**Group 10 Preliminary Data Review: Harry Potter Collection Review**

**Group Members**

Jenna Hendrickson and Ashley Rasmussen.

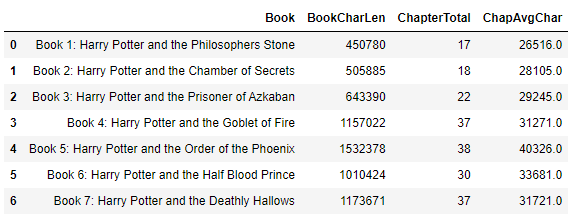
**Data Cleaning**

The seven Harry Potter books were entered into individual data frames. Each data frame contains the book name, chapter title and the content of the chapter. Because chapters are known it was easy to check for accuracy and some code that is very specific to books 5 and 7 were needed to accurately define each chapter. Otherwise a general code was sufficient in most cases.

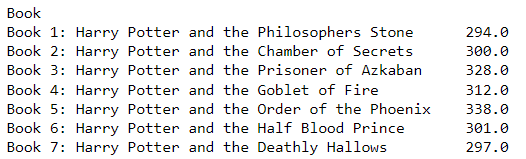
We had hoped that because this data was of well established and edited books that the errors would be minimal. We still feel that they are minimal however we did find a misspelling in a title and a substitution of an “B” for an “E” in another title. We will however not need to worry about emoji’s because they are not present in the books.

Process – extract chapter titles, break down into sentences, find sentence level sentiment and then average sentiment score compared to sentiment of the chapter title.

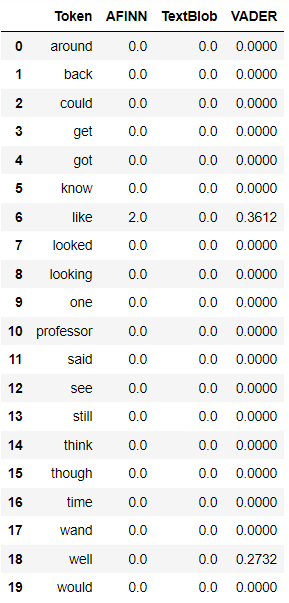
**Summary Statistics**



1. Top ten common words in each book
2. Top ten 3 word phrase in each book
3. Average number of sentences per chapter in each book



1. 20 most frequently used terms (not including character names) across all documents and their sentiment scores:



**Moving Forward**

Moving forward the chapters will be broken down into sentences. A model will be evaluated and chosen and the aggregate sentence sentiment will be determined for each chapter. This will be compared to the sentiments of each chapter to see if there is a general trend. Given enough time the sentiments from a character list obtained in the original data will be evaluated. Then the sentences containing those character names from each book will be aggregated into a general sentiment of the character by book through the series. These will then be compared.