



CZ4034 INFORMATION RETRIEVAL

GROUP 15

Group Members & Roles



Wilson



Mona



Kenneth

Agenda

1. Introduction



2. Web Crawling

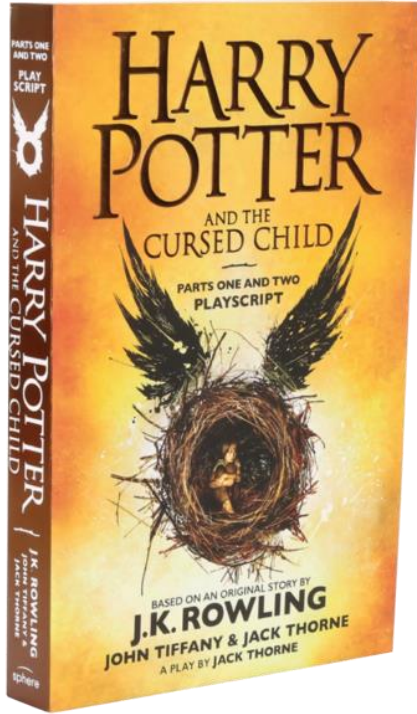


3. Indexing & Querying



4. Classification





Reasons for our Corpus Choice



One of the **most controversial and least rated** book of the Harry Potter series

Came with **high expectations** but received mixed reviews from readers

Has an **overall rating of 3.9** which is extremely low compared to all the other books of the Harry Potter series

Applications of the Analysis



**Read critical
reviews**



**Share the
same opinion**



**Decide whether
to buy book**

Web Crawling

Web Scrapping



- Data crawled from Amazon Customers' Review Page using Amazon Review Scraper
- Crawled 10175 book reviews

The screenshot displays the Amazon product page for 'Harry Potter and the Cursed Child, Parts One and Two: The Official Playscript of the ...' by Rowling, J.K. The page shows a 3.9 out of 5 star rating from 20,415 customer ratings. A bar chart indicates the distribution of star ratings: 5 stars (51%), 4 stars (16%), 3 stars (13%), 2 stars (8%), and 1 star (12%).

Below the reviews, the 'Top positive review' and 'Top critical review' sections are visible. The 'Top positive review' section shows a review by a user named 'Sitemaps' with a rating of 5 stars. The 'Top critical review' section shows a review by a user named 'Sitemaps' with a rating of 1 star.

At the bottom of the screenshot, a web scraper interface is shown, displaying a table of extracted data. The table has columns for ID, Selector, type, Multiple, Parent selectors, and Actions. The data is as follows:

ID	Selector	type	Multiple	Parent selectors	Actions
author	span a-profile-name	SelectorText	no	review	Element preview Data preview Edit Delete
title	a-size-base-review-title	SelectorText	no	review	Element preview Data preview Edit Delete
date	span a-size-base a-color-secondary	SelectorText	no	review	Element preview Data preview Edit Delete
content	div a-row-review-data span a-size-base	SelectorText	no	review	Element preview Data preview Edit Delete
rating	span a-icon-alt	SelectorText	no	review	Element preview Data preview Edit Delete



Web Crawling

author

title

date

rating

content

helpful

A.J. Garcia

★★★★★

Spell Binding

Reviewed in the United States on January 2, 2019

Format: Kindle Edition | **Verified Purchase**

I almost didn't read this because the reviews were so harsh, but the plot on wiki sounded like something I wanted to happen since book one. This epic story comes in the form of a play manuscript and leaves much up to the imagination! If you are strong with your own HP imagery, then this will be a pleasure to read. It will take a couple of chapters to get used to the format, no doubt, but press on! Just when you think you have it figured out, there is a plot twist I had not seen coming.

As a writer, I become skeptical when authors don't push plots to their full interest. This story reveals layer by layer a question I have been asking for eons: "What would happen if Potter got sorted into Slytherin?" But then it goes a step further than you can imagine. This is all I can say without revealing spoilers about this spellbinding book. The whole cast and crew is present to assist with an epic tale. Prepare yourself for the ride of a lifetime and (PLEASE) don't knock it till you've tried it.

8 people found this helpful

Web Crawling



SORT BY

FILTER BY

Top rated

All reviewers

All stars

5 star only

4 star only

3 star only

2 star only

1 star only

All positive

All critical

All formats

Text, image, video

Showing 1-10 of 12,700 reviews



Biana



And that becomes terribly clear very early

Reviewed in the United States on August 5, 2016

Format: Kindle Edition | **Verified Purchase**

Ian Malcolm: I'll tell you the problem with the scientific po

next step. You didn't earn the knowledge for yourselves, a

didn't require any discipline to attain it. You read what others had done

it for it. You stood on the shoulders of geniuses to accomplish some

**Multiple filters
crawled**

Indexing & Querying

Indexing



- Apache version solr 8.5.0 is used for indexing .csv file
- Inverted indexing is used as it is more efficient
- Web interface is hosted on the Apache HTTP Server by XAMPP

Querying Demo

Output Analysis of Querying



QUERY	SPEED OF QUERYING IN MS
content:harry	97
rating: "3.0 out of 5 stars"	18
rating: "3.0 out of 5 stars" AND (content: "Harry" OR content: "Ron")	3
(rating: "4.0 out of 5 stars")^1.5 OR (content: "cursed child")	7
content: "harry child"~4	149

Classification

Merging Data



amazon_positive_reviews.csv

amazon_critical_reviews.csv

amazon_reviews_1star.csv

amazon_reviews_2star.csv

amazon_reviews_3star.csv

amazon_reviews_4star.csv

amazon_reviews_5star.csv

amazon_image_reviews.csv

1. Merge

all_amazon_reviews.csv

- web-scraper-order
- web-scraper-start-url
- author
- title
- date
- content
- rating
- next
- next-href
- helpful
- image-src
- image_2-src

2. Remove duplicates & redundant columns

reviews_df

- author
- title
- date
- content
- rating
- helpful
- image-src

Data Pre-processing



Data Frame

- Removal of duplicates and irrelevant columns
- Formatting columns
- Creating sentiment feature

Review Column

- Case-folding
 - Stripping of whitespaces
 - Removal of empty strings
- Stemming
 - Lemmatization
 - Stop words removal

Vectorization

- Count Vectorization
- TF-IDF Vectorization
- Unigrams + Bigrams
- Bigrams
- Bigrams + Trigrams
- Unigrams + Bigrams + Trigrams

Data Pre-processing



Data Frame

- Removal of duplicates and irrelevant columns
- Formatting columns
- Creating sentiment feature

Review Column

- Case-folding
- Stripping of whitespaces
- Removal of empty strings

- Stemming
- Lemmatization
- Stop words removal

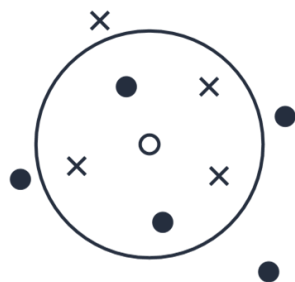
Vectorization

- Count Vectorization
- TF-IDF Vectorization
- Unigrams + Bigrams
- Bigrams
- Bigrams + Trigrams
- Unigrams + Bigrams + Trigrams

Classification Models



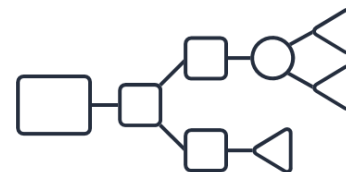
Naïve Bayes



K-Nearest Neighbour



Support Vector Machines



Decision Trees

Enhancing Classification



Innovations for enhancing classifications:

- Ensemble Classification - Random Forest Classification
- k-Fold Cross Validation
- GridsearchCV
- Error Analysis

Results



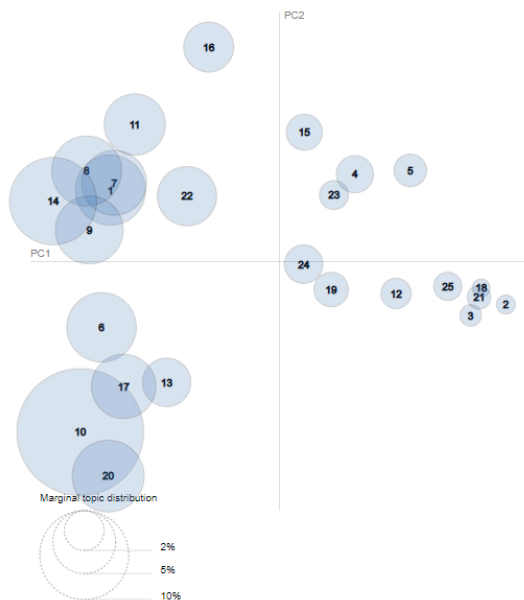
MODEL	F1 SCORE	PRECISION	RECALL	CV FOLDS
K-Nearest Neighbour	● 0.754	● 0.813	● 0.705	5
Decision Tree	● 0.782	● 0.737	● 0.812	5
Naïve Bayes	● 0.863	● 0.901	● 0.830	5
Random Forest	● 0.867	● 0.838	● 0.899	5
Linear Support Vector Machines	● 0.873	● 0.849	● 0.899	5

Topic Modelling with LDA



Selected Topic:

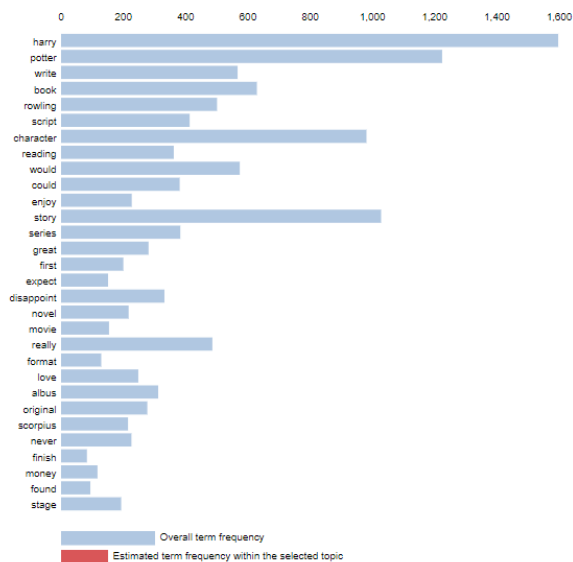
Intertopic Distance Map (via multidimensional scaling)



Slide to adjust relevance metric: (2)
 $\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1

Top-30 Most Salient Terms¹



1. $\text{saliency}(\text{term}, w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$ for topics t ; see Chuang et. al. (2012)

2. $\text{relevance}(\text{term}, w | \text{topic } t) = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w)$; see Sievert & Shirley (2014)

Summary

We built our information retrieval system based on the customer reviews of the book, *Harry Potter and the Cursed Child Part 1 and 2*, on Amazon.

Crawling

- Crawled the Amazon review page of Harry Potter and the Cursed Child
- Used the Amazon Review Scraper, a Google Chrome extension by Scrapehero
- Exported the data in CSV
- Merged the data from multiple CSV files

Querying and Indexing

- Created a web interface consisting of HTML and PHP files and integrated them with Solr
- Tested various query methods such as boost query and proximity query in Solr

Classification

- Conducted data pre-processing methods such as count vectorization, n-grams, case-folding
- Classify the sentiment of reviews into positive or negative
- Used models such as Naïve Bayes, KNN, SVM, decision trees and random forest
- Improved classification results using GridSearchCV and error analysis
- Implemented custom pre-processing methods specific to the domain

END