The goal of my project was to scrape all of the 5-star Amazon.com reviews for all of the books in Lemony Snicket's *A Series of Unfortunate Events* and to perform a frequency count of all of the words used in these reviews in order to determine why people like these books. The first step of my project involved web scraping the 5-star reviews of the books from Amazon.com. I thought I would be able to obtain all of the reviews that I wanted using "wget", but this method proved unsuccessful. Instead, I was able to find a Google Chrome browser extension that was simply called "Web Scraper", and I followed the instructions at https://www.scrapehero.com/amazon-review-scraper/ to scrape the reviews that I desired. Once "Web Scraper" has been set up, it just needs to know the first page of reviews to be scraped, and it finds all of the following pages of reviews and scrapes them as well. After scraping the reviews, I saved the reviews for each book in a separate .csv file; for example, all of the reviews for the first book were saved to book1.csv.

For the next part of my project, I planned on cutting out the fields of the .csv files that contained the actual reviews and then using "egrep" to obtain word lists for reviews from each book. However, many of the reviews had commas within them, so cutting out fields using a comma as a delimiter did not work as well as I had assumed that it would. Thus, I had to save the .csv files in Excel in a way that replaced the comma with some other character as a delimiter; I chose to use the pipe character as my new delimiter, and I did the delimiter replacement using the instructions here: https://www.howtogeek.com/howto/21456/export-or-save-excel-files-with-pipe-or-other-delimiters-instead-of-commas/. The fields of the .csv files that contained the actual reviews were field 6 of each file, which was labeled "content." I wanted to cut out field 6 of each pipe-delimited .csv file, to remove the line only containing "content", and to obtain a word list for each file using "egrep"; these steps were saved to the shell script word_list.sh. When generating word lists, I set the minimum word length to four letters in order to remove very common English words that were not very relevant to the books. Since I wanted to obtain frequency counts, I made another shell script called word_count.sh that simply performs the classic frequency count idiom on any file that it is given.

Then it was time to create a makefile. My ultimate goal was to make a frequency count of all of the words (with four or more letters) from all of the reviews across all of the books, but I also wanted to be able to make frequency counts of the words of the reviews from each book. Therefore, I decided to make word lists from reviews of each book separately and a word list containing all of the words from all of the reviews across all of the books. For instance, book1_words.txt contains a word list of only words that were used in the reviews for the first book, and all_words.txt contains all of the words from all of the reviews across all of the books. The end goal of the makefile is to generate a frequency count of the words in the word list containing all of the words from all of the reviews across all of the books; this frequency count is contained in all_words_count.txt. I also set up options within the makefile in order to generate the frequency counts of the words in the word lists of the reviews from each book; *e.g.*, book1_words_count.txt is a frequency count of the words in the word list book1_words.txt.

With all of the word frequency counts generated, all that was left to do was to analyze the results. The first book had the greatest amount of reviews. In book 1, the most used words are related to books, children, and reading, which makes sense, as this is a children's book. The book 1 reviewers frequently mentioned the main characters, but they seemed to mention Count Olaf and Violet more than Sunny and Klaus. All of the other books in the series did not have quite as many reviews as the first book. In book 2, the most used words revolve around books, series, reading, and love, most likely because this book and all of the other books in the series have

sparked a love of reading in many children. In fact, books 3-13 all had similar results, with words about books, series, reading, and love having some of the highest frequency counts. Hence, it comes as no surprise that these same words are all near the top of the list in all_words_count.txt; even Count Olaf, the most discussed character, is not referred to as often as these words. It seems that the main reason people enjoy the books in this series is that these books cause kids to love reading.