**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.

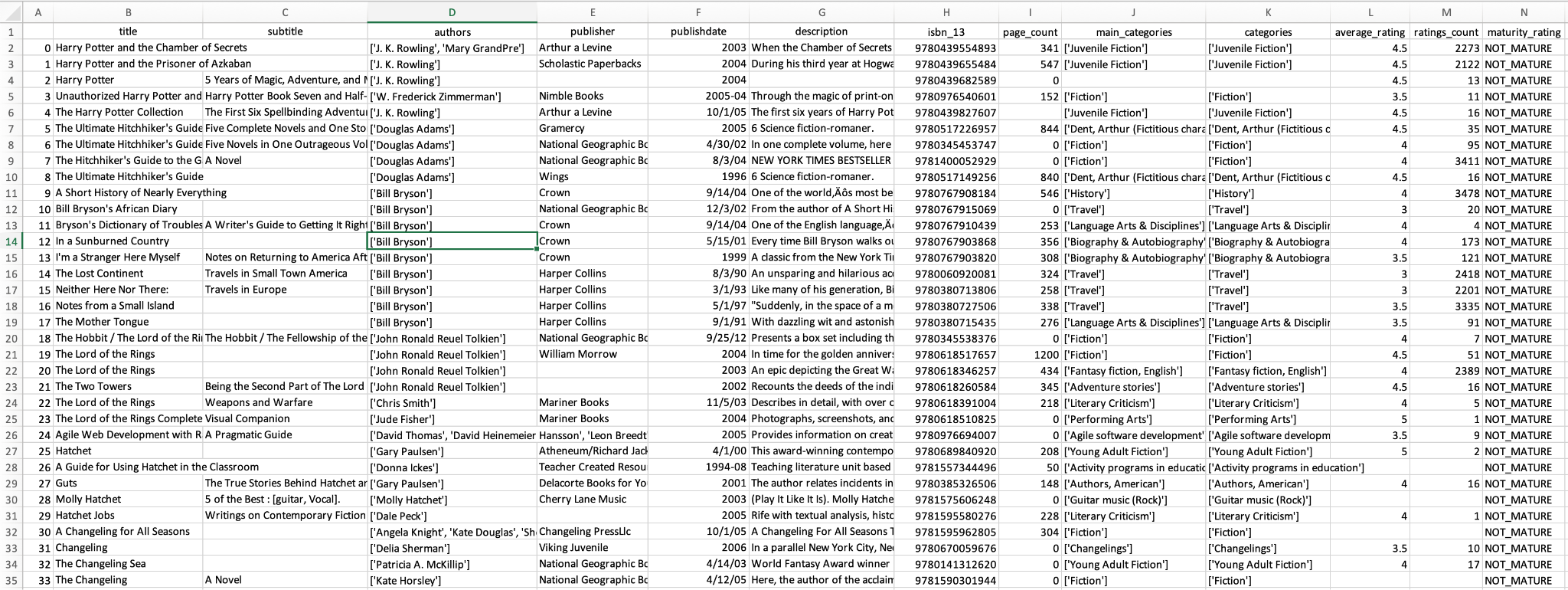
A challenge we have run into along with absent data is confusion with cross referencing the ISBN numbers between data sources. The issue with this process was that certain ISBNs were returning empty responses from the Google Books API. We are in the process of determining how to iterate through the results 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.



**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.