INST414
Team members - Ben Griffith, Daniel Gonzalez, Prince Okpoziakpo
03/17/2023

Project - 2 Data Collection Report

How did you obtain your data?

So far, our data is coming from two main sources. We are sourcing our list of books from a large Kaggle dataset found by Daniel, then using the Google Books API to cross reference the books and obtain further data which we will be using to conduct similarity tests between the books. We are planning to calculate similarity using potential variables such as book description, genre, author, and ratings. The Kaggle dataset contains ~11,000 rows of book data, but many of the records in this dataset have missing fields. This is where the Google Books API comes into play. Through the API, we were able to iterate over the dataset, take each ISBN, then search the API by ISBN to obtain more data about the books.

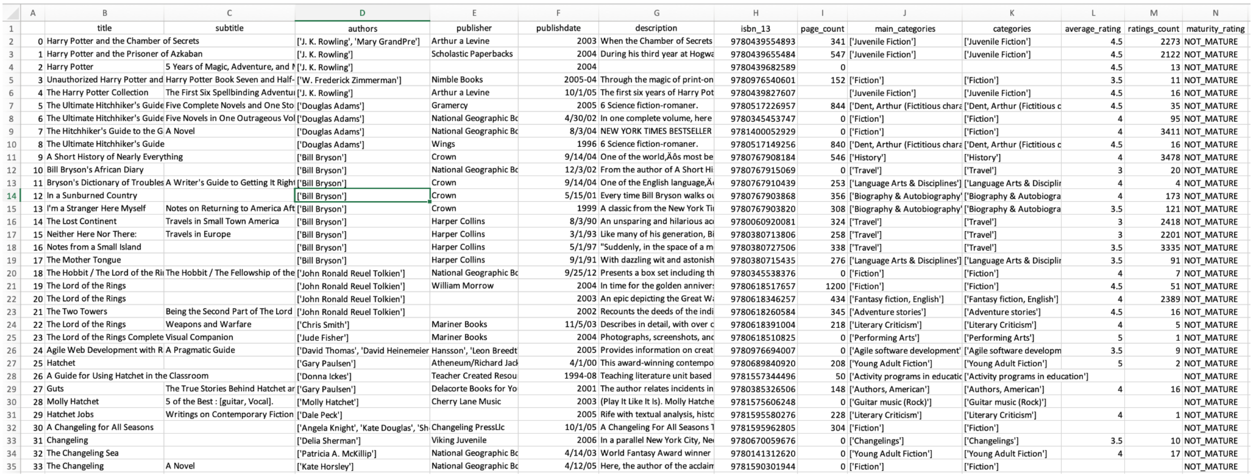
A challenge we have run into along with absent data is confusion with cross referencing the ISBN numbers between data sources. The issue with this process was that certain ISBNs were returning empty responses from the Google Books API. We are in the process of determining how to iterate through the results and remove these items so that we only store books that exist in both the Kaggle dataset and the Google Books API. Another issue is that certain ISBNs are returning multiple values. This is likely a result of duplicate entries in the Kaggle dataset for some books. This will require us to parse through the results and remove any duplicate books from the final dataset.

How large is your data?

As aforementioned, the Kaggle dataset from which we are basing our initial search contains ~11,000 rows. However, we found that only 25% of the Kaggle dataset books exist in the Google Books API; after filtering out duplicates, and removing records that existed in the Kaggle dataset but not in the Google Books API database, we have 2,670 books in our records. We understand that this could be a result of using the ISBN-13 number for each book. We plan on making another round of API calls to test if we get more books by searching by the ISBN-10 number, then merge unique results. Meanwhile, we will only be using the data set of books we have right now for our analysis in order to make progress in our analysis, and continue to collect high quality data as we move forward. Our main CSV contains a row for each book with ~12 attributes. We plan to have other data frames as well for our matrices for similarity analysis. Each of these will contain a row for each book. Considering they do not yet exist, we do not know the storage size for each. At the moment, all of our current CSV files take up ~3.5 MB of storage.

In what format(s) are you storing your data? Describe the abstract data types, not just the file format.

As mentioned above, we will be using similarity for our analysis as well as graphs. We are planning to possibly use Jaccard similarity with genres to get an idea of which books have the most similar category classifications. For this analysis our data will be organized into a matrix, with the ISBN-10/13 number as the index, and different features, i.e., Description, and Genres, as the columns. Our analysis will consist of several layers, so we will use different abstract data types required for each analysis. For example, we will be using Jaccard Similarity to the similarity between books, relative to their genre. For this layer of analysis, we will use Sets to get the intersection between book genres, and determine other membership qualities. In future layers of analysis, we will use graphs to model the relationships between several books by generating weighted edges between books. The weight of these edges will depend on the variables that we determine are high indicators of correlation between books. Once the graph has been modeled, we will create a visualization of the graph.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		title	subtitle	authors	publisher	publishdate	description	isbn_13	page_count	main_categories	categories	average_rating	ratings_count	maturity_rating
2	0	Harry Potter and the Chamber of Secrets		['J. K. Rowling', 'Mary GrandPre']	Arthur a Levine	2003	When the Chamber of Secrets	9780439554893	341	['Juvenile Fiction']	['Juvenile Fiction']	4.5	2273	NOT_MATURE
3	1	Harry Potter and the Prisoner of Azkaban		['J. K. Rowling']	Scholastic Paperbacks	2004	During his third year at Hogwa	9780439655484	547	['Juvenile Fiction']	['Juvenile Fiction']	4.5	2122	NOT_MATURE
4	2	Harry Potter	5 Years of Magic, Adventure, and M	['J. K. Rowling']		2004		9780439682589	0			4.5	13	NOT_MATURE
5	3	Unauthorized Harry Potter and	Harry Potter Book Seven and Half-	['W. Frederick Zimmerman']	Nimble Books	2005-04	Through the magic of print-on	9780976540601	152	['Fiction']	['Fiction']	3.5	11	NOT_MATURE
6	4	The Harry Potter Collection	The First Six Spellbinding Adventu	['J. K. Rowling']	Arthur a Levine	10/1/05	The first six years of Harry Pot	9780439827607		['Juvenile Fiction']	['Juvenile Fiction']	4.5	16	NOT_MATURE
7	5	The Ultimate Hitchhiker's Guide	Five Complete Novels and One Sto	['Douglas Adams']	Gramercy	2005	6 Science fiction-romaner.	9780517226957	844	['Dent, Arthur (Fictitious chara	['Dent, Arthur (Fictitious c	4.5	35	NOT_MATURE
8	6	The Ultimate Hitchhiker's Guide	Five Novels in One Outrageous Vol	['Douglas Adams']	National Geographic Bc	4/30/02	In one complete volume, here	9780345453747	0	['Fiction']	['Fiction']	4	95	NOT_MATURE
9	7	The Hitchhiker's Guide to the G	A Novel	['Douglas Adams']	National Geographic Bc	8/3/04	NEW YORK TIMES BESTSELLER	9781400052929	0	['Fiction']	['Fiction']	4	3411	NOT_MATURE
10	8	The Ultimate Hitchhiker's Guide		['Douglas Adams']	Wings	1996	6 Science fiction-romaner.	9780517149256	840	['Dent, Arthur (Fictitious chara	['Dent, Arthur (Fictitious c	4.5	16	NOT_MATURE
11	9	A Short History of Nearly Everything		['Bill Bryson']	Crown	9/14/04	One of the world,Äôs most be	9780767908184	546	['History']	['History']	4	3478	NOT_MATURE
12	10	Bill Bryson's African Diary		['Bill Bryson']	National Geographic Bc	12/3/02	From the author of A Short Hi	9780767915069	0	['Travel']	['Travel']	3	20	NOT_MATURE
13	11	Bryson's Dictionary of Troubles	A Writer's Guide to Getting It Right	['Bill Bryson']	Crown	9/14/04	One of the English language,Ä	9780767910439	253	['Language Arts & Disciplines']	['Language Arts & Disciplir	4	4	NOT_MATURE
14	12	In a Sunburned Country		['Bill Bryson']	Crown	5/15/01	Every time Bill Bryson walks ou	9780767903868	356	['Biography & Autobiography'	['Biography & Autobiogra	4	173	NOT_MATURE
15	13	I'm a Stranger Here Myself	Notes on Returning to America Aft	['Bill Bryson']	Crown	1999	A classic from the New York Ti	9780767903820	308	['Biography & Autobiography'	['Biography & Autobiogra	3.5	121	NOT_MATURE
16	14	The Lost Continent	Travels in Small Town America	['Bill Bryson']	Harper Collins	8/3/90	An unsparing and hilarious ac	9780060920081	324	['Travel']	['Travel']	3	2418	NOT_MATURE
17	15	Neither Here Nor There:	Travels in Europe	['Bill Bryson']	Harper Collins	3/1/93	Like many of his generation, Bi	9780380713806	258	['Travel']	['Travel']	3	2201	NOT_MATURE
18	16	Notes from a Small Island		['Bill Bryson']	Harper Collins	5/1/97	"Suddenly, in the space of a m	9780380727506	338	['Travel']	['Travel']	3.5	3335	NOT_MATURE
19	17	The Mother Tongue		['Bill Bryson']	Harper Collins	9/1/91	With dazzling wit and astonish	9780380715435	276	['Language Arts & Disciplines']	['Language Arts & Disciplir	3.5	91	NOT_MATURE
20	18	The Hobbit / The Lord of the Ri	The Hobbit / The Fellowship of the	['John Ronald Reuel Tolkien']	National Geographic Bc	9/25/12	Presents a box set including th	9780345538376	0	['Fiction']	['Fiction']	4	7	NOT_MATURE
21	19	The Lord of the Rings		['John Ronald Reuel Tolkien']	William Morrow	2004	In time for the golden anniver	9780618517657	1200	['Fiction']	['Fiction']	4.5	51	NOT_MATURE
22	20	The Lord of the Rings		['John Ronald Reuel Tolkien']		2003	An epic depicting the Great Wa	9780618346257	434	['Fantasy fiction, English']	['Fantasy fiction, English']	4	2389	NOT_MATURE
23	21	The Two Towers	Being the Second Part of The Lord	['John Ronald Reuel Tolkien']		2002	Recounts the deeds of the indi	9780618260584	345	['Adventure stories']	['Adventure stories']	4.5	16	NOT_MATURE
24	22	The Lord of the Rings	Weapons and Warfare	['Chris Smith']	Mariner Books	11/5/03	Describes in detail, with over c	9780618391004	218	['Literary Criticism']	['Literary Criticism']	4	5	NOT_MATURE
25	23	The Lord of the Rings Complete	Visual Companion	['Jude Fisher']	Mariner Books	2004	Photographs, screenshots, anc	9780618510825	0	['Performing Arts']	['Performing Arts']	5	1	NOT_MATURE
26	24	Agile Web Development with R	A Pragmatic Guide	['David Thomas', 'David Heinemeier Hansson', 'Leon Breedt'		2005	Provides information on creat	9780976694007	0	['Agile software development'	['Agile software developm	3.5	9	NOT_MATURE
27	25	Hatchet		['Gary Paulsen']	Atheneum/Richard Jack	4/1/00	This award-winning contempo	9780689840920	208	['Young Adult Fiction']	['Young Adult Fiction']	5	2	NOT_MATURE
28	26	A Guide for Using Hatchet in the Classroom		['Donna Ickes']	Teacher Created Resou	1994-08	Teaching literature unit based	9781557344496	50	['Activity programs in educatic	['Activity programs in education']			NOT_MATURE
29	27	Guts	The True Stories Behind Hatchet ar	['Gary Paulsen']	Delacorte Books for Yo	2001	The author relates incidents in	9780385326506	148	['Authors, American']	['Authors, American']	4	16	NOT_MATURE
30	28	Molly Hatchet	5 of the Best : [guitar, Vocal].	['Molly Hatchet']	Cherry Lane Music	2003	(Play It Like It Is). Molly Hatche	9781575606248	0	['Guitar music (Rock)']	['Guitar music (Rock)']			NOT_MATURE
31	29	Hatchet Jobs	Writings on Contemporary Fiction	['Dale Peck']		2005	Rife with textual analysis, histc	9781595580276	228	['Literary Criticism']	['Literary Criticism']	4	1	NOT_MATURE
32	30	A Changeling for All Seasons		['Angela Knight', 'Kate Douglas', 'Sh	Changeling PressLlc	10/1/05	A Changeling For All Seasons T	9781595962805	304	['Fiction']	['Fiction']			NOT_MATURE
33	31	Changeling		['Delia Sherman']	Viking Juvenile	2006	In a parallel New York City, Ne	9780670059676	0	['Changelings']	['Changelings']	3.5	10	NOT_MATURE
34	32	The Changeling Sea		['Patricia A. McKillip']	National Geographic Bc	4/14/03	World Fantasy Award winner	9780141312620	0	['Young Adult Fiction']	['Young Adult Fiction']	4	17	NOT_MATURE
35	33	The Changeling	A Novel	['Kate Horsley']	National Geographic Bc	4/12/05	Here, the author of the acclaim	9781590301944	0	['Fiction']	['Fiction']			NOT_MATURE

Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?

We have conducted some data cleaning in order to convert our API responses into json and eventually CSV, but these CSV files have not shown themselves to be overwhelmingly large in terms of storage. The 11,000 rows hasn't been difficult to process, given our current computational constraints, so we do not plan to convert our data into further compressed formats. Eventually if we plan to graph our networks in Gephi or related software, we may need to consider the size of the resulting files and how to properly store them to keep them accessible. Additionally, if we plan to use applications like Gephi we will have to consider converting our files to file types like json to be able to import the data correctly.