011).Agrawala, Maneesh, Li, Wilmot, & Berthouzoz, Floraine. Commun. ACM, 54(4), 60-69. doi: 10.1145/1924421.1924439

[10] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5-53.

[11] https://www.telegraph.co.uk/technology/google/7930273/Google-counts-total-number-of-books-in-the-world.html

[12] https://www.ibm.com/blogs/business-analytics/data-is-the-new-bacon/

[13] http://jmcauley.ucsd.edu/data/amazon/

[14] https://nces.ed.gov/datatools/