CZ4034 INFORMATION RETRIEVAL

GROUP 15

Group Members & Roles

Wilson

Mona

Kenneth

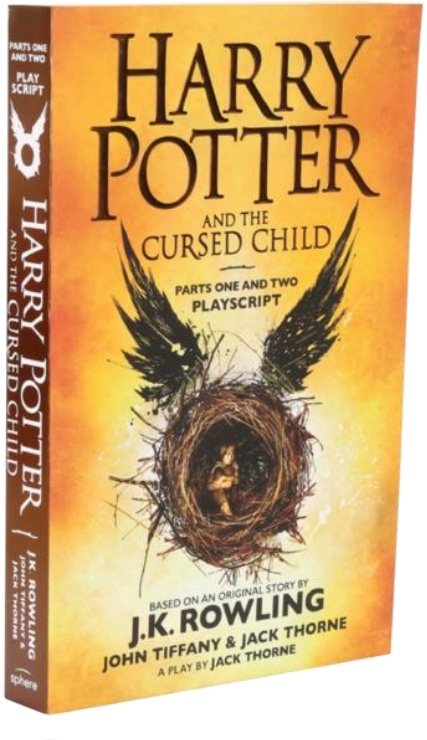
Agenda

1. Introduction

2. Web Crawling

3. Indexing & Querying

4. Classification


PARTS ONE AND TWO
PLAY SCRIPT
HARRY POTTER
AND THE CURSED CHILD
J.K. ROWLING & JACK THORNE
sphere
HARRY POTTER
AND THE CURSED CHILD
PARTS ONE AND TWO
PLAYSCRIPT
BASED ON AN ORIGINAL STORY BY
J.K. ROWLING
JOHN TIFFANY & JACK THORNE
A PLAY BY JACK THORNE

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

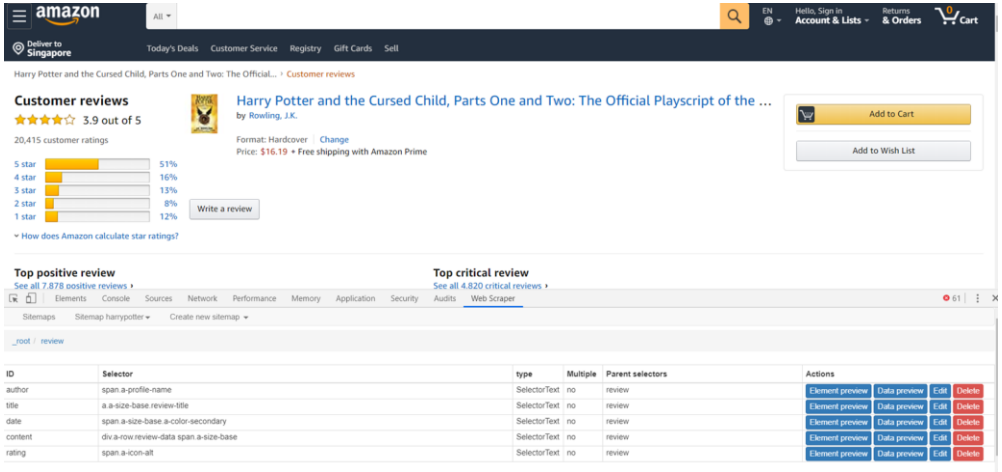
Share the same opinion

Decide whether to buy book

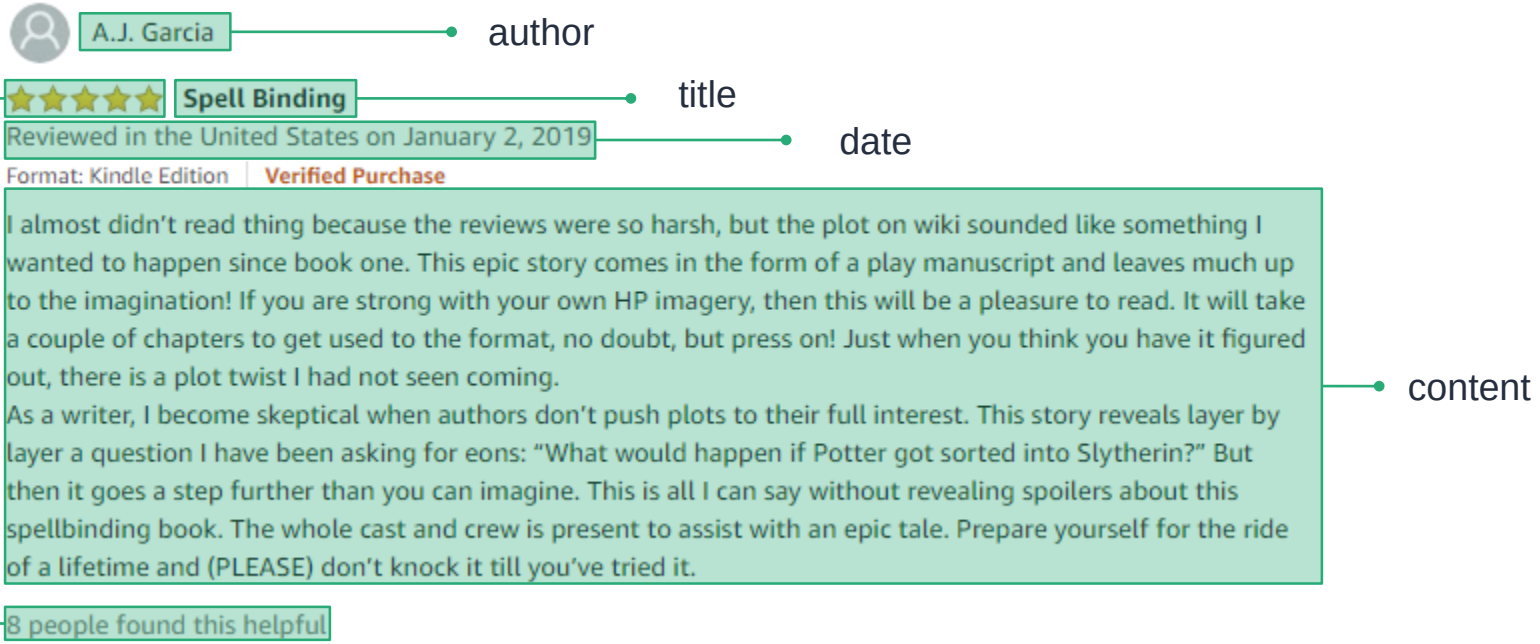
Web Crawling

Web Scraping

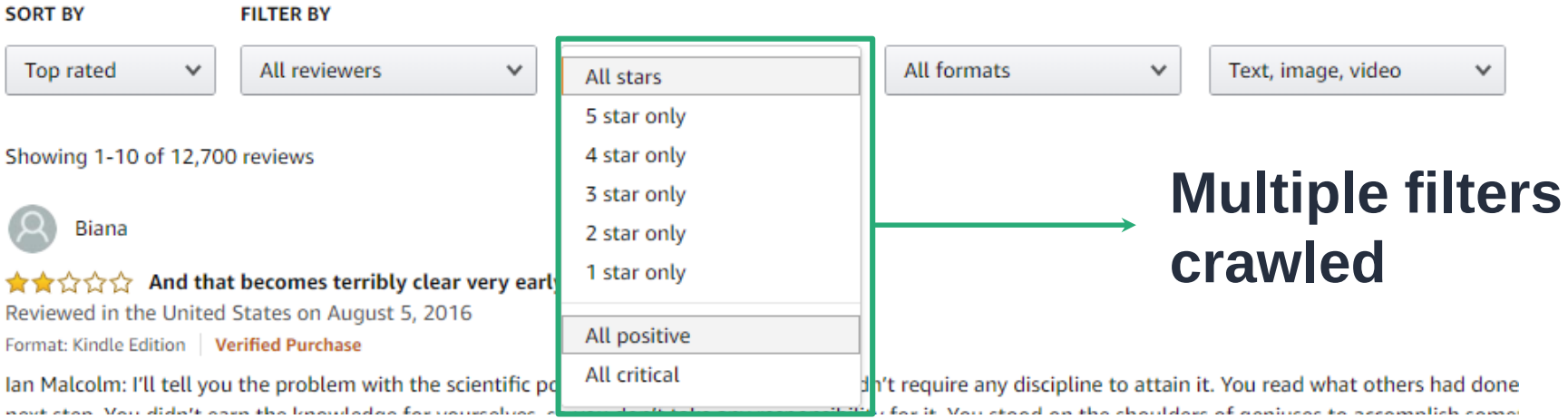
- Data crawled from Amazon Customers' Review Page using Amazon Review Scraper
- Crawled 10175 book reviews



Web Crawling



Web Crawling



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

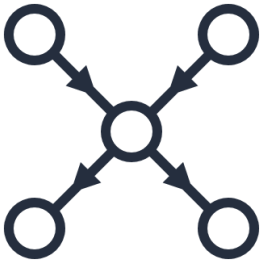
- Case-folding
- Stripping of whitespaces
- Removal of empty strings

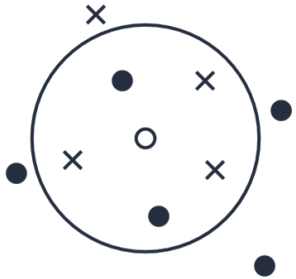
- Stemming
- Lemmatization
- Stop words removal

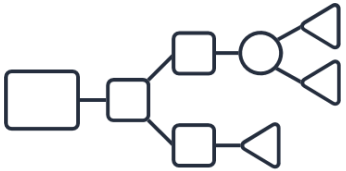
Vectorization

- 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

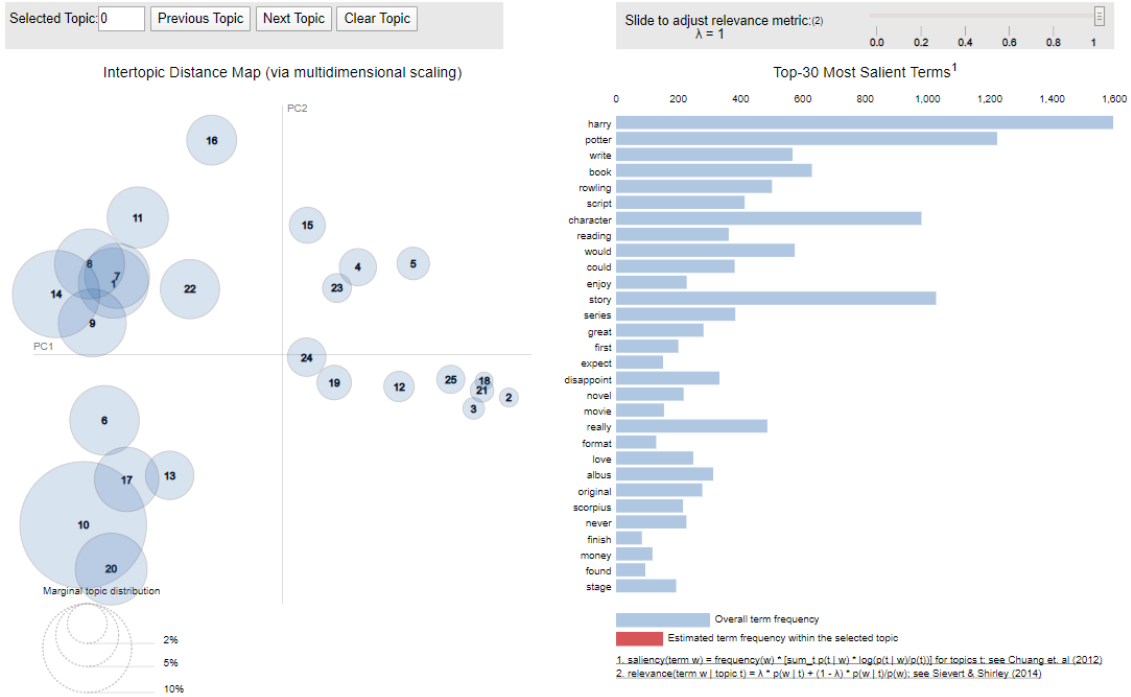
Innovations for enhancing classifications:

- 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



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