Abigail Tee Ren Hui (A0146695N) | Chua Wei Yang Rex (A0206487X) | Muhamad Imran Bin Mohd Yusof (A0206485B) | Kuah Jun Wei Jerome (A0052238L) | Seah Fang Ying (A0206532N)

Improving Book Searches on Online Bookstores

Abstract

This paper examines the concept of a contextual search engine as well as an automatic reader-centric book genre tagging to help online bookstores improve their users book search experience and in turn increase book sales on their website. Overall, the results of the models served as good proof of concept and could be followed-up upon for further development and implementation.

Github Link: https://github.com/Abigail-Tee/Group-Project-BT5153.git

1. Introduction

With the introduction of the internet, companies have set-up or moved their businesses online to increase revenue by extending their reach and product offerings while reducing overhead costs for inventory management, rental and manpower.

An example is the introduction of online bookstores. Amazon, launched in 1995, was one of the first bookstore with the business model of selling a large (almost infinite) catalogue of books online. Even though it has since grown into a multi-billion-dollar revenue generating company selling almost everything, in 2014, Forbes reported that book sales still accounted for 7% of the company's annual yearly revenue approximated at $75 billion (Bercovici, 2014). By 2018, Amazon was dominating the book market in the United States with 807 million books and 560 million e-books sold (Day & Gu, 2019).

Online bookstores such as bookdepository.com also entered the market to cover regions such as Europe and Asia and increasing the variety of books available from different regions online, raking in millions in revenues before being acquired by Amazon in 2011. In Canada, almost 50% of books purchased amounting to ~$980M were bought online in 2016 (Shaw, 2017).

1.1 Problem Statement

Online bookstores have become the way in which people search for and purchase books. However, with the large number of titles available, we observed some teething issues when searching for books online which may adversely affect online book purchases. The online bookstore browsing experience is tailored from a business perspective rather than the customers.

- **Keyword search requires a customer to have knowledge about a book.** Search engines on these online bookstores retrieve results based on keywords, title, author or International Standard Book Number (ISBN). Such search algorithms require customers to have prior knowledge about a book and limits explorative searches based on the content of a book.

- **Book categorisation does not reflect customers' perception of a book's genre.** Customers can explore the book collections by browsing through book categories; however, these categories are often provided from a business or publishers' perspective and does not reflect the customer's perception of a book's genre. For example, the book "Harry Potter and the Sorcerer's Stone" is categorised under "Children's General Story Books" or "Funny Books for Kids" on bookdepository.com, as compared to customer-curated genres from Goodreads.com which classifies the book as 'Fantasy' or 'Fantasy, Magic'.

1.2 Purpose of Project

This project aims to revamp the book search experience for customers visiting an online bookstore and in turn increase sales through the following:

1. **Contextual Search Engine**
   Customers may only recall some context of the books such as the plot or themes of the book but not the typical features required by the bookstore search function. Hence, in order to improve customers' experience on an online bookstore, we intend to prototype a search function for books based on the books' content. This search algorithm would surface popular book titles based on the text description provided by the customer, introducing an alternative method for customers to search for the books they want as well as potentially discover new books.

2. **Reader-centric Book Genre Tagging**
   Customer-curated genre tagging is already a feature at Goodreads.com. However, the tagging is done by readers, hence, new books or less read books may not be well-tagged and harder for customers to search. Based on the corpus of data of user-defined genre tags, we aim to predict the genre tags of new and less read books based on their synopsis.

2. Data Pre-processing and Exploration

2.1 Dataset

The dataset used comes mainly from 2 sources:

- **Book Depository.** The first dataset is a Book Depository dataset that is available on Kaggle (Simakis, 2020). Data parameters of this Kaggle dataset is listed in Appendix A. The primary fields that are relevant to our project objectives are the ISBNs, title and number of ratings.

- **Goodreads.** In addition, using the ISBNs as unique identifiers from the Book Depository dataset, additional data on the books' synopsis/description and customer-curated genres, which are also referred to as shelves on Goodreads.com, were scraped from Goodreads.com using their API client in Python or via a web scraper.

Data used for contextual search engine and prediction of book genres. The data from Book Depository and Goodreads were combined and the combined dataset is used for the contextual search function algorithm as well as the prediction of genre for new and less read books. With reference to *Table 1* below, after combining the Kaggle dataset with the information scraped from Goodreads.com, and removing null values, the total number of observations is 344,639.

Table 1. Project Dataset Description

Data	Type	Source
ISBN13 [A]	Int	Book Depository
Title [A]	Str	Book Depository
Synopsis [A]	Str	Goodreads
Total No. Ratings [A]	Float	Book Depository
User-Defined Genres [A]	List of str	Goodreads
Reviews [B]	Str	Goodreads

(A) 344,639 unique books
(B) 30,000 user reviews, about 15 user reviews for each randomly selected book

Data used for contextual search engine model evaluation. To evaluate the accuracy of the contextual search engine (refer to Section 4.1 for methodology and model evaluation), user reviews for the books were required. Customer reviews were scraped from Goodreads.com. The reviews were subsequently curated based on the number of words to ensure that the reviews selected contain more description of the book instead of a mere expression of sentiment such as "This book is a page-turner, a must read", and are of an acceptable length to mimic actual book searches.

2.2 Data Pre-processing

2.2.1 Text Cleaning of Book Description

For data pre-processing, the text of the book description was cleaned using methods such as lower casing and removal of tags, punctuations and stop words. In addition to the stop words in the Gensim library, words such as 'bestseller', 'bestselling' and 'prize' that appear quite often in a book's synopsis but do not describe and distinguish the books were also removed. Lemmatisation was also used to remove inflectional endings and return the base form of the words while maintaining certain meaning of the words. Finally, all non-alphabet characters were removed.

2.2.2 Filtering of Books

Books with at least 50 words of description and at least 5 genres. *Figure 1* below shows the distribution of description length while *Figure 2* shows the distribution of number of genres for the books. Books with fewer than 50 words of description or 5 genres were removed as they may not adequately describe the book.

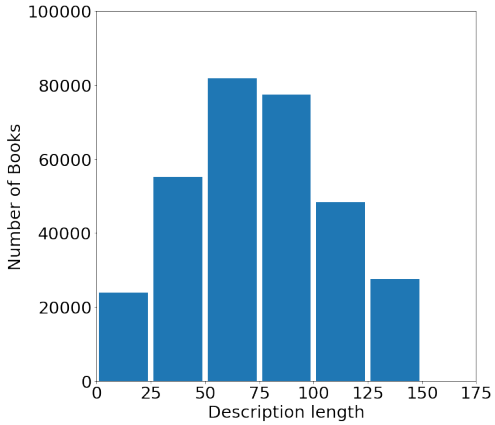

Figure 1. Distribution of Books by Description Length

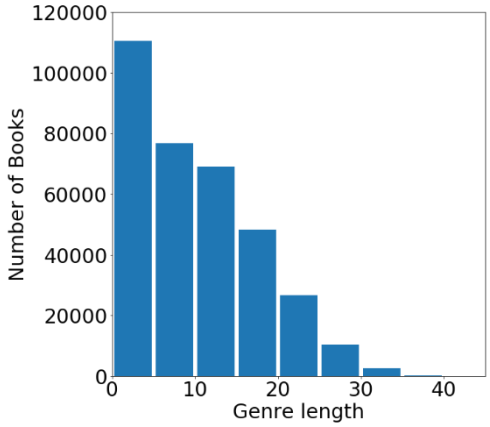

Figure 2. Distribution of Books by Genre Length

Fiction Books. Fiction and Non-fiction books are the two broad genre categories for books. The books under these two genres have a vastly different range of sub-genres and distinct descriptions. Hence, different models might be required for the contextual search engine and genre prediction models for these two broad genre categories so as to achieve better search relevance and prediction accuracy.

Also, due to limited computational capabilities, we were not able to run the proposed models for all the books. As such, we focused on Fiction books to test our proposed model. A total of 50,000 books were used for training and 10,000 books were used for testing.

3. Extracting Text Features

Using two different methods, Latent Dirichlet Allocation (LDA) algorithm and Doc2Vec (Gensim), we extracted text features from the synopsis for each book and created 2 matrices for modelling. The text features were used for the Contextual Search Engine and Genre Prediction.

LDA. The model for LDA algorithm was trained using the count vectors of the description of 50,000 books and based on 2,000 topics. The algorithm generated a topic probability distribution for each book, with each book being assigned the topic category with the highest probability. We selected 2,000 topics due to limited computational capabilities. As previously mentioned, for actual implementation on an online bookstore, it would be ideal to run the LDA model on a wider selection of books and with higher number of topics.

Doc2Vec. A Doc2Vec model was trained using word tokens for each of the description of the same 50,000 books and based on 2,000 features. Doc2Vec employs a similar unsupervised learning technique as Word2Vec to learn the meaning of the entire document instead of the words. Hence, instead of creating a feature vector for every word in the corpus, the Doc2Vec method computes a feature vector for every document in the corpus.

3.1 Exploring Results of LDA

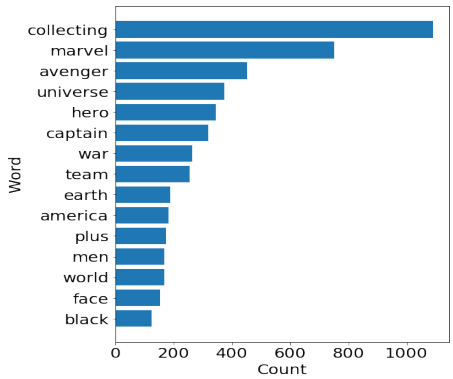

Figure 3. Distribution of the top 15-word occurrences for the most popular topic

We looked at the word occurrences of the top topic identified by the LDA to provide a sense of the coherence of the topics. A histogram of the top 15-word occurrences for top topic with the greatest number of books assigned to it is shown in *Figure 3*. We observe that the top 15-word occurrences cover words such as Marvel, Avenger, Captain and America, a clear indication that these books relate to the Marvel Universe.

3.2 Exploring Results of Doc2Vec

From the Doc2Vec model results, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of the Doc2Vec document vectors for visualization. A scatter plot of the feature vectors for 1000 books belonging to 3 popular genres, Mystery, Fantasy, and Romance, is shown in *Figure 4*. From the plot, we observe that clustering of books is generally quite distinct for the different genres, with "Fantasy" being the most distinct.

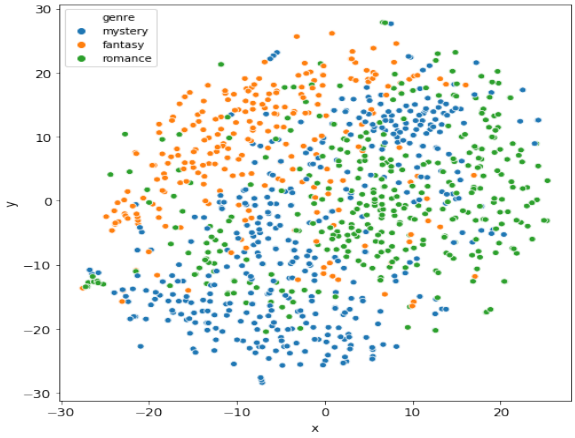

Figure 4. Scatter plot of the feature vectors for 1000 books (mystery, fantasy, and romance)

4. Contextual Search Engine

A contextual search engine provides an alternative method for users to search for a book based on a string of word that matches the description of the book as opposed to traditional search engines which requires users to search for a book based on the author, title or ISBN13.

4.1 Methodology

4.1.1 Similar Books

Based on the search input text, books which have topics that are most similar to the search text are recommended to the users. The LDA topic vectors were used to search for similar books based on the search terms and the search results were ranked based on their cosine similarity. The vectors from Doc2Vec were not used as the text searches are typically short and summarising them into feature vectors may not produce meaningful representations of the text search to match to similar books.

Search Text Pre-processing. Each search text is pre-processed to obtain a clean text for the search engine. Text cleaning steps employed were the same as the steps outlined in Section 2.2.1.

Topic Probability Vector. Based on the trained LDA model, the search text is converted to a topic probability vector which effectively represents the topic distribution of the search text.

Pairwise Cosine Similarity. Cosine similarity between the search text's topic probability vector and all the books' topic probability vector is computed and books with the highest cosine similarity are recommended to the users.

Search Options. Users have the flexibility to filter the search results based on popular books, defined as having more than 100,000 ratings.

4.1.2 Evaluation

The results of the contextual search engine are evaluated using 2 different methods.

Comparison with Online Websites. Search results from the contextual search engine are compared to other online websites and results are evaluated using human judgement. 3 main websites are used in the evaluation. Firstly, Goodreads, a book social cataloguing website that allows users to share reviews and opinions about books. Secondly, Amazon, a popular e-commerce website that sells books and e-books. Finally, Open Library, an internet archive of books. The search methodology employed by Goodreads and Amazon are based on string search of the title, author and/or ISBN13, while Open Library has the option to search for a book based on a string match of the book's content or description.

Book Reviews as Search Text. Book reviews are used as search inputs based on the proposed theory that the reviews could mimic search text and represents the content of the book. The book for which the review was left for was considered as the actual book label. Reviews were used for 4 main genres, (a) Fantasy, (b) Thriller, (c) Science Fiction and (d) Romance. Reviews were scrapped from Goodreads (Section 2.1) and approximately 200 reviews were tested for each genre.

4.2 Results and Discussion

4.2.1 Comparison with online websites

Since the contextual search engine is an alternative search method for users to search for a book based on the book's description, the search input text used for comparison are search texts that represent general descriptions of a story outline. 2 different search texts were used for evaluation, mainly (a) Wizarding Magic and (b) Dystopian End of World.

Table 2. Top 3 unique books from various search engines

Search Text 1: Wizarding Magic
Contextual Search Engine Results
1. Harry Potter and the Prisoner of Azkaban
2. Harry Potter 1-7 Audio Collection
3. Harry Potter and the Goblet of Fire
Goodreads Results
1. Off to be the Wizard
2. Wizard at Large
3. Winder of the Ice Wizard
Amazon Results
1. Wizarding World: Hidden Creatures Scratch Magic
2. J.K. Rowling's Wizarding World: Movie Magic Volume One: Extraordinary People and Fascinating Places: 1
3. Wizarding for Beginners
Open Library Results
1. The Complete Idiot's Guide to the World of Harry Potter
2. The Plot Thickens... Harry Potter Investigated by Fans for Fans
3. Mugglenet.com's Harry Potter should have died: controversial views from the #1 fan site

The results for the book searches for the different search engines can be seen in *Table 2* above (refer to Table B-1 in Appendix B for full book synopsis of the books). With reference to *Table 2* above, based on the search text "Wizarding magic", the books recommended by the contextual search engine were mainly books from the Harry Potter Series, which is a popular fantasy series about wizardry and magic. Harry Potter books are very relevant to the search term and this shows that our search engine successfully finds and recommends books with content that matches the search term.

In contrast, based on the search results of Goodreads and Amazon, the recommended books contain the keywords "Wizard" or "Magic" in the title of the book. While the content of the books is somewhat relevant to the search term, this constitutes merely a superficial match of the book's content. Open Library's search engine matches the search term to the content of the book based on a string match. Although the books' content may contain the search term, the books may not necessarily be relevant to the search term. For example, the third recommended book, "Mugglenet.com's Harry Potter should have died: controversial views from the #1 fan site", is a fan-fiction book that discusses certain ideas in the Harry Potter book such as "Should we pity Voldermort or hate him?" and is not a story about wizarding or magic per se. Therefore, matching the search text to the book's content based on a string search does not necessarily surface relevant books for the user.

Table 3. Top 3 unique books from various search engines

SEARCH TEXT 2: DYSTOPIAN END OF WORLD
Contextual Search Engine Results
1. Nineteen Eighty-Four
2. Allegiant
3. Divergent
Goodreads Results
1. Babylon Working: The End Of the World is just a Beginning (book1)
2. Babylon Working – Part II: A Dystopian Sci-Fi Fantasy Horror
3. Babylon Working: A Dystopian Sci-Fi Horror
Amazon Results
1. The End of the World
2. The End: A Postapocalyptic Novel
3. The End of the World Running Club: Library Edition
Open Library Results
1. Ten to one: selected poems
2. The great inversion and the future of the American city
3. Killer Critique

Another search text, "Dystopian end of world" was used to further illustrate the results of the search engine. The results for the book searches for the different search engines can be seen in *Table 3* above (refer to Table B-2 in Appendix B for full book synopsis of the results). With reference to *Table 3* above, based on the search text "Dystopian end of world", the contextual search engine recommended the book 1984, a popular dystopian novel written by George Orwell, set in a repressive and totalitarian society. The second and third recommended book are books that belong to the divergent series, also set in a dystopian city. The content of these books is highly relevant to the search term and shows that the contextual search engine is able to provide a good match of the search term and book's content.

On the other hand, the books surfaced from Goodreads and Amazon are books with the keywords "end of world" or "dystopian" in the title. This limits the search results to only books that have the relevant keywords in the title. Open Library's search algorithm matches the key words in the search term to the book's content. Truncated words such as "end of" are matched to the content of the book, without considering the other words such as "dystopian" and "world", which discards the whole meaning of the search term. Consequently, the recommended books are irrelevant to the search term. For example, the first recommended book, Ten to one: selected poems, is merely a collection of poem without any dystopian storyline.

Overall, the contextual search engine provides a good alternative method for users to search for a book based on the book's content which is unparalleled by existing search engines.

4.2.2 BOOK REVIEWS AS SEARCH TEXT

In addition to comparison with other websites, book reviews were used as search inputs to evaluate the contextual search engine.

Accuracy Score. The top 10 books surfaced from the contextual search engine were compared to the actual book for which the review was left for. An accuracy score of 1 is assigned if the top 10 search results contain the actual book and 0 otherwise.

Table 4. Accuracy Scores for various Genres

GENRE	FANTASY	THRILLER	SCI-FICTION	ROMANCE
SCORE	5%	5%	5.5%	4.5%

Book Reviews are a Poor Evaluation Method. With reference to *Table 4* above, there is generally a poor match between the top 10 search results from the contextual search engine and the reviews. Upon further analysis, this is mainly attributed to the content of the reviews whereby reviews usually contain opinions of the characters and authors and are not representative of the content of the book. It is thus unsurprising that there is a poor match between the book review and the contextual search engine's recommended books. Therefore, it should not be used as a method to evaluate search engines and the low accuracy scores is not a good reflection of the relevance of the search engine.

5. Genre Prediction

The purpose of the genre prediction model is to tag new and less read books with genres that can adequately capture the essence of the book.

5.1 Methodology

5.1.1 MULTI-LABEL CLASSIFICATION METHODS

Books are generally multi-dimensional and often belong to more than one genre. For example, the Harry Potter book series can be tagged as 'Fantasy', 'Magic' and 'Adventure', which all accurately capture the essence of the book. In order to achieve this, multi-label classification methods were used to develop a model to predict the top 5 genres of each book.

Predictor Variables. The 2 sets of vectors from the book descriptions derived in Section 3 through LDA and Doc2Vec were used as the predictor variables.

6 multi-label classification methods explored for the genre prediction model are listed below.

Baseline Model. The top 5 most common genres for all books were used as the predicted genres for all the books.

This model will be used as a baseline model for comparison with other multi-label classification methods.

Classifier Chains. This model uses problem transformation to transform a multi-label classification problem into *n* multi-class classification problem, where *n* is the total number of unique genres in our case. In the first problem, the input data is used to predict the first class. The predictions are then added to the input data for the prediction of the next class. The process is repeated until all classes are predicted, thus forming a chain of classifiers, as illustrated by the name of the method.

Label Powerset. Another model which uses problem transformation method, where the multi-label classification problem is transformed into a multi-class classification problem. The genre labels are combined to form unique combinations of different genres, and each combination is then treated as a unique class. Here, the total number of genre combinations forms the total number of classes. Then, a multi-class classifier is fitted to predict the best combination of genres, thus giving its name, label powerset.

Adapted Algorithm. This model uses a base classification algorithm and adapts it for a multi-label classification problem. We used Multilabel k Nearest Neighbours (MLkNN), where k-NearestNeighbours classification is used to find nearest examples to a test class and then Bayesian inference is used to select the assigned labels.

Ensemble Method. The ensemble method used in this paper is the Distinct RAndom k-labELsets multi-label classifier (RAkELd). This method divides the label space into equal partitions of size k, trains a Label Powerset classifier per partition and makes its predictions by summing the result of all trained classifiers.

Neural Network. Neural networks and convolutional neural networks were explored for our multi-label genre classification model.

5.1.2 Transforming Genre to Labels

The actual labels for the genre prediction refer to the user-defined genre tagging tied to each book. As mentioned in Section 2, the book genres obtained from Goodreads.com are tagged by users, i.e. user defined. Allowing customers to tag book genres can result in (i) rare occurrences of book-specific genres such as 'star trek the next generation' and 'buffy the vampire slayer, (ii) occurrences of uncommon genres such as 'Georgian romance' and 'fractured fairy tales', and (iii) occurrences of irrelevant genres like 'm f m' and 'm m m m'. All these add up to 977 unique genres in the training data.

Given our project purpose, we are only concerned with tagging new and less read books with genres that users can relate with. Therefore, we retained only the most popular 50 genres in the dataset and each book is then labelled using the top 5 user-defined genre tags that are part of the 50 common genres. As only fiction books are included in the dataset, the 'Fiction' genre tag was removed for the top 50 list. *Figure 5* below shows the top 50 genres based on the number of books assigned with the genre.

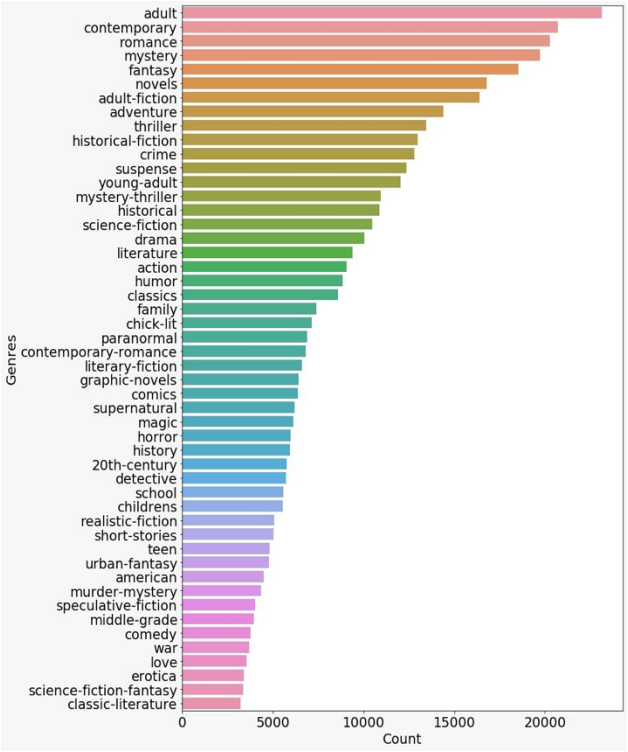

Figure 5. Distribution of books based on top 50 Genres

5.1.3 Evaluation Metric

All the models explored requires the use of multi-label classification methods to produce the predicted probabilities for all 50 genres. Only the top 5 genres with the highest probabilities for each book is retained as the predicted genre for evaluating the models.

For the evaluation, each model is evaluated based on the percentage of books with at least 60% of matched genres. Based on the project purpose, we assess that it is not necessary to match all genres of each book since the genres tagged to books can be subjective and having at least 60% of the genres correct is sufficient to enable users to know the content or locate new and lesser known books.

5.2 Model Selection

5.2.1 Stage 1: Model Selection using default model Settings

As mentioned in Section 5.1.1, we explored several models using the text feature extracted from LDA and Doc2Vec as the predictor variables to assess which vector would provide the best genre prediction.

For the assessment, we took a two-stage approach. In the first stage, models were employed based on their default settings. The better text feature extraction method and

models are then brought over to Stage 2 for further hyper-parameter tuning to select the best model.

The results from Stage 1 is shown in *Table 5* below (Refer to Table C-1 in Appendix C for model details and results). From the results in Stage 1, we observe that the document embedding method, Doc2Vec, consistently performs better than LDA. The results also highlight that Classifier Chains, Ensemble and Neural Network methods generally performs better among the 5 models. As Label Powerset is slow to train and the Ensemble Method is also a representation of a Label Powerset method, we did not explore it further in Stage 2. Hence, only the 3 models highlighted were explored further in Stage 2 to determine the best model.

Table 5. Results of Stage 1 Prediction Models using Default Settings

	% of books with at least 60% of genres matched (e.g. 3 out of 5)	
Model	LDA	Doc2Vec
Baseline	15.1%^	
Classifier Chains	15.6%	34.8%
Label Powerset	14.3%	35.6%
Adapted Algorithm	14.7%	13.2%
Ensemble (rakeld)	12.5%	24.1%
Neural Network	45.6%	50.0%

^ not based on either LDA or Doc2Vec

5.2.2 Stage 2: Model Selection with Hyper-Parameter tuning

For Stage 2, we tune the models by trying several base classifiers for the different methods and adjusted the hyper-parameters of the base classifiers.

The results from Stage 2 is shown in *Table 6* (Refer to Table C-2 in Appendix C for model details and results). From the results of Stage 2, we observed that the models selected from Stage 1 generally performed better post tuning. Out of the various models, the Neural Network model (based on Convolutional Neural Network with word embeddings) performed better by a significant margin with an accuracy score of 72.4%.

Table 6. Results of Stage 2 Prediction Models with Parameter Tuning

		% of books with at least 60% of genres matched (e.g. 3 out of 5)
Model	Base Classifier	Doc2Vec
Classifier Chains	Stochastic Gradient Descent	45.5%
	Logistic Regression	46.4%
Ensemble (rakeld)	Stochastic Gradient Descent	51.8%
	Logistic Regression	53.1%
	Random Forest	50.6%
Neural Network	8 dense layer (deep) with dropout	49.1%
	CNN with word embeddings	**72.4%^**

^ not based on Doc2Vec

5.2.3 Convolutional Neural Network Model

This section provides an assessment of the CNN model in greater detail.

When tuning the CNN model, we explored several embedding sizes ranging 20 to 500 and achieved the best results with an embedding size of 500. The large embedding size required could be due to the size of our word corpus (descriptions for 50,000 books) and significant number of genre (50 genres) to predict. Hence, a larger dimensional space works better at distinguishing the words and book description.

For the model, we also applied 5 filter sizes with 200 filters each. Each filter represents an n-gram which allows the model to learn from phrases up to 6 words (2 to 6-grams), enabling the model to better learn the semantics of the words in the book description.

Table 7 below shows the results of the CNN model. In terms of performance, the CNN model was able to predict at least 60% of book genres for 72.4% of the books in our test dataset of 10,000 books. It was able to predict all genres for 20.4% of the books and 98.1% of the books when considering predicting at least 1 of the genres correct.

Table 7. Overall Performance of CNN Model

Model	% of books with at least xx% of genres matched				
	20%	40%	60%	80%	100%
CNN 2-6 grams 200 filters Embedding size = 500	98.1%	90.4%	72.4%	42.4%	20.4%

5.3 Results and Discussion

5.3.1 Embeddings versus LDA

Comparing the performance of the different text feature extraction methods, we found that embeddings performed better than LDA for genre predictions. Both document level (i.e. Doc2Vec) and word embeddings used in the CNN model proved superior over LDA, possibly due to the embeddings being better able to extract the semantic behind each word. The models highlighted that word embeddings in conjunction with CNN filters that allowed the model to capture both words and phrases produced the highest accuracy results.

5.3.2 Choice of Evaluation Metric

Table 8 highlights some examples on how the matching of genres is done. From the examples, we observed that the incorrect predictions of the model can be fairly divergent from the actual genres. Taking the case with only 1 matched genre for example, the predicted genres included 'literature' and 'classics' whereas the actual genre had 'horror' and 'fantasy'. This observation highlights the need for models to be evaluated based on its ability to get the majority of genres correct rather than measuring a successful outcome based on 1 or 2 accurate genre predictions. Neither is an exact match necessary in our use case. We found the 60% genre prediction rate used as our model evaluation metric to be adequate.

Table 8. Examples of Predicted vs Actual Top 5 Genre Matching

Predicted	Actual	Matched	
		no.	%
[classics, novels, adventure, literature, science-fiction]	[science-fiction, mystery, fantasy, horror, adult]	1	20%
[science-fiction, speculative-fiction, thriller, adventure, short-stories]	[science-fiction, fantasy, novels, speculative-fiction, science-fiction-fantasy]	2	40%
[science-fiction, historical-fiction, classics, adventure, fantasy]	[classics, fantasy, humor, literature, science-fiction]	3	60%
[historical-fiction, romance, historical, mystery, contemporary]	[historical-fiction, mystery, romance, historical, suspense]	4	80%
[suspense, romance, mystery, thriller, contemporary-romance]	[mystery, suspense, romance, thriller, contemporary-romance]	5	100%

5.3.3 Recall of Genre Prediction

Figure 6 shows the recall of each genre ordered by descending accuracy. We observed that the model performed well for most genres such as 'Romance' (88%) and 'Historical-fiction' (87%) whereas genres such as 'Murder-mystery' (4%) and 'Love' (3%) performed poorly.

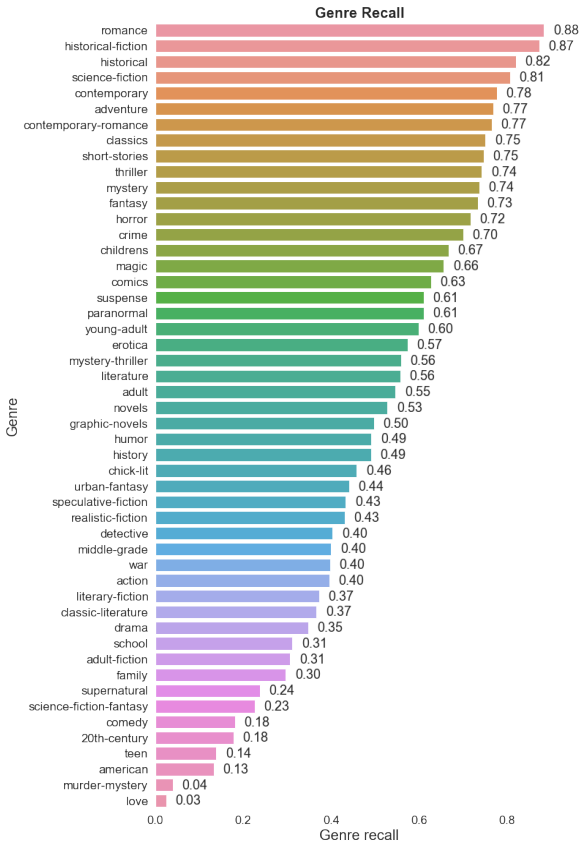

Figure 6. Genre recall

A cross reference of the respective genre frequencies in the training dataset is shown in *Figure 7* below. We observed that the genres that performed poorly (i.e. 'Murder' and 'Love') were less frequent in the training dataset whereas top performing genres (i.e. 'Romance') were more frequent. This suggests a positive correlation between a genre's frequency in the training dataset and its recall. Therefore, the poor results of certain genres could be due to the small number of genre samples in the training dataset.

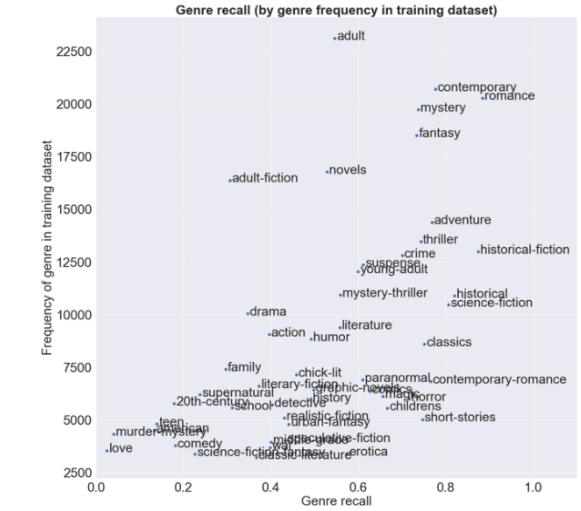

Figure 7. Genre recall versus training data frequency

6. Limitations and Future research

6.1 Contextual Search Engine

In Section 4, we demonstrated that the concept of a contextual search engine performs well when compared with the current market leaders for online bookstores such as Amazon and Goodreads. Based on the results, we were able to produce a model that is able to provide a unique alternative to the customer book exploration and search experience. However, due to the lack of good benchmarks, we acknowledge that there are some limitations in our assessment of the model's effectiveness.

Human Judgement is Costly and Inefficient. Relying on human judgement to evaluate the search engine is costly and inefficient.

Future Research. In order to establish the value of the contextual search engine, it is more practical to conduct A/B testing on an e-commerce website and compare the search-to-purchase conversion rate between the new and old search model.

6.2 Reader-centric Book Genre Tagging

In Section 5, we explored various models to automate the book genre tagging process to enable users to find new and less read books. Our model achieved a high level of accuracy. However, we acknowledge that there are limitations to our model and opportunities for further research as elaborated below.

Limited Scope of Genres. The genre prediction model focused on only Fiction books and was based on a subset of 50k books in order to reduce computational cost in developing this proof of concept.

Infrequent Genre Labels. The analysis on the genre specific recall suggests that the arbitrary choice of top 50 genres could still lead to 'rare' genres that affects the overall model accuracy.

Future Research. There is room to further improve the model accuracy by using a larger corpus of book descriptions to obtain a training dataset with a balanced distribution of genres for the training of our model, or by optimising the number of top genres selected for tagging so as to reduce rare occurrences of genres, or both. In addition, a separate assessment on the model's performance on non-fiction books can help solidify the practicality of the reader-centric book genre tagging concept.

7. Conclusion

Overall, this project achieved both its objective and produced workable models that are good proof of concepts for an alternative search engine and an automated genre tagging system. Further developments could build on these models for possible implementation.

References

Bercovici, J. (2014). Amazon Vs. Book Publishers, By The Numbers. Forbes.

Day, M., & Gu, J. (2019, March 27). The Enormous Numbers Behind Amazon's Market Research.

Shaw, H. (2017, January 30). Book sales continued to migrate online in 2016, but paperback purchases beat e-books.

Simakis, P. (2020). *Kaggle*. Retrieved from https://www.kaggle.com/sp1thas/book-depository-dataset

Appendix A: Data Description for Kaggle Book Depository dataset

Table A: Description of Data for the Kaggle Book Depository dataset

Field Name	Description	Data Type
authors	Book's author(s) name	List of tr
bestsellers-rank	Bestsellers ranking	Int
categories	Book's categories.	List of int
description	Book description	Str
dimension_x	Book's dimension X	Float (in cm)
dimension_y	Book's dimension Y	Float (in cm)
dimension_z	Book's dimension Z	Float (in mm)
edition	Edition	Str
edition-statement	Edition statement	Str
for-ages	Range of ages	Str
format	Book's format.	Int
id	Book's unique id	Int
illustrations-note		
imprint		
index-date	Book's crawling date	Date
isbn10	Book's ISBN-10	Str
isbn13	Book's ISBN-13	Str
lang	List of books' language(s)	
publication-date	Publication date	Date
publication-place	Publication place	ID
publisher	Publisher	Str
rating-avg	Rating average [from 0-5]	Float
rating-count	Number of ratings	
title	Book's title	Str
url	Book relative url	https://bookdepository.com + url
weight	Book's weight	Float (in grams)

Table A above shows the dataset description for the Book Depository dataset which is available on Kaggle. The dataset contains around 1 million items listed on Book Depository, which includes books, CDs, calendars, paperback, hardback etc. Only items that are paperback and hardback, which are usually fiction and non-fiction books, are used for our project.

Appendix B: Book Results from Contextual Search Engine and Other Online Websites

Table B-1 and B-2 below shows the descriptions for the various books. All the book descriptions were sourced from Amazon.com or Goodreads.com.

Table B-1: Search Results for "Wizarding Magic"

Search Text: **Wizarding Magic**		
Book Search Engine	**Title**	**Description**
Contextual Search Engine	Harry Potter and the Prisoner of Azkaban	**Book Description from Goodreads:** Harry Potter's third year at Hogwarts is full of new dangers. A convicted murderer, Sirius Black, has broken out of Azkaban prison, and it seems he's after Harry. Now Hogwarts is being patrolled by the dementors, the Azkaban guards who are hunting Sirius. But Harry can't imagine that Sirius or, for that matter, the evil Lord Voldemort could be more frightening than the dementors themselves, who have the terrible power to fill anyone they come across with aching loneliness and despair. Meanwhile, life continues as usual at Hogwarts. A top-of-the-line broom takes Harry's success at Quidditch, the sport of the Wizarding world, to new heights. A cute fourth-year student catches his eye. And he becomes close with the new Defense of the Dark Arts teacher, who was a childhood friend of his father. Yet despite the relative safety of life at Hogwarts and the best efforts of the dementors, the threat of Sirius Black grows ever closer. But if Harry has learned anything from his education in wizardry, it is that things are often not what they seem. Tragic revelations, heartwarming surprises, and high-stakes magical adventures await the boy wizard in this funny and poignant third installment of the beloved series.
	Harry Potter 1-7 Audio Collection	**Book Description from Goodreads:** Enjoy the complete Harry Potter series performed by the Grammy Award-winning Jim Dale. This complete unabridged audiobook collection contains the following: Harry Potter and the Sorcerer's Stone Harry Potter and the Chamber of Secrets Harry Potter and the Prisoner of Azkaban Harry Potter and the Goblet of Fire Harry Potter and the Order of the Phoenix Harry Potter and the Half-Blood Prince Harry Potter and the Deathly Hallows
	Harry Potter and the Goblet of Fire	**Book Description from Goodreads:** Harry Potter is midway through his training as a wizard and his coming of age. Harry wants to get away from the pernicious Dursleys and go to the International Quidditch Cup. He wants to find out about the mysterious event that's supposed to take place at Hogwarts this year, an event involving two other rival schools of magic, and a competition that hasn't happened for a hundred years. He wants to be a normal, fourteen-year-old wizard. But unfortunately for Harry Potter, he's not normal - even by wizarding standards. And in his case, different can be deadly.

Goodreads	Off to be the Wizard	**Book Description from Goodreads:** Martin Banks is just a normal guy who has made an abnormal discovery: he can manipulate reality, thanks to reality being nothing more than a computer program. With every use of this ability, though, Martin finds his little “tweaks” have not escaped notice. Rather than face prosecution, he decides instead to travel back in time to the Middle Ages and pose as a wizard. What could possibly go wrong? An American hacker in King Arthur’s court, Martin must now train to become a full-fledged master of his powers, discover the truth behind the ancient wizard Merlin… and not, y’know, die or anything.
	Wizard at Large	**Book Description from Goodreads:** It all began when the half-able wizard Questor Thews announced that finally he could restore the Court Scribe Abernathy to human form. All went well until the wizard breahed the magic dust of his spell and suddenly sneezed. Then, where Abernathy had stood, there was only a bottle containing a particularly evil imp, who soon escapes.
	Winter of the Ice Wizard	**Book Description from Amazon:** JACK AND ANNIE, joined by Teddy and Kathleen (from earlier books), travel in the Magic Tree House to a land of snow where the Ice Wizard has captured Morgan and Merlin. The four friends must find the Ice Wizard’s missing eye . . . or is it really his heart that is missing?
Amazon	Wizarding World: Hidden Creatures Scratch Magic	**Book Description from Amazon:** This Wizarding World scratch art book is full of activities to unleash your imagination! Reveal what beasts Newt Scamander keeps in his suitcase, draw what a dragon could be guarding inside your own Gringotts vault, and find the Basilisk that is slithering around Hogwarts undetected! Including magical creatures from both the Harry Potter and Fantastic Beasts films.
	J.K. Rowling's Wizarding World: Movie Magic Volume One: Extraordinary People and Fascinating Places: 1	**Book Description from Amazon:** Featuring all eight Harry Potter movies and the upcoming movie Fantastic Beasts and Where to Find Them, this magical book is the ultimate insider's guide to the films from J.K. Rowling's Wizarding World for young fans. From the gilded halls of Gringotts and Hogwarts to the New York City of Fantastic Beasts and Where to Find Them, each page of this book delivers a fun, interactive experience for young readers as they discover how the extraordinary places and fascinating characters of the wizarding world took shape onscreen. Filled with lift-the-flaps, stickers, and other engaging inserts, this engrossing book overflows with captivating facts about the movie magic used to create a world fit for witches and wizards. Including insights from the actors who played Harry Potter, Professor Dumbledore, Newt Scamander, and many more, this book is a must-have for young fans of the Wizarding World.
	Wizarding for Beginners	**Book Description from Amazon:** Best friends Dave (now a knighted dragon) and Albrecht (Dave's German-speaking, trusty steed, life coach, and a goat) from Knighthood for Beginners are back--and they're going undercover! They must disguise themselves as wizards to enter the notoriously secretive Wizarding Guild, in order to free their kidnapped, talking-animal friends and stop Terrence, the most evil wizard of them all. Luckily, they have the perfect book to help them on their quest, the amazing, the brilliant, Wizarding for Beginners! Copious black-and-white illustrations by the author help bring all the hilarity to life in this eagerly anticipated follow-up to Knighthood for Beginners.

OpenLibrary	The Complete Idiot's Guide to the World of Harry Potter	**Book Description from Amazon:** An entertaining overview of the spellbinding series of fantasy novels by J. K. Rowling covers all aspects of the wizards' world, furnishes detailed facts about all seven books in the Harry Potter series, and examines the books in terms of their historical, literary, religious, scientific, and mythological roots. Original.
	The Plot Thickens... Harry Potter Investigated by Fans for Fans	**Book Description from Amazon:** The Ultimate Unofficial Guide to Harry Potter broke new ground as a book by fans for fans. Now, Wizarding World Press takes that one step further by collecting the best pieces from 50 budding authors, ages 7 to 47 -- and representing 9 countries -- from the most popular Harry Potter fan site, MuggleNet.com. Sure to spark lively debate among Harry Potter sleuths and fans, these essays are as entertaining as they are thought-provoking.
	Mugglenet.com's Harry Potter should have died: controversial views from the #1 fan site	**Book Description from Amazon:** K. Rowling not only told a wonderful story of a boy wizard; she also created an endless' magical world filled with millions of what - ifs. Whether the reader is a ten - year - old in the school lunchroom or an adult posting on one of the countless Potter fansites' debating what ifs is the love of millions of Harry Potter fans. Now Mugglenet.com's Unofficial' Unauthorized and Unequaled Harry Potter Debates brings the original and entertaining views of the experts behind Mugglenet.com to 100 of the most interesting and contentious debate topics in the Potter world. Digging into every reference in all seven books' the authors apply the same unmatched insight that allowed them to be so incredibly accurate with their predictions in Mugglenet.com's What Will Happen in Harry Potter 7. Readers find plenty to ponder as they consider whether Snape or Dumbledore is a better wizard' whether it would be better to own Harry's invisibility cloak or his flying broomstick' and many more magical conundrums.

Table B-2: Search Results for “Dystopian end of world”

Book Search Engine	Title	Description
Contextual Search Engine	Nineteen Eighty-Four	**Book Description from Goodreads:** It is 1984. The world is in a state of perpetual war and Big Brother sees and controls all. Winston Smith, a member of the Outer Party and propaganda-writer at the Ministry of Truth, is keeping a journal he should not be keeping and falling in love with Julia, a woman he should not be seeing. Outwardly compliant, Winston dreams of rebellion against the oppressive Big Brother, risking everything to recover his lost sense of individuality and control of his own future. One of the bestselling books of the twentieth century, 1984 is the dystopian classic that introduced such Orwellian terms as ‘Big Brother,’ ‘doublethink,’ ‘Newspeak,’ and ‘thoughtcrime’ to the collective consciousness, giving official terminology to state-sanctioned deception, surveillance, and historical revisionism.
	Allegiant	**Book Description from Goodreads:** The faction-based society that Tris Prior once believed in is shattered - fractured by violence and power struggles and scarred by loss and betrayal. So when offered a chance to explore the world past the limits she's known, Tris is ready. Perhaps beyond the fence, she and Tobias will find a simple new life together, free from complicated lies, tangled loyalties, and painful memories.

		But Tris's new reality is even more alarming than the one she left behind. Old discoveries are quickly rendered meaningliess. Explosive new truths change the hearts of those she loves. And once again, Tris must battle to comprehend to complexities of human nature - and of herself - while facing impossible choices about courage, allegiance, sacrifice, and love. Told from a riveting dual perspective, ALLEGIANT, by #1 New York Times best-selling author Veronica Roth, brings the DIVERGENT series to a powerful conclusion while revealing the secrets of the dystopian world that has captivated millions of readers in DIVERGENT and INSURGENT.
	Divergent	**Book Description from Goodreads:** In Beatrice Prior's dystopian Chicago, society is divided into five factions, each dedicated to the cultivation of a particular virtue--Candor (the honest), Abnegation (the selfless), Dauntless (the brave), Amity (the peaceful), and Erudite (the intelligent). On an appointed day of every year, all sixteen-year-olds must select the faction to which they will devote the rest of their lives. For Beatrice, the decision is between staying with her family and being who she really is--she can't have both. So she makes a choice that surprises everyone, including herself. During the highly competitive initiation that follows, Beatrice renames herself Tris and struggles to determine who her friends really are--and where, exactly, a romance with a sometimes fascinating, sometimes infuriating boy fits into the life she's chosen. But Tris also has a secret, one she's kept hidden from everyone because she's been warned it can mean death. And as she discovers a growing conflict that threatens to unravel her seemingly perfect society, she also learns that her secret might help her save those she loves . . . or it might destroy her.
Goodreads	Babylon Working: The End Of the World is just a Beginning (book1)	**Book Description from Goodreads:** THE END OF THE WORLD IS JUST A BEGINNING…The year is 2066. The world is embroiled in a perpetual war. The authoritarian Union, a dictatorship that spans from America to the borders of Russia, uses the “Global Defence Authority” army as a replacement for the old United Nations, sending endless lines of troops, male and female, to fight The African Islamic League, a quasi-Caliphate that spans much of Africa and the middle East, known commonly as “The Terrorists”. The Frontline extends from China to the edges of Europe. Britain is run by neo-Fascist party the B.F.P. and is under the firm control of aging Prime Minster Mark Collins. In London society is downtrodden and lost. Naive seventeen year old Aaron Styles is thinking about his future and has decided to join the Global Defence Authority forces, even though he is a black kid in a society where having anything but white skin severely impacts your life choices. Ghettos, known as Estates, are the common symbol of racial divide, but Aaron has been brought up in secrecy away from the notorious ghetto by elderly guardian Doctor Andrew Forrester, an 82 year old middle class white historian. The Doctor’s life has been one of regrets and inaction against the murderous political thugs that have taken society over, and now he is left devastated by Aaron’s decision to leave. Finally, together with his life long friend and ex-resistance fighter Shirley Barnes, he decides he’s going to do something about it. Meanwhile, on the front line in the Egyptian desert something very strange is happening. Twenty-three year old Californian Jake Kochowski, a Corporal in the Global Defence Authority Marine infantry, finds himself at the heart of a horrifying supernatural event, involving deities, demons and monsters, that will change the world, and its warring occupants, forever. Using the distinct voices of Aaron Styles, Doctor Forrester and Corporal Kochowski, “Babylon Working” is a dystopian dark fantasy sci-fi horror, with a contemporary take on themes explored by works such as Orwell and Heller with the horror tones of H. P. Lovecraft and a strong dose of the postmodern visionary genius of George A. Romero and Robert Kirkman.
	Babylon Working – Part II: A Dystopian Sci-Fi Fantasy Horror	
	Babylon Working: A Dystopian Sci-Fi Horror	

		With elements of apocalyptic horror and speculative science fiction, and an underlying threat of Lovecraftian monsters and strange deities, Babylon Working is a unique story in three parts, set in a world where the armor of democracy and free expression has been destroyed and removed. Enjoy Parts 1, 2 and 3: find them in the Kindle store today!
Amazon	The End of the World	**Book Description from Amazon:** Created during sleepless nights while he worked on his animated films, The End of the World was illustrated entirely on Post-It notes over the course of several years, slowly taking shape from all the deleted scenes, bad dreams, and abandoned ideas that were too strange to make it to the big screen, including essential early material that was later developed into the animated classic World of Tomorrow. Hertzfeldt's visually striking work transcends its unusual nature and taps into the deeply human, universal themes of mortality, identity, memory, loss, and parenthood . . . with the occasional monstrous biting eel descending from the sky.
	The End: A Postapocalyptic Novel	**Book Description from Amazon:** Young Gordon Van Zandt valued duty and loyalty to country above all, so after 9/11, he dropped out of college and joined the Marine Corps. This idealism vanished one fateful day in a war-torn city in Iraq. Ten years later, he is still struggling with the ghosts of his past when a new reality is thrust upon him and his family: North America, Europe and the Far East have all suffered a devastating Super-EMP attack, which causes catastrophic damage to the nation's power grid and essential infrastructures. Everything from cell phones to cars to computers cease to function, putting society at a standstill. With civilization in chaos, Gordon must fight for the limited and fast dwindling resources. He knows survival requires action and cooperation with his neighbors, but as the days wear on, so does all sense of civility within his community--and so he must make some of the most difficult decisions of his life in order to ensure his family's safety.
	The End of the World Running Club	**Book Description from Goodreads:** Perfect for fans of The Martian, this powerful post-apocalyptic thriller pits reluctant father Edgar Hill in a race against time to get back to his wife and children. When the sky begins to fall and he finds himself alone, his best hope is to run – or risk losing what he loves forever. When the world ends and you find yourself forsaken, every second counts. No one knows this more than Edgar Hill. Stranded on the other side of the country from his wife and children, Ed must push himself across a devastated wasteland to get back to them. With the clock ticking and hundreds of miles between them, his best hope is to run -- or risk losing what he loves forever.
OpenLibrary	Ten to one: selected poems	**Book Description from Amazon:** Bob Perelman's poetry is alternately lyrical, polemical, funny, self-critical and didactic, and consistently engaging. In this, his first selected volume, the author of The Marginalization of Poetry delivers a stunning poetic portrait of late 20th-century America. This volume includes work from all of Perelman's previously published collections.
	The great inversion and the future of the American city	**Book Description from Amazon:** Eye-opening and thoroughly engaging, this is an indispensible look at American urban/suburban society and its future. In The Great Inversion, Alan Ehrenhalt, one of our leading urbanologists, reveals how the roles of America's cities and suburbs are changing places--young adults and affluent retirees moving in, while immigrants and the less affluent are moving out--and addresses the implications of these shifts for the future of our society.

		Ehrenhalt shows us how the commercial canyons of lower Manhattan are becoming residential neighborhoods, and how mass transit has revitalized inner-city communities in Chicago and Brooklyn. He explains why car-dominated cities like Phoenix and Charlotte have sought to build twenty-first-century downtowns from scratch, while sprawling postwar suburbs are seeking to attract young people with their own form of urbanized experience.
	Killer Critique	**Description from Goodreads:** Now a seasoned member of the Paris Police Judiciaire, the ever fashionable and whip smart Commissaire Capucine Le Tellier is finding murder to be more than just the specialite du jour. . . When the senior food critic for Le Figaro is found face-first in a plate of Ravioles d'horand, there seem to be as many suspects as there are restaurants in the City of Light. Yet Capucine feels she'll solve the case quicker than it takes to serve up an omelet aux fines herbes. Un problem, murders of food critics have become an epidemic. As the bodies pile up, un, deux, trois, so do the suspects, including a sexy starlet, an award-winning novelist, and a smorgasbord of aggravated chefs. While Capucine struggles to zero in on the murderer's tastes, she is confronted with a false dilemma: file and forget the case, leaving restaurant critics across France vulnerable to a killer's episodic cravings, or use her husband, Alexandre, himself a famous food journalist, as irresistible bait. Filled with delectable intrigue and ripe with quirky suspects, with a dash of Freudian idees de grandeur, "Killer Critique " is a feast worth killing for. . ."

Appendix C – Genre Prediction Models Results

Table C-1: Stage 1 (Pre-tuned models) Results

Model			% of books with at least 60% of genres matched (e.g. 3 out of 5)	
	Base Classifier	Hyper Parameters	LDA	Doc2Vec
Baseline	NA	NA	15.1%^	
Classifier Chains	GaussianNB	Default	15.6%	34.8%
Label Powerset	GaussianNB	Default	14.3%	35.6%
Adapted Algorithm	Mlknn	k=20	14.7%	13.2%
Ensemble (rakeld)	GaussianNB	Labelset_size = 10	12.5%	24.1%
Neural Network	NA	3 Dense layers with 10% dropout activation = Selu Final activation = sigmoid Optimiser = Adamax	45.6%	50.0%

^ not based on either LDA or Doc2Vec

Table C-2: Stage 2 (Post-tuned models) Results

Model			% of books with at least 60% of genres matched (e.g. 3 out of 5)
	Base Classifier	Hyper Parameters	Doc2Vec
Classifier Chains	Stochastic Gradient Descent	Loss = modified_huber Max_iter = 10,000	45.5%
	Stochastic Gradient Descent	Loss = Log Max_iter = 10,000	45.3%
	Logistic Regression	Max_iter = 10,000	46.4%
Ensemble (RakelD)	Stochastic Gradient Descent	Labelset_size = 10 Loss = modified_huber	48.2%
	Stochastic Gradient Descent	Labelset_size = 10 Loss = Log	51.8%
	Random Forest	Labelset_size = 10 n_estimators = 200	49.4%
	Random Forest	Labelset_size = 10 n_estimators = 200 criterion = entropy	49.2%
	Random Forest	Labelset_size = 10 n_estimators = 300 criterion = entropy	50.6%
	Logistic Regression	Labelset_size = 10 max_iter = 10,000	53.1%
	Logistic Regression	Labelset_size = 10 max_iter = 10,000 Solver = Newton-cg	53.1%
Neural Network	8 dense layer (deep) with dropout	8 Dense layers with 10% dropout activation = Relu Final activation = sigmoid Optimiser = Adamax Final activation = sigmoid Optimiser = Adam	49.1%
	CNN with word embeddings	Embedding size = 500 Filter sizes = [2,3,4] Number of filters = 30	66.8%
Neural Network	**CNN with word embeddings**	**Embedding size = 500** **Filter sizes = [2,3,4,5,6]** **Number of filters = 200**	**72.4%**