Harry Potter Exploratory Analysis in Spark: Ten Years of Magic

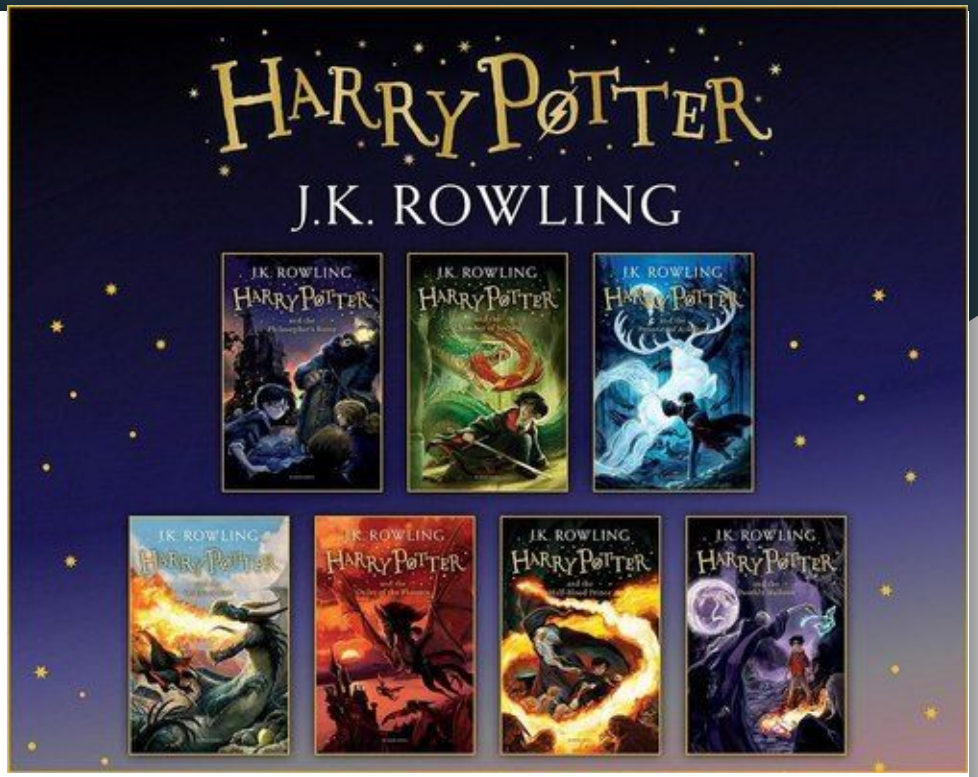
By: Michaella Steinruck

Introduction

- The entire *Harry Potter* series –

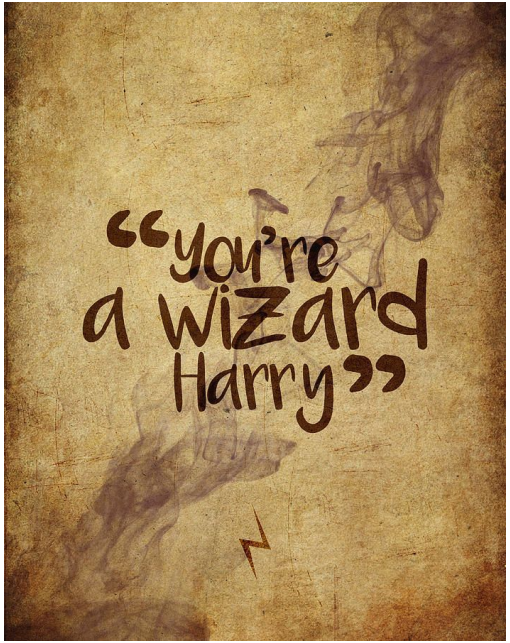
  1,084,170 words

- 7 books published over 10 years
- Over 80 spells
- Over 700 characters
- Interwoven and stand alone plot devices



https://5.imimg.com/data5/SELLER/Default/2020/8/PE/PX/MO/54836353/harry-potter-books-collection-j-k-rowling-bloomsbury-publishing-500x500.jpg

Wizarding World



https://images.fineartamerica.com/images/artworkimages/mediumlarge/1/youre-a-wizard-harry-samuel-whitton.jpg

- Came to an end when the last movie in 2011
- *Cursed Child* is cursed
- JK Rowling is problematic
- A lot happens over the course of the series
- We want more

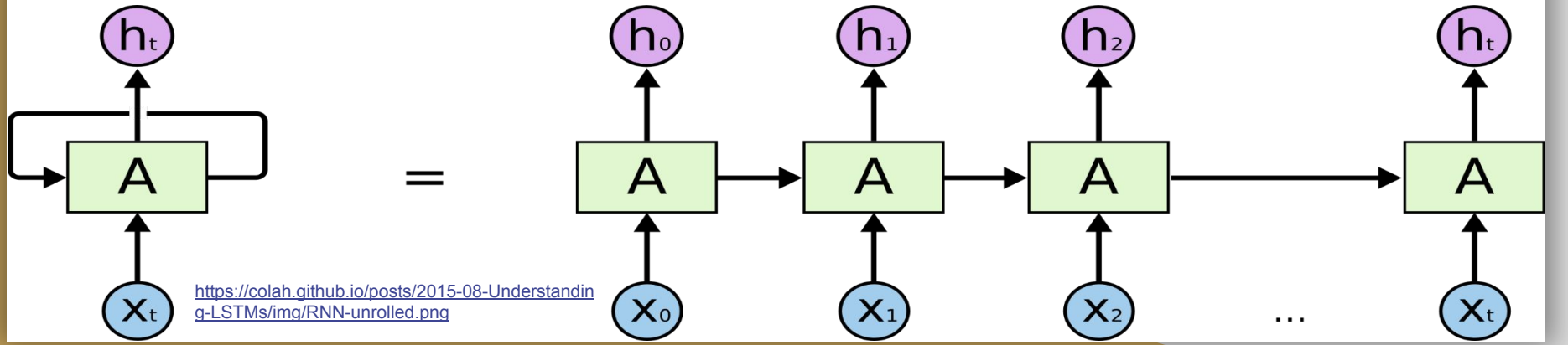
Magic of the Future

- Basic natural language processing to get
  - Character map
  - Spell references
  - Most common words
- Further Analysis
  - Plot summaries for each book

https://external-preview.redd.it/6LRBne5rxCZWCGYrcc61zhIboHAvrKkc0qsS4hXlJA0.jpg?auto=webp&s=d112ab32e0b49b049609f2a4a7813da01b1b96eb

And the Magic Continues

- Advanced machine learning
  - Train a neural network to read the books
  - Write a new novel (maybe not a full one)



https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

The Data

- 7 .txt files containing the contents of the novels pulled from github user
  - https://github.com/formcept/whiteboard
- Split into batches by average words per page
  - ~250 words per page
  - ~4000 rows, each with a book label to identify book and location

How I arrange it might change once I know more about what I need

Sources

- https://blog.fostergrant.co.uk/2017/08/03/word-counts-popular-books-world/#:~:text=Harry%20Potter%20and%20the%20Half,Harry%20Potter%20series%20%E2%80%93%201%2C084%2C170%20words
- https://github.com/formcept/whiteboard/tree/master/nbviewer/notebooks/data/harrypotter