Mini Project 3: Text Mining

Investigating Harry Potter and J. K. Rowling

Samantha Young

## Project Overview

The driving motivation of this literary analysis was to both analyze the Harry Potter series as a whole as well as J.K. Rowlings growth as a writer over time.  To do this I analyzed the sentiment of 3 novels in the 7 book series, the first one the fifth one and the last book. By doing this, I hoped to capture the full arc of emotion in the series. By analyzing word frequency and average word length I hoped to better understand Rowling's changing writing style.

## Implementation

The data of the books I analyzed came from archive.org. This required me to request from websites and use the beautifulsoup package to extract text from HTML pages. I did the majority of my text analysis by looking at each individual word. To do this, I  created a list of string elements that could be easily searched and manipulated. This list contained no punctuation so it could be implemented in several "word based" functions later. In order to find the most frequent and least frequent words in the texts I implemented dictionaries to keep track of the how many times a certain word was in the list. By sorting by value, I could determine the most and least frequent words.  In this assignment, I took a large string, broke it up into a list and analyzed those list elements by associating the contents with dictionary keys.

The design decision to analyze only words longer than 4 letters and that were not included in the title had positives and negatives. By looking at only long words, I removed the common particles of speech like "it, the, an," however I also lost some character names ex. Ron, Cho. I felt like this was a worthy compromise because it allowed the data that I analyzed to be less littered with unimportant words even if I missed a few meaningful ones. I also removed any words that were also in the title of the work, this was to filter for more meaningful words in the novel.

## Results

The Harry Potter Series is a children's series however it was released over the course of 10 years therefore the target audience grew and matured as each new book was released. I inferred that this increasing level of maturity would allow Rowling to write about darker topics as her series progressed. I equated darkness of material to negative words which were analyzed using the Sentiment Analyzer. The prediction that the later books would have more negative content was supported after running the Sentiment Analyzer.

Harry Potter 1 = The Philosopher’s Stone

```
{'compound': 1.0, 'neg': 0.078, 'neu': 0.833, 'pos': 0.089}
```

Harry Potter 5 = The Order of the Phoenix

```
{'compound': 0.9999, 'neg': 0.086, 'neu': 0.826, 'pos': 0.088}
```

Harry Potter 7 = The Deathly Hallows

```
{'compound': -1.0, 'neg': 0.098, 'neu': 0.817, 'pos': 0.085}
```

The values for positive words remained relatively consistent between the books. The negative words shows an apparent shift as the series progresses. a percent increase of 25.64% of negative words  from the first book to the last book.. This data corresponds to the increase number of deaths and disasters that occur in the 7th book as opposed to the first book.

To analyze Rowling as an author I looked at average word length of the works and most frequent/infrequent words used. I inferred that the increase in word length correlated to Rowlings growing experience as a writer, however to fully support this claim more data would need to be analyzed.

The difference in average length of the first and last book is 0.16218050693266495 words.

Book 1:
average word length of Philospher's Stone is
6.473334580838324
Most Frequent:
['Harry', 'Potter', 'Hagrid', 'Hermione', 'about', 'their', 'didn’t', 'could', 'there', 'looked']
Least Frequent:
['Coming', 'offhand', 'Whenever', 'Millicent”', 'tugged', 'concern', 'holes', 'Baruffio', '“Aaah', 'swarmin’']

Book 5:

```
average word length of Order of The Phoenix is
6.7154522555946565
Most Frequent:
['Harry', 'Potter', 'Hermione', 'could', 'their', 'about', 'around', 'looked', 'Professor', 'Sirius']
Least Frequent:
['“Ffine', 'recital', 'Wilfred', 'monument', 'walnut', 'EVANS”', 'mansion', 'overcast',
'parthumans', 'sincerity']

Book 7:
average word length of Deathly Hallows is
6.67025322549586
Most Frequent:
['Harry', 'Hermione', 'Potter', 'could', 'Dumbledore', 'their', 'there', 'would', 'looked', 'about']
Least Frequent:
['siblings', 'recommences', 'judgments', 'floorlength', 'Stripping', 'venture', 'admire',
'“CCrabbe', 'creatures’', 'agreement']
```

## Reflection

Aside from idea selection, the process of which I took to complete this project went well. Unfortunately I could not decide on an idea and wasted numerous hours starting and abandoning projects. Once I decided on a project I went about completing it in a systematic and buildable manner. I started by just loading HTML files and finding data in a web page. Then separately wrote string parsing and text manipulation functions, which I tested using a small sample .txt file before merging the HTML imported text and the functions I wrote. This aided in the unit testing process. To improve, I would have written better doctests and thoroughly investigated a project before starting it. I expended a great deal of time and energy in projects I never completed. The information I learned about pulling data off the internet and reading from web pages will be essential in my next project, interactive data visualization.