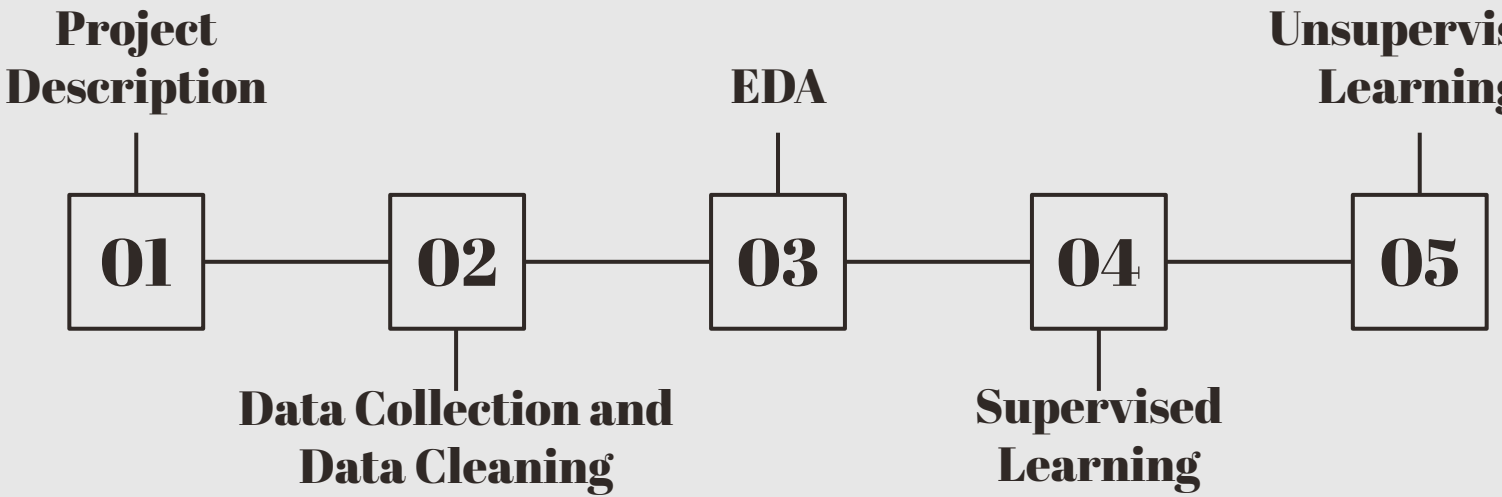
CS105 Final Project

Lillian Xiao

Analysis on Lolita Book Reviews

Outline


Project Description
01
02
Data Collection and Data Cleaning
EDA
03
04
Supervised Learning
Unsupervised Learning
05

01 Project Description

Project Description

Introduction

- Mining and Analyzing reviews of “Lolita” by Vladimir Nabokov

Description of Project

- Evaluate reader opinion on the novel using two techniques using data gathered from Goodreads

Techniques

- Logistic Regression and K-means Clustering

Hypotheses

- Lolita is a controversial novel that elicits mixed reactions

02 Data Collection And Data Cleaning

Data Collection

- Scraped data from GoodReads using BeautifulSoup
  - Parse using BeautifulSoup
  - Create dictionary
  - Navigate web page’s HTML structure
  - Find the information needed and append to dictionary

Data Cleaning

- Filtered to English only
- Tokenize, lemmatize
  - removed stop words, correct typos, etc
- Ratings: replace string w/ numerical value

	reviews	rating
0	Between the CoversAfter re-reading "Lolita", I...	Rating 5 out of 5
1	Nymph. Nymphet. Nymphetiquette. Nymphology. Ny...	Rating 5 out of 5
2	Now, this is going to be embarrassing to admit...	Rating 5 out of 5
3	I wasn't even going to write a review ofLolita...	Rating 4 out of 5
4	when i first read this book, i hated every sec...	Rating 4 out of 5

	reviews	rating
0	local bookseller ever read firmly going either...	5.0
1	nymph nymphet never think year old way stain b...	5.0
2	going embarrassing know reading enjoying book ...	5.0
3	even going write review finishing honestly man...	4.0
4	first read book every second pride reader dist...	4.0

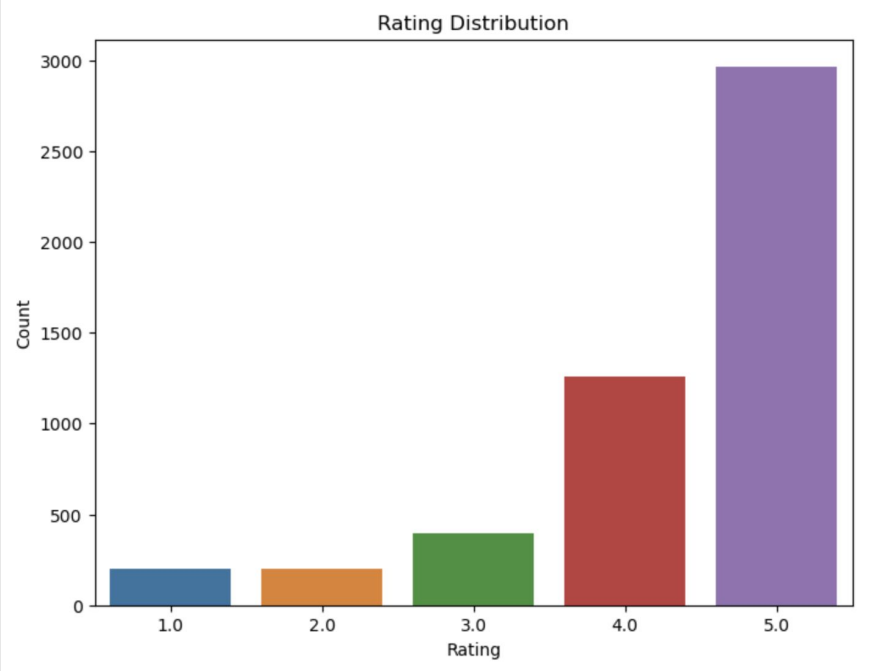
Vectorization

- Used TF-IDF to measure how prevalent each word is within a review relative to the corpus

	aback	abandon	ability	abject	able
0	0.02469	0.02469	0.0	0.0	0.016814
1	0.00000	0.00000	0.0	0.0	0.023185
2	0.00000	0.00000	0.0	0.0	0.000000
3	0.00000	0.00000	0.0	0.0	0.000000
4	0.00000	0.00000	0.0	0.0	0.000000

03 EDA

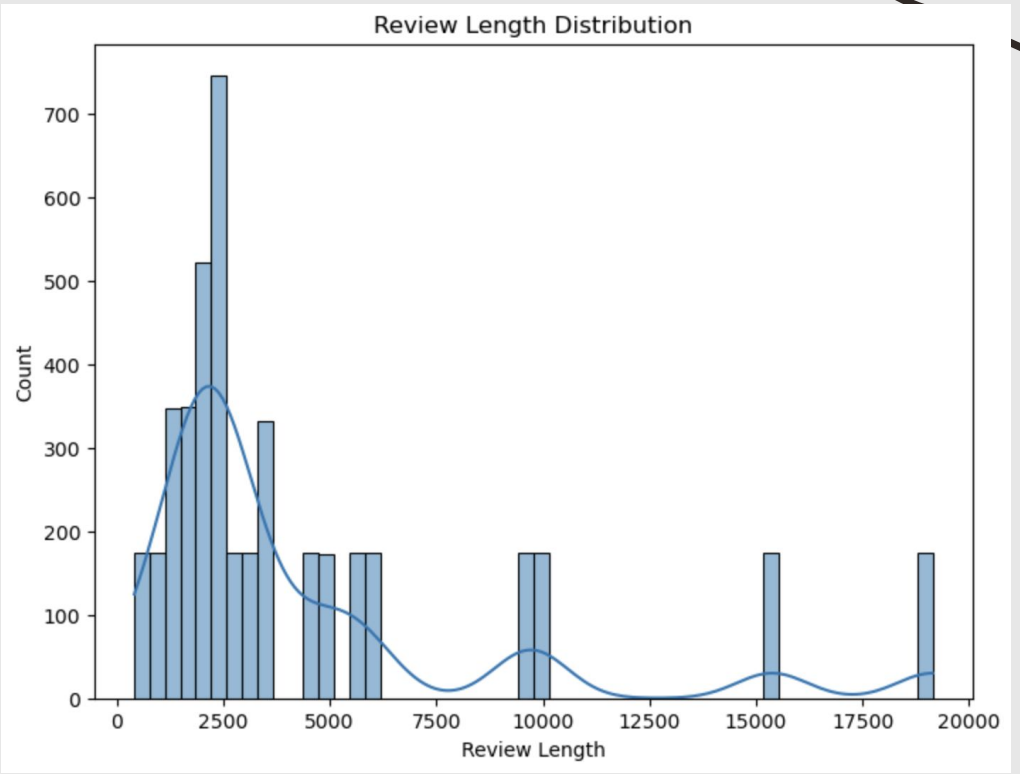
Analyze Frequency of Ratings



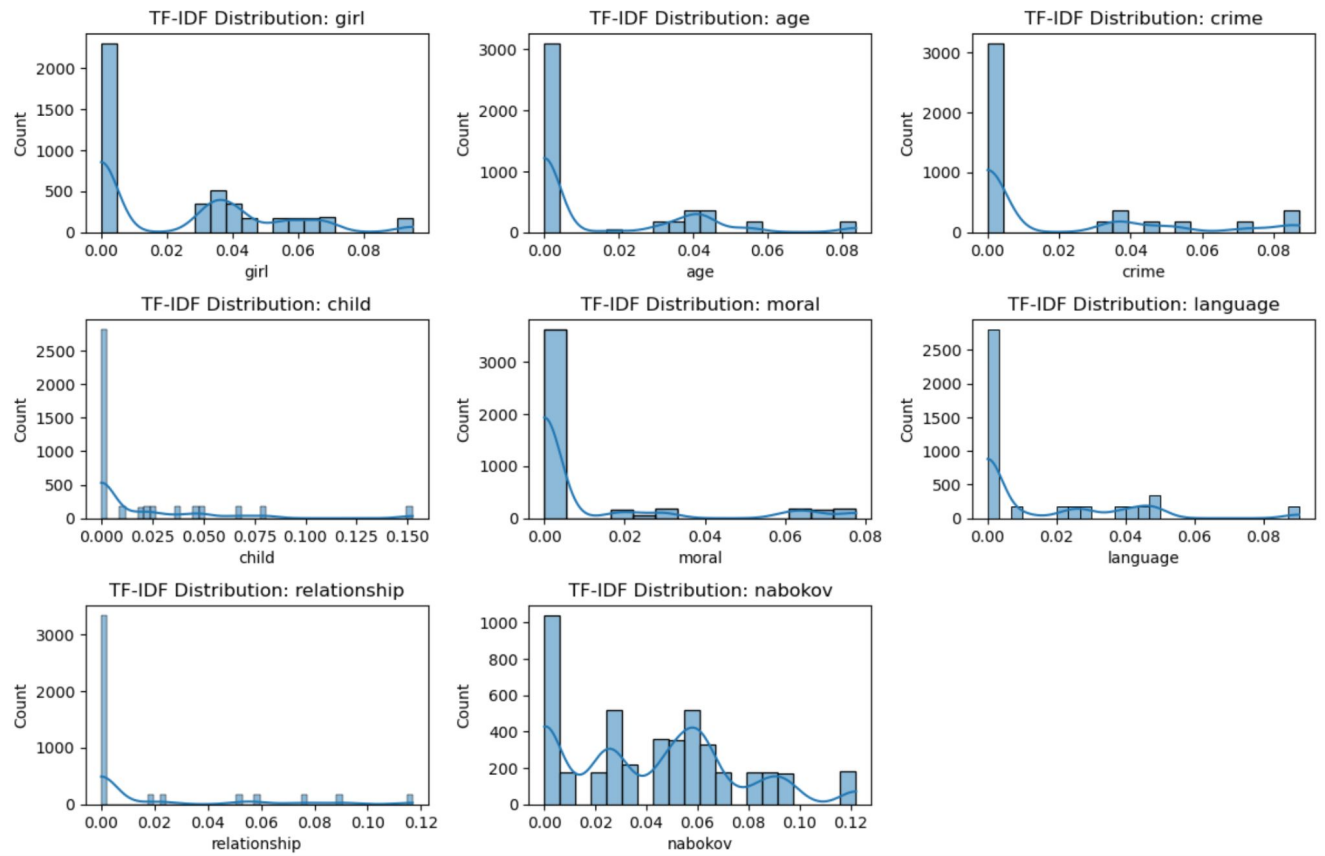
- Mostly 5 star reviews
- Almost same amount of 1 and 2 star reviews

Analyze Length of Reviews

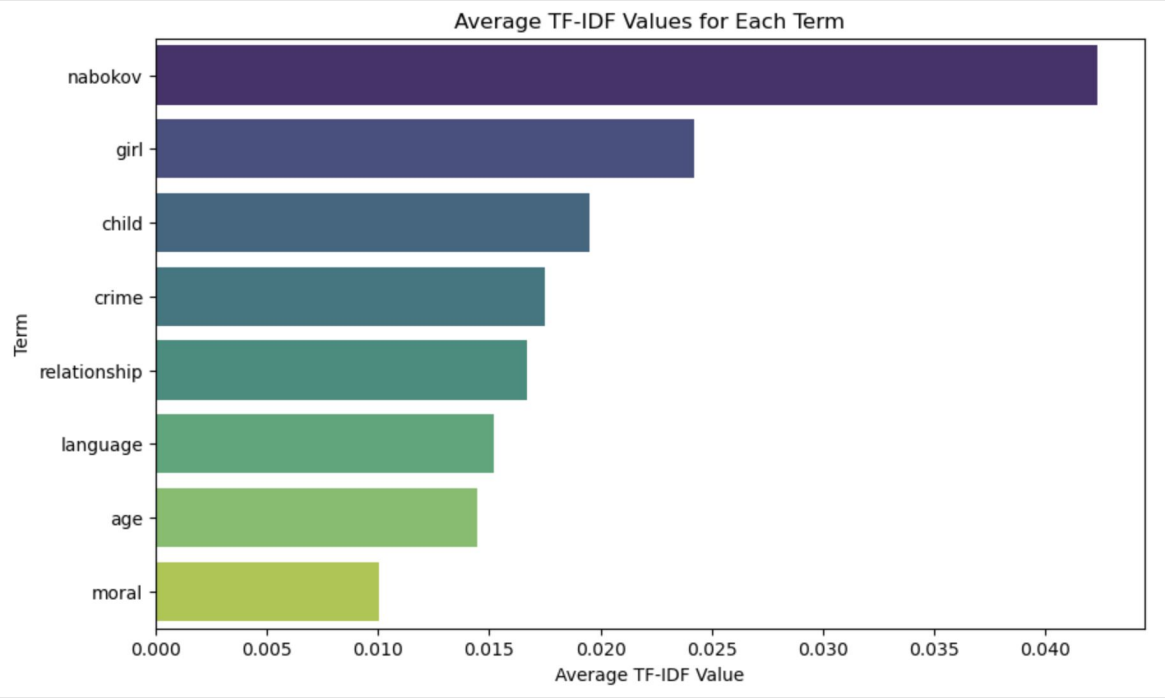
- Most reviews were about 2,500 characters
- Right/Positively skewed



TF-IDF Distribution for Each Word

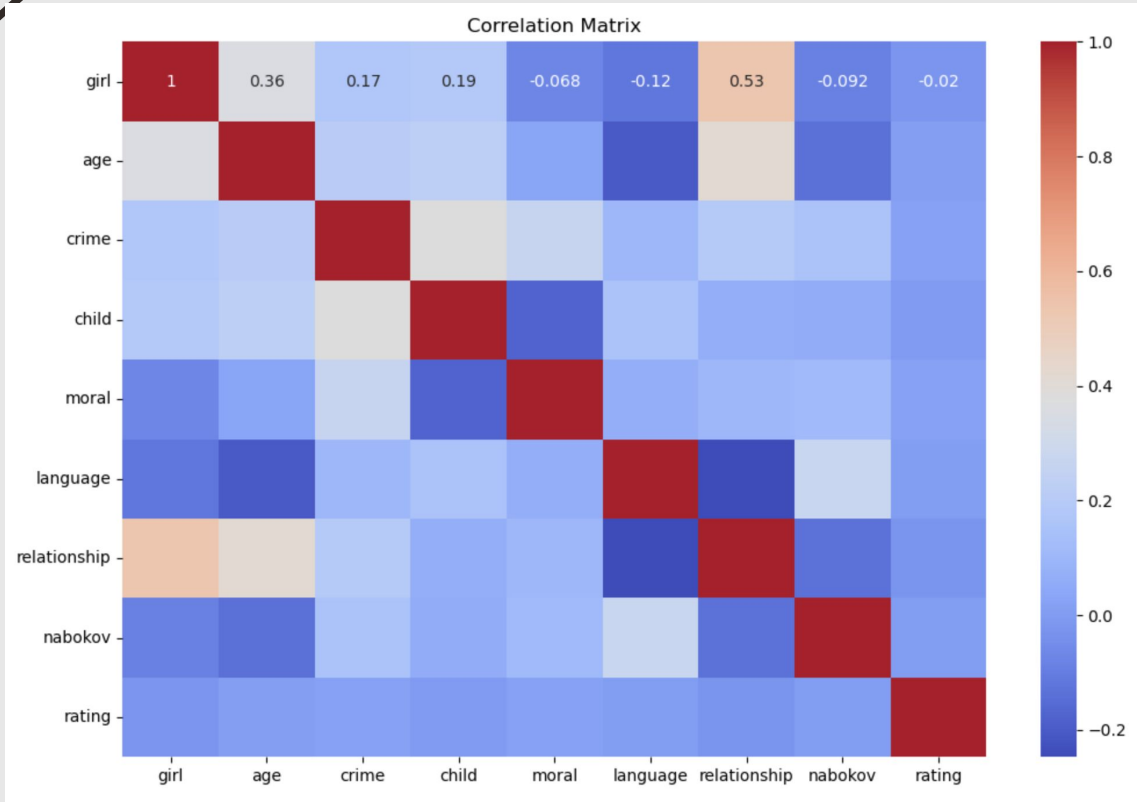


Average TF-IDF Values



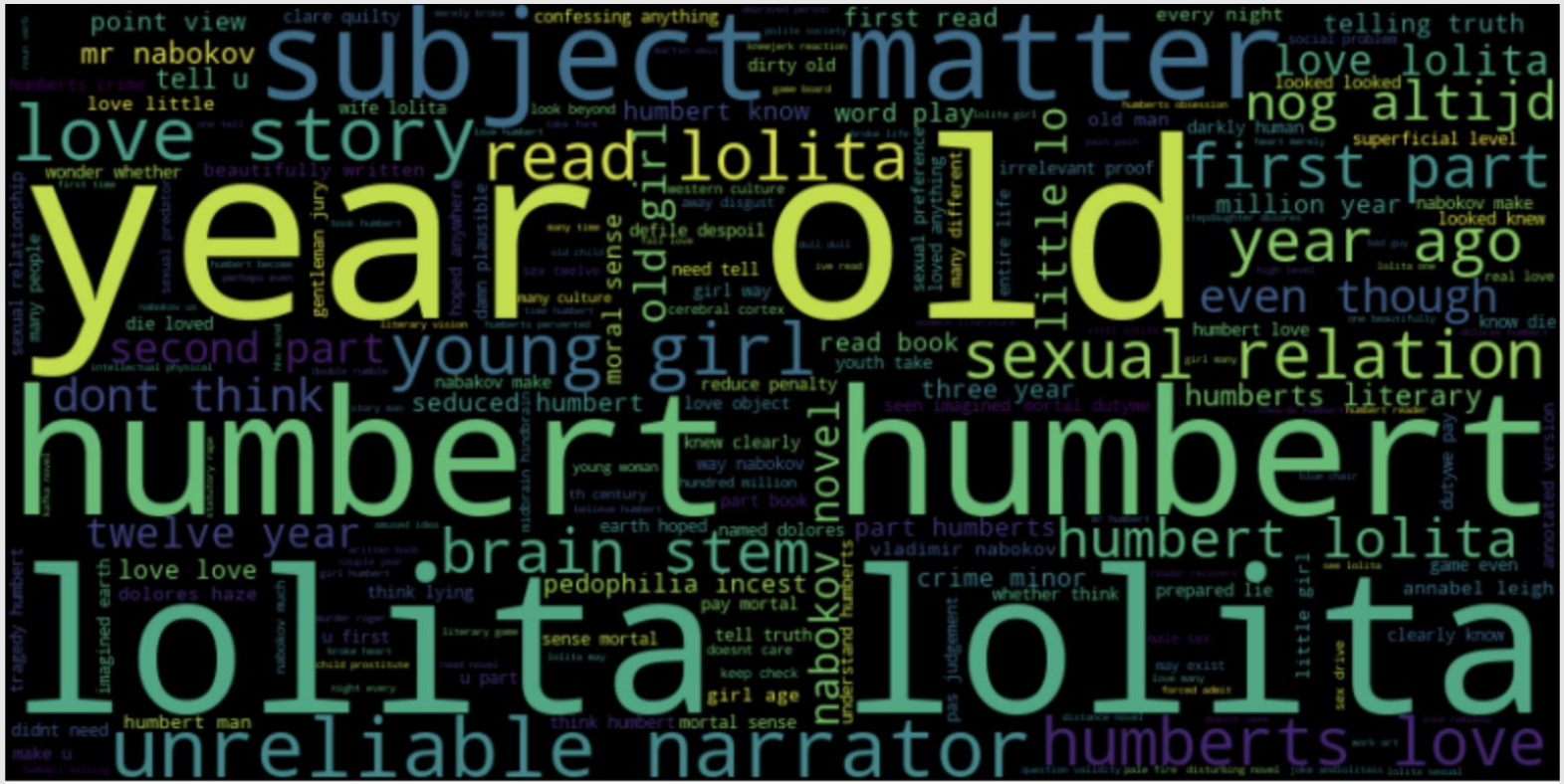
- Ranking of word importance

Correlation Matrix

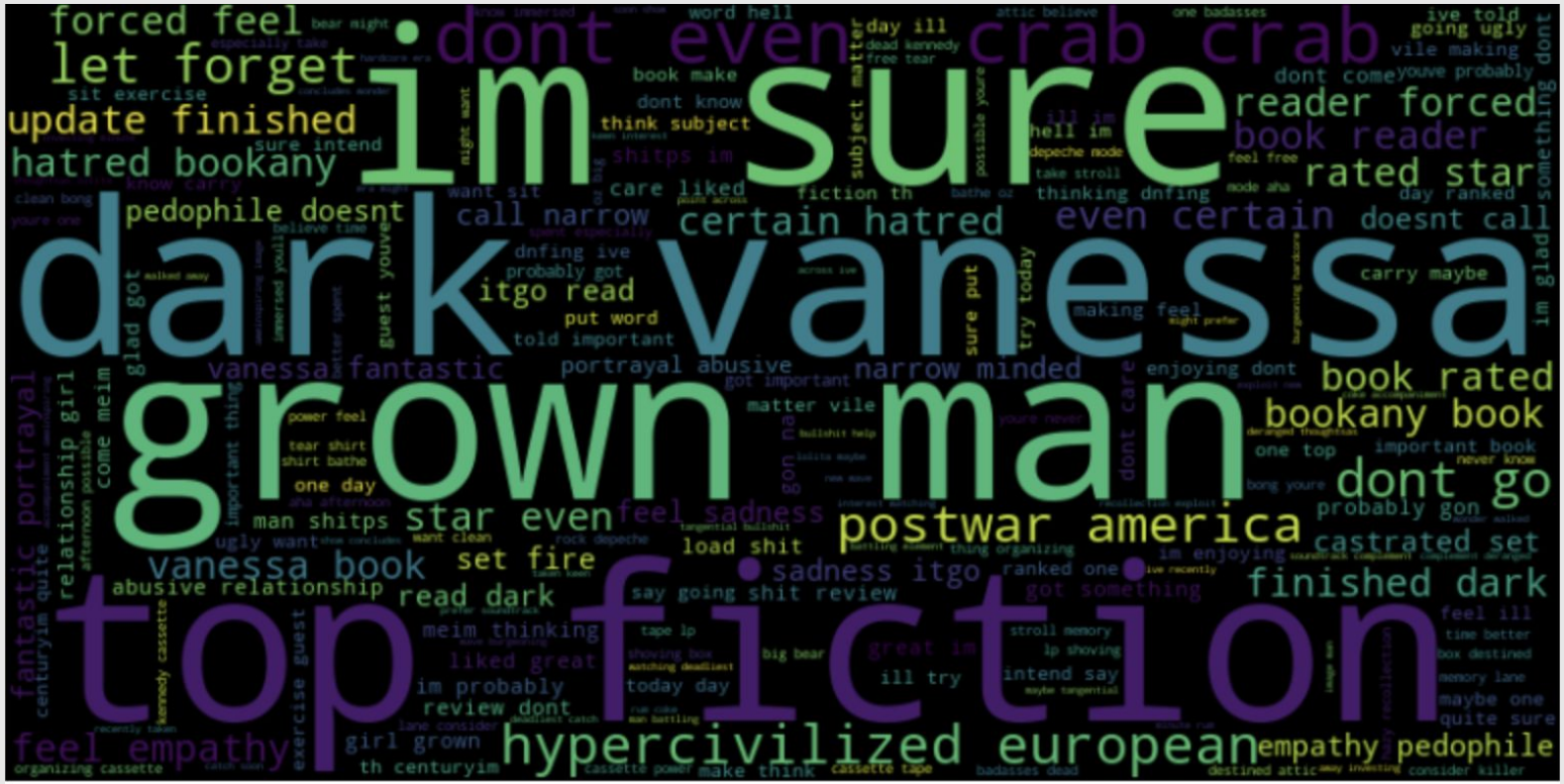


- Values: 1 to -0.2
- Positive relation
- Negative relation

Word Cloud: Positive Reviews



Word Cloud: Negative Reviews



04 Supervised Learning: Logistic Regression

Goal

Create a logistic regression model to help us predict whether a reviewer gave Lolita five stars based on how their review uses certain words.

Preparing the data

- Selected features for predictors: `['year', 'old', 'subject', 'unreliable', 'grown', 'girl', 'sure', 'crime']`
  - Chosen based off of word clouds and most common terms
- Merged reviews & ratings dataframe with selected feature tf-idf columns
- Converted ratings column from numeric to boolean
  - 5 stars = 1
  - > 5 stars = 0

Preparing the data

Over half of all ratings were 5/5, so in an effort to make the data more balanced, we turned ratings into a binary variable.

	reviews	rating	year	old	subject	unreliable	grown	girl	sure	crime
0	local bookseller ever read firmly going either...	1.0	0.068522	0.072439	0.036109	0.02865	0.000000	0.000000	0.00000	0.071625
1	nymph nymphet never think year old way stain b...	1.0	0.047265	0.049967	0.066419	0.00000	0.000000	0.000000	0.00000	0.019762
2	going embarrassing know reading enjoying book ...	1.0	0.078394	0.062156	0.000000	0.00000	0.062931	0.074245	0.00000	0.049166
3	even going write review finishing honestly man...	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000
4	first read book every second pride reader dist...	0.0	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.07359	0.000000

Stratified split & validation

- Stratified train-test split to retain ratio of 5 stars vs not 5 stars
  - 0.75 train, 0.25 test
- Stratified 10-fold cross-validation (k = 10 is a common choice)

```
Iteration  1
accuracy:  0.784          intercept:  [0.32543076]
 coefficients:  [[14.06878504 -1.68371884 -3.53534617  5.93223952 -5.30898066  1.69848181
  -7.29736809 -4.24443198]]

Iteration  2
accuracy:  0.784          intercept:  [0.32543076]
 coefficients:  [[14.06878504 -1.68371884 -3.53534617  5.93223952 -5.30898066  1.69848181
  -7.29736809 -4.24443198]]

Iteration  3
accuracy:  0.784          intercept:  [0.32543076]
 coefficients:  [[14.06878504 -1.68371884 -3.53534617  5.93223952 -5.30898066  1.69848181
  -7.29736809 -4.24443198]]
```

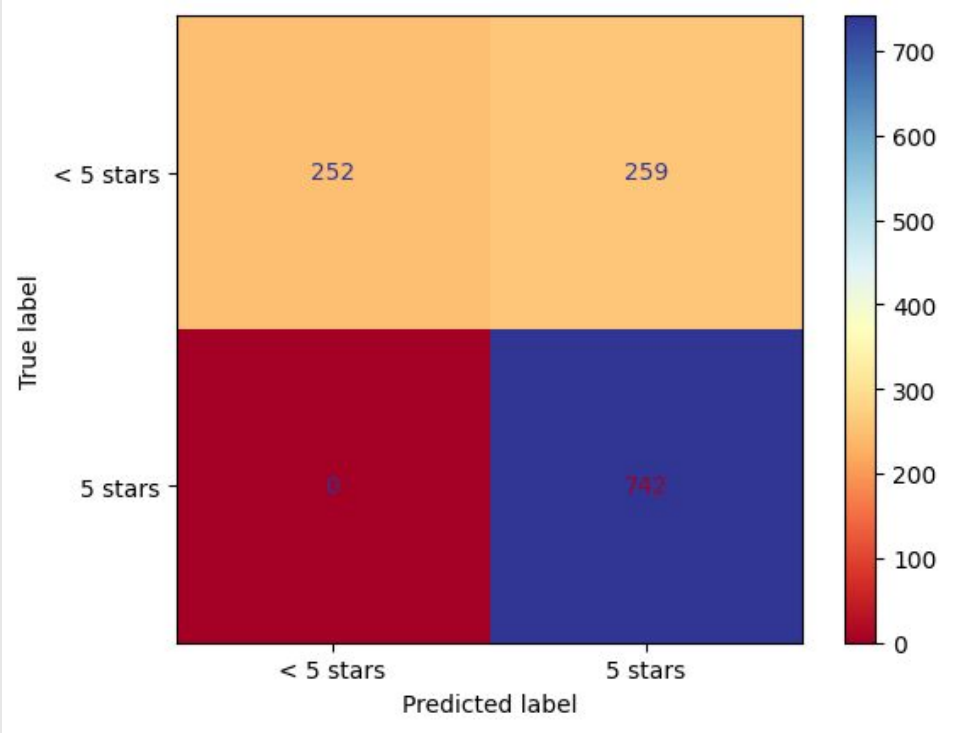
Result & interpretation

- Averaged scores and parameter estimates from 10-fold cross-validation
- Generally it looks like, on average, reviews with more mention of the words “year” & “unreliable” were more likely to rate Lolita a 5/5, and those with more mention of “subject” and “sure” were more likely to rate Lolita lower than 5/5
- Average accuracy: 0.79

```
avg accuracy: 0.7895702127659575
avg intercept est: [0.32450044]
average coefficient estimates:
'year': 14.064844822327064
'old': -1.6937995952632168
'subject': -33.6227908512463
'unreliable': 5.816364014786968
'grown': -5.400658674393614
'girl': 1.8542619484600293
'sure': -7.337653670075892
'crime': -4.28760369091506
```

Confusion matrix

- Likely due to unbalanced samples, model was 100% successful predicting 5-star reviews but ~50% success rate predicting > 5-star reviews
  - Perfect recall (= 1)
- In the future we could try getting a more balanced sample and/or using a different method of feature selection



05 Unsupervised Learning Technique - K-means Clustering

Goal

Use k-means clustering in analyzing overall sentiment of Lolita.

clean data >> calculate sentiment >> k-means >> visualize/analyze

Preparing the Data

- nltk to tokenize + lemmatize + remove stop words
- remove words incorrectly combined during vectorization
- remove non-sentiment words

	review,rating
0	December 7, 2017Between the CoversAfter re-rea...
1	March 15, 2017Nymph. Nymphet. Nymphetiquette. ...
2	September 15, 2023Now, this is going to be emb...
3	April 5, 2020I wasn't even going to write a re...
4	December 13, 2023when i first read this book, ...
...	...
5965	August 26, 2023Prof. Harry Levin of Harvard sa...
5966	November 9, 2018Αυτό το επί πολλά χρόνια απαγο...
5967	May 12, 2014Warning: contains spoilers forThe ...
5968	January 24, 2023In this sulfurous and scandalo...
5969	March 28, 2024When Humbert Humbert, (his pare...

	reviews
0	local bookseller ever read firmly going either...
1	march nymphet never think year old way stain b...
2	going embarrassing know reading enjoying book ...
3	even going write review finishing honestly man...
4	first read book every second pride reader dist...
...	...
5965	august harry levin great book darkly symbolica...
5966	light life fire
5967	may murder roger de remember seeing interview ...
5968	sulfurous scandalous novel reader ethic bring ...
5969	march little imagination thirteen fell love gi...

Sentiment Analysis w/ VADER

- classify sentiment of each review based on calculated sentiment score
- categorizes sentiment based off score
    - extremely negative (-1) - extremely positive(1)
- negative, neutral, positive and compound (normalize neg+neu+pos values)

```
if score >= 0.75:
    return 'extremely positive'
elif score >= 0.25:
    return 'positive'
elif score >= 0.05:
    return 'slightly positive'
elif score > -0.05:
    return 'neutral'
elif score > -0.25:
    return 'slightly negative'
elif score > -0.75:
    return 'negative'
else:
    return 'extremely negative'
```

	reviews	neg	neu	pos	compound	sentiment
0	local bookseller ever read firmly going either...	0.196	0.547	0.258	0.9980	extremely positive
1	march nymphet never think year old way stain b...	0.152	0.694	0.154	-0.2615	negative
2	going embarrassing know reading enjoying book ...	0.215	0.501	0.285	0.9640	extremely positive
3	even going write review finishing honestly man...	0.152	0.557	0.291	0.9623	extremely positive
4	first read book every second pride reader dist...	0.191	0.550	0.260	0.8639	extremely positive
...	...	...	...	...	...	...
5965	august harry levin great book darkly symbolica...	0.252	0.524	0.224	-0.7783	extremely negative
5966	light life fire	0.545	0.455	0.000	-0.3400	negative
5967	may murder roger de remember seeing interview ...	0.196	0.594	0.210	-0.6396	negative
5968	sulfurous scandalous novel reader ethic bring ...	0.262	0.522	0.217	-0.7200	negative
5969	march little imagination thirteen fell love gi...	0.157	0.596	0.246	0.9657	extremely positive

Goals of Analysis

Approach 1: perform k-means on compound column

- analyze distribution of compound values within the clusters to understand overall sentiment

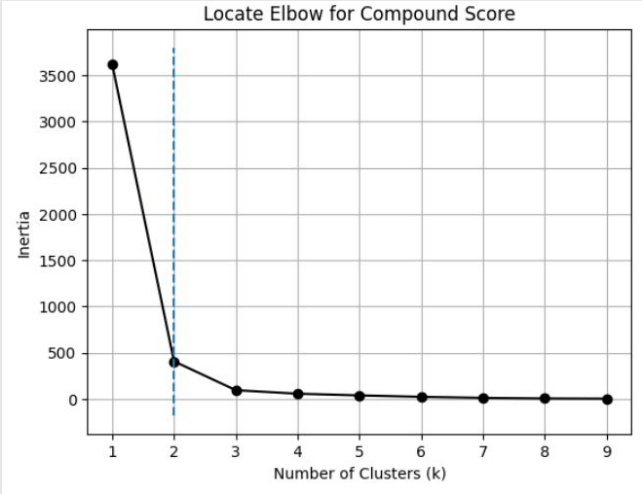
Approach 2: perform k-means on grouped scores

- combine neg, neu, and pos scores for k-means clustering
- takes into account more variables
- allows for more in depth analysis

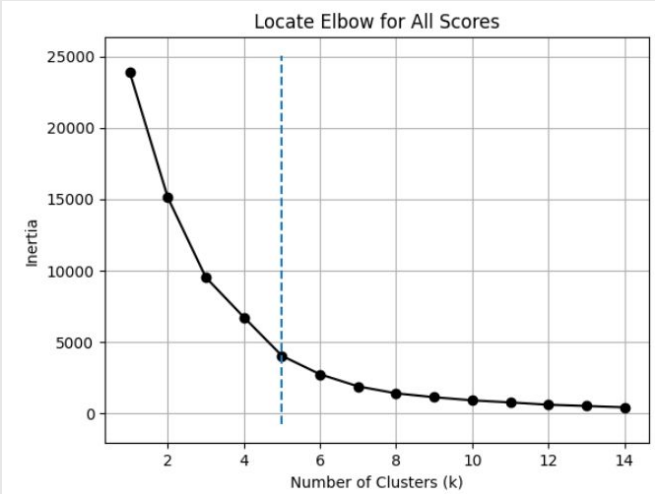
K-Means (Elbow Method)

- compute inertia for k values and plot
- KneeLocator - mathematically calculate optimal k-value

Approach 1: k val = 2



Approach 2: k val = 5



K-Means cont.

Approach 1

- k=2
- run K-Means algorithm on compound column
- create cluster column

	reviews	neg	neu	pos	compound	sentiment	cluster
0	local bookseller ever read firmly going either...	0.196	0.547	0.258	0.9980	extremely positive	0
1	march nymphet never think year old way stain b...	0.152	0.694	0.154	-0.2615	negative	1
2	going embarrassing know reading enjoying book ...	0.215	0.501	0.285	0.9640	extremely positive	0
3	even going write review finishing honestly man...	0.152	0.557	0.291	0.9623	extremely positive	0
4	first read book every second pride reader dist...	0.191	0.550	0.260	0.8639	extremely positive	0
...	...	...	...	...	...	...	...
5965	august harry levin great book darkly symbolica...	0.252	0.524	0.224	-0.7783	extremely negative	1
5966	light life fire	0.545	0.455	0.000	-0.3400	negative	1
5967	may murder roger de remember seeing interview ...	0.196	0.594	0.210	-0.6396	negative	1
5968	sulfurous scandalous novel reader ethic bring ...	0.262	0.522	0.217	-0.7200	negative	1
5969	march little imagination thirteen fell love gi...	0.157	0.596	0.246	0.9657	extremely positive	0

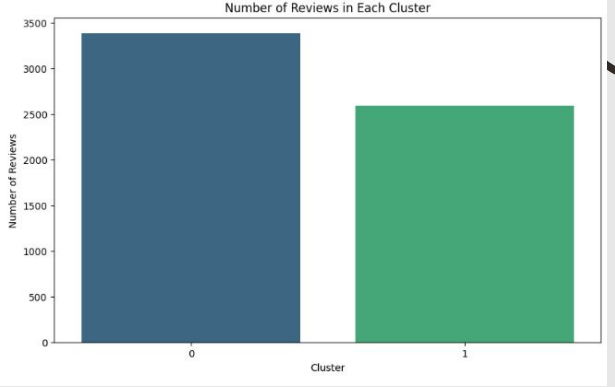
Approach 2

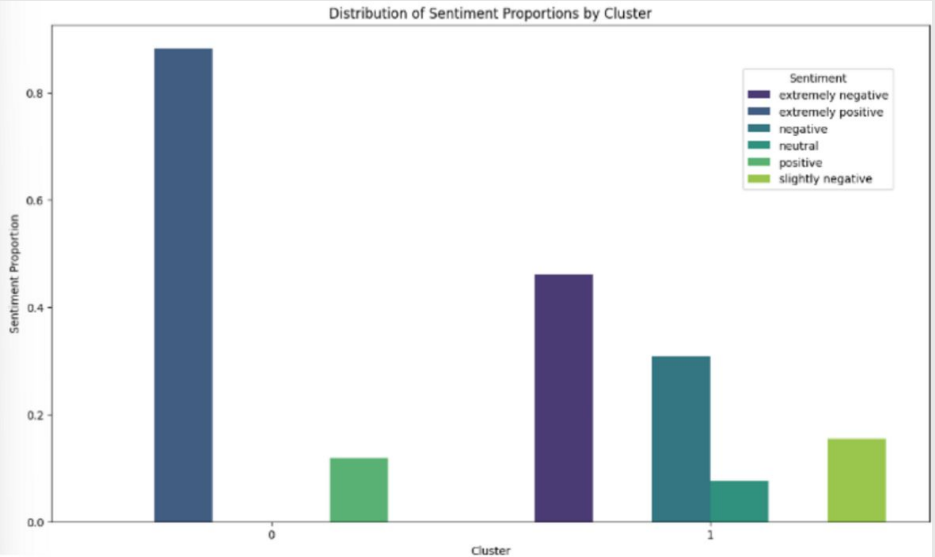
- k=5
- run K-Means algorithm on features [neg, neu, pos, compound]
- create cluster column

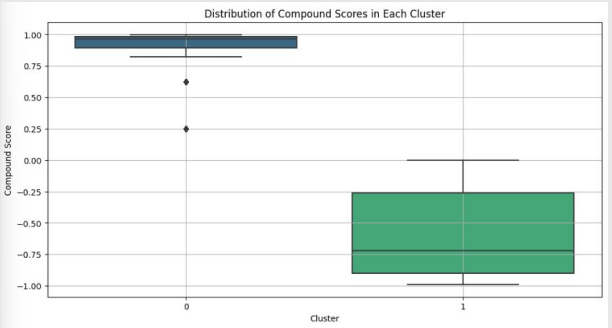
	reviews	neg	neu	pos	compound	sentiment	cluster
0	local bookseller ever read firmly going either...	0.196	0.547	0.258	0.9980	extremely positive	0
1	march nymphet never think year old way stain b...	0.152	0.694	0.154	-0.2615	negative	1
2	going embarrassing know reading enjoying book ...	0.215	0.501	0.285	0.9640	extremely positive	0
3	even going write review finishing honestly man...	0.152	0.557	0.291	0.9623	extremely positive	0
4	first read book every second pride reader dist...	0.191	0.550	0.260	0.8639	extremely positive	0
...	...	...	...	...	...	...	...
5965	august harry levin great book darkly symbolica...	0.252	0.524	0.224	-0.7783	extremely negative	1
5966	light life fire	0.545	0.455	0.000	-0.3400	negative	3
5967	may murder roger de remember seeing interview ...	0.196	0.594	0.210	-0.6396	negative	1
5968	sulfurous scandalous novel reader ethic bring ...	0.262	0.522	0.217	-0.7200	negative	1
5969	march little imagination thirteen fell love gi...	0.157	0.596	0.246	0.9657	extremely positive	0

Approach 1 - Analysis

- reviews sentiments are mixed
- more readers maintain an overall positive opinion
    - extremely positive sentiments > negative sentiments

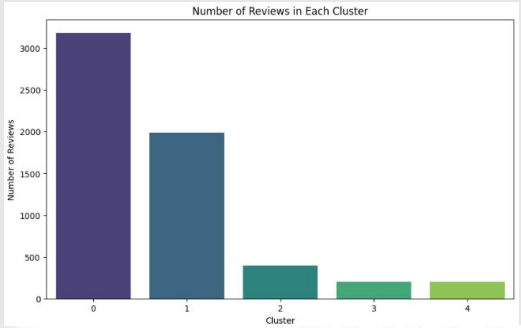


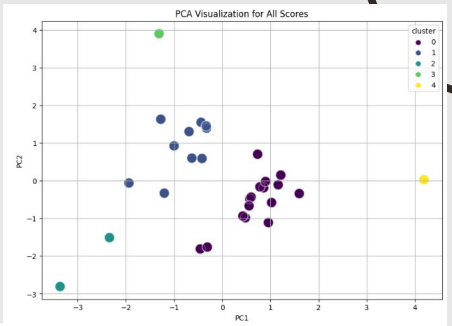


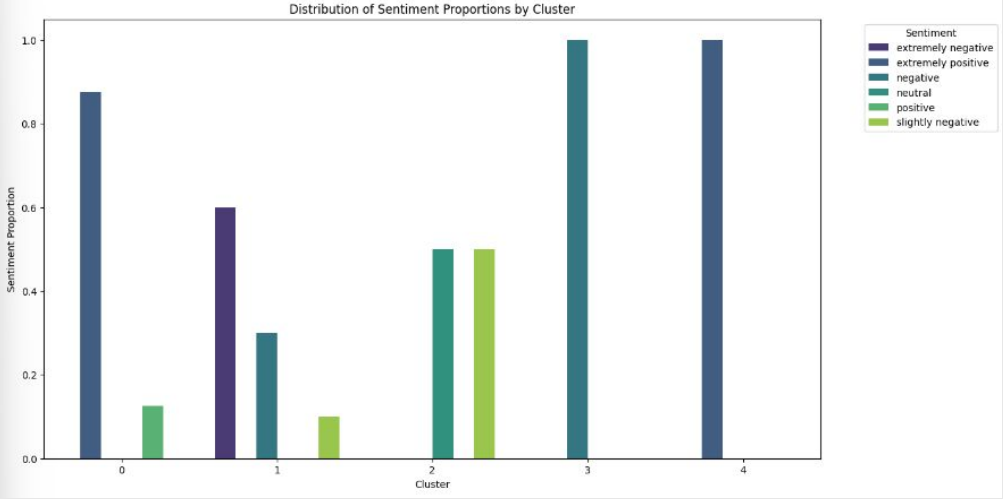


Approach 2: Analysis

- prioritize cluster 1 and 2 for analysis
- more readers have overall extremely positive sentiment towards the novel than negative sentiment
- confirms findings of approach 1







Conclusion

Our initial hypothesis of Lolita being a controversial novel with mixed reviews was correct.

Thank You!