Harry Potter

And the Mystery of NLP

A Text Classification Project By

Jack Roach

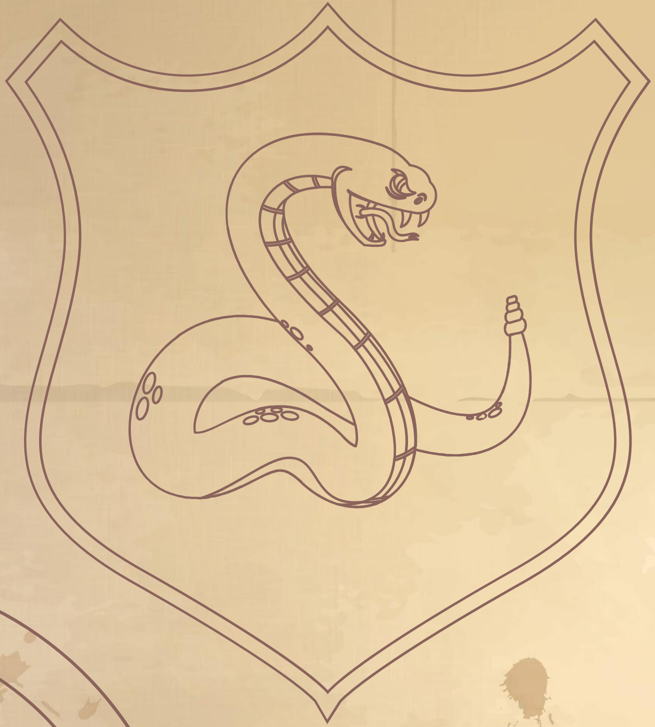
The Harry Potter Novels

- Ten year long fantasy series written by British author, J. K. Rowling
- Initially published in 1997, consisting of seven books
- Arguably the most popular fictional book series in the world
  - Over 500 million copies sold, the most of any book series in history
  - Translated into 80 languages
  - Over 600,000 works of Fanfiction

Harry Potter and the...

1. Philosopher's Stone
2. Chamber of Secrets
3. Prisoner of Azkaban
4. Goblet of Fire
5. Order of the Phoenix
6. Half-Blood Prince
7. Deathly Hallows

Slytherin to Python and…



Analyze the writing of the original seven Harry Potter books

Train a model to classify which book a chosen body of text belongs to

NLP

Natural Language Processing

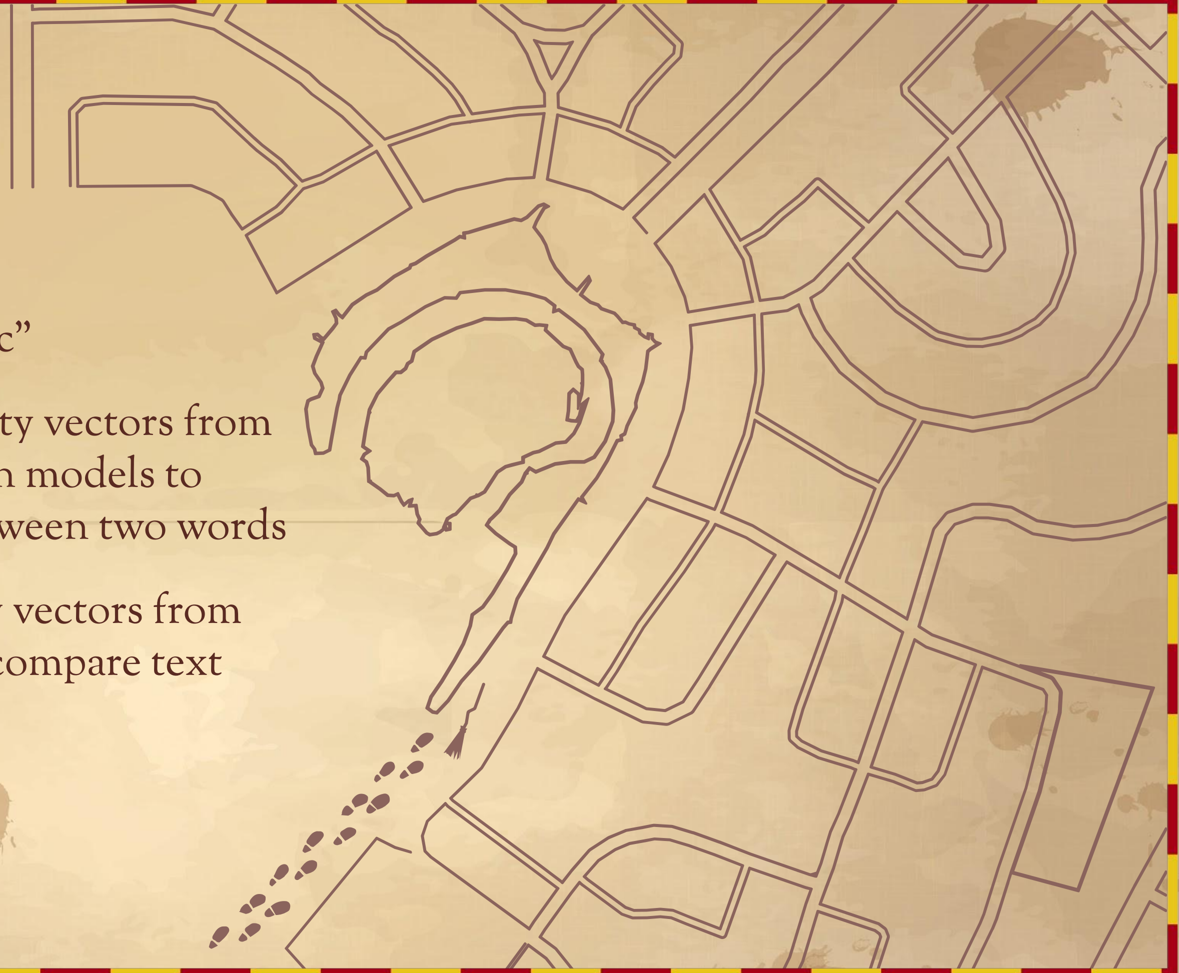
- Computers don’t inherently understand human language
- NLP helps a computer to understand the relationship and context of words

Mapping Text with Doc2Vec

MLP tool based off of “Word2Vec”

Word2Vec models create similarity vectors from individual words for classification models to understand the relationships between two words

Doc2Vec models create similarity vectors from groups of words (documents) to compare text similarity



Handling the Text Data

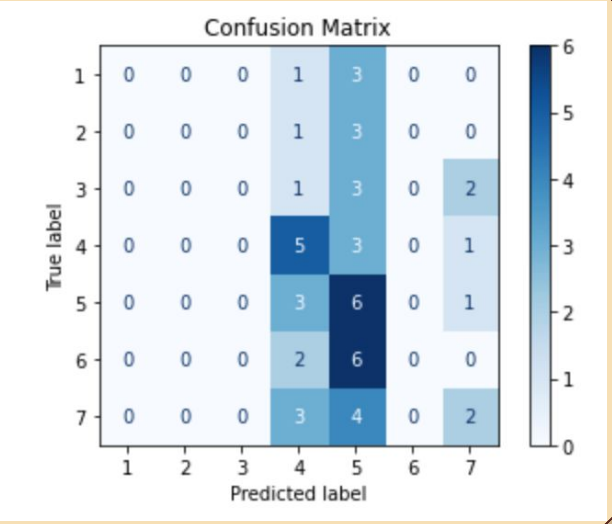
- Harry Potter corpora obtained from Kaggle as text files
- Seven Books
- 1.17 million words
- 199 chapters (our training data)
    - Book five is the longest with 38 chapters
    - Book one is the shortest with 17

Modeling

- Baseline accuracy is 19% if we predict book 5 every time
- Best Classification model so far is a random forest classifier
- Accuracy of 32%
- F-one score of 29%

Modeling Observations

- ❖ A lot harder to distinguish the writing style of each book than I expected
- ❖ J.K.'s Rowling's writing style is surprisingly consistent despite the different tones in each book
- ❖ Some classification models were shockingly terrible
  - ➢ Especially Logistic Regression shown here:



Future Goals

- Implement an app to classify the writing style of fanfiction
- Grid search on a stronger computer
- Further Sentiment analysis
- Compare various NLP vectorization tools to improve model performance

Thank you