**PROJECT SUMMARY : TEXT ANALYSIS OF HARRY POTTER BOOKS**

**COURSE**

DATASCI 350 DS: Methods for Data Analysis, Spring 2016.

**AUTHOR**

Iswarya Murali

**PROJECT OBJECTIVE**

The objective of the project is to perform Textual Analysis of the 7 Harry Potter books to generate a Social Network of the characters and test some hypothesis. (refer to the *Hypothesis Testing and Analysis* section for details)

**DATA SOURCES**

* The [Harry Potter Books](-%09https:/github.com/iswaryam/UWDataScience-Spring2016/tree/master/Books) in text format:
* [Lexicon of positive and negative words](-%09https:/github.com/iswaryam/UWDataScience-Spring2016/tree/master/Lexicon) (taken from <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>)
* [Top Named Characters in Harry Potter](http://harrypotter.answers.wikia.com/wiki/Top_200_most_named_harry_potter_characters_s)
* [List of Spells in Harry Potter](https://en.wikipedia.org/wiki/List_of_spells_in_Harry_Potter)

**DATA PREPARATION AND CLEANING**

1. Read the books in text file format
2. Normalization: Clean Up non-ascii, non-alphabet/non-numeric characters, remove all quotes and other extraneous punctuation except period.
3. Tokenization: Replace alternate nicknames for characters.
4. Remove stop words

**HYPOTHESIS TESTING AND ANALYSIS**

For each of the 7 books, do the following:

1. **Social Network Analysis**:  
   Generate a graph of the top 25 characters, per book. If 2 characters occur within a set threshold (15 non stopwords) of each other, count as 1 edge between them.
2. Analyze the graph object to get the following:
   1. Node with Highest **degree** centrality (most connected character), **eigenvector** centrality and **betweenness** centrality (<https://en.wikipedia.org/wiki/Centrality>).
   2. Mean Degree
   3. Pair of nodes with highest weighted edge (strongest relationship)
   4. Hypothesis Testing: Do the Edge Weights of the graph follow *Power Law fit*?
3. **House Popularity**
   1. Plot mentions of the 4 houses (Gryffindor, Hufflepuff, Ravenclaw and Slytherin), per book
   2. Hypothesis Testing: Are the Houses equally represented per *Chi-Squared Goodness of Fit Test*?
4. **Word Cloud**: Generate a word cloud per book
5. **Spell popularity**: Plot mentions of the occurrence of spells (scraped from Wikipedia)
6. **Sentiment Analysis**: Check occurrences of negative words and positive words (based on the Lexicon), as a progression between book 1 and book 7

All the results (plots/hypothesis tests) are saved as pdf/txt file in the Results folder.

**CODE**

<https://github.com/iswaryam/UWDataScience-Spring2016>

* [HarryPotter-Analysis.R](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/HarryPotter-Analysis.R) : Main Entry Point File. Reads the text of the 7 books and loops through all functions
* [hp\_helper.R](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/hp_helper.R) : Helper file with all functions required to perform analysis.
* [hp\_unit\_tests.R](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/hp_unit_tests.R) : Unit Tests for helper functions.
* [hp\_global\_vars.R](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/hp_global_vars.R) : File that defines some global variables that are used for the analysis.

**RESULTS FILES**

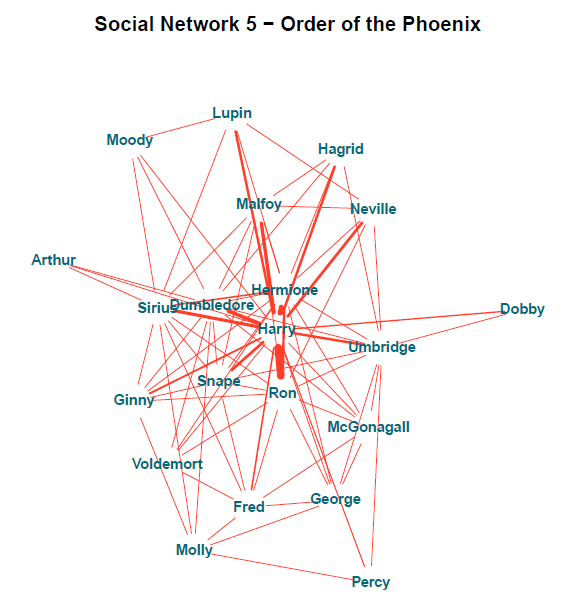
All Results are stored in the "Results" folder.

* [HarryPotter\_Plots.pdf](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/Results/HarryPotter_Plots.pdf)
* [HarryPotter\_GraphAnalysis\_HypothesisTesting.txt](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/Results/HarryPotter_GraphAnalysis_HypothesisTesting.txt)
* Log File: [harrypotterlog.log](https://github.com/iswaryam/UWDataScience-Spring2016/blob/master/Results/harrypotterlog.log)

**SAMPLE RESULTS**

Sample results for one book (ie Book 5). The Results folder has plots/hypothesis tests for all of the 7 books.

**Social Network**



**Graph Analysis Results**

**GRAPH ANALYSIS**

Mean Degree: 7.80952380952381

Highest Degree Centrality: Harry

Strongest Relationship: Harry and Ron

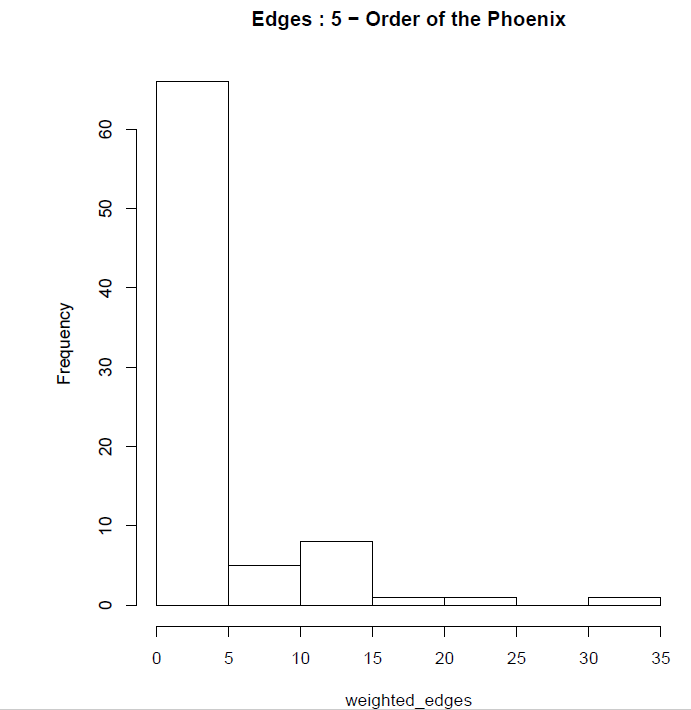
Highest EigenVector Centrality: Harry

Highest Betweennes Centrality: Dumbledore

**Graph Analysis – Hypothesis Testing**

**HYPOTHESIS - DO THE EDGE WEIGHTS FOLLOW POWER LAW?**

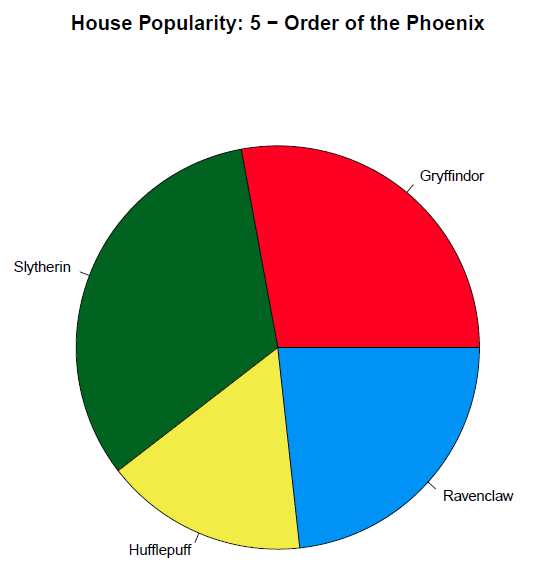
P-Value for Power Law Fit 0.928332635775458. Failed to Reject Null Hypothesis; Edge Weights fit Power Law.



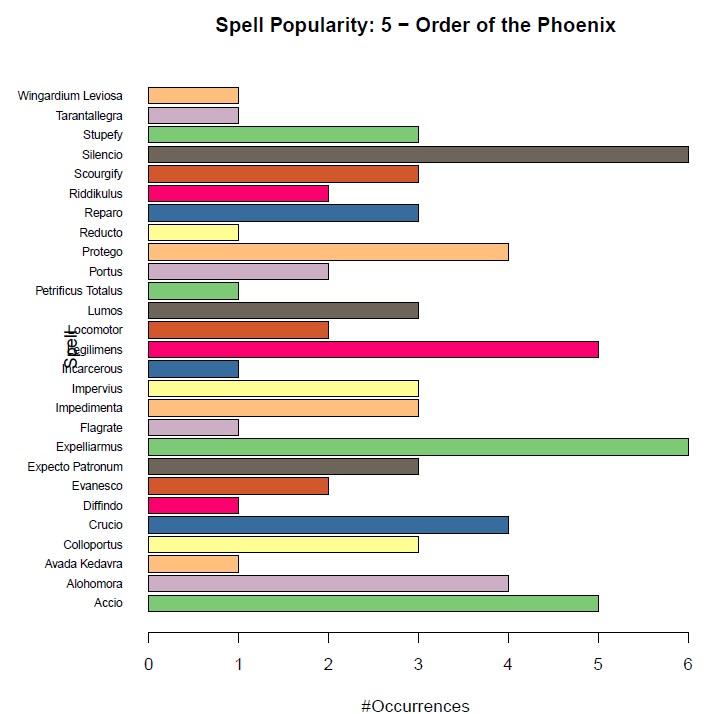
**House Popularity**

**HYPOTHESIS - ARE THE 4 HOUSES REPRESENTED EQUALLY PER CHI SQUARED GOODNESS OF FIT TEST?**

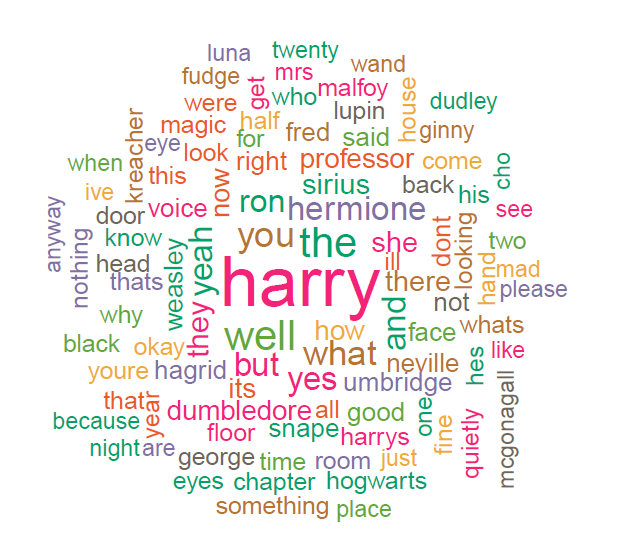
P-Value for Chi Squared Test: 0.477396172245043 . Failed to Reject Null Hypothesis; the Houses are equally represented.



**Spell Popularity**



**Word Cloud**



**Sentiment Analysis**

