# Text Mining

## A sentiment analysis of Amazon book reviews

Team Autodesk Adoption

Doris Zhang
Emma Yu
Kaki Liu
Spyros Protoulis
Yifu Yan
Xuhui Zhang

# The dataset

| reviewTime | unixReviewTime | summary | overall | reviewText | reviewerName | asin | reviewerID |
|---|---|---|---|---|---|---|---|
| 02 26, 2010 | 1267142400 | "With Wings Like Eagles" Soars! | 5 | Michael Korda's hi | Gurman Singh Bal | 61125369 | A1UBELZ7KJCE4Z |
| 06 27, 2010 | 1277596800 | Scarpetta on the case | 4 | I really thought Sc | Janet M. Gruen "bookworm51" | 006112740X | A1EVO3EVXEQTMC |

**Table Info**

| | |
|---|---|
| Table ID | machinelearning-196501:machineLearningDataset.amazonBooks |
| Table Size | 7.78 GB |
| Number of Rows | 8,898,041 |
| Creation Time | Feb 27, 2018, 3:10:32 PM |
| Last Modified | Feb 27, 2018, 3:10:32 PM |
| Expiration Time | Never Edit |
| Data Location | US |
| Labels | None Edit |

| | |
|---|---|
| reviewTime | STRING |
| unixReviewTime | INTEGER |
| summary | STRING |
| overall | FLOAT |
| reviewText | STRING |
| reviewerName | STRING |
| helpful | INTEGER |
| asin | STRING |
| reviewerID | STRING |

# Sentiment Analysis

Steps:

1. Tokenize documents
2. Stem words and remove stop words
3. Join a dictionary that includes sentiment scores
4. Calculate score

# Step 1: Tokenize Documents

| | asin | reviewerID | overall | reviewText |
|---|---|---|---|---|
| 54751 | 1935627880 | A3E2ET9QG1723S | 5 | What I loved about this book was the way that Chris has bo... |
| 43237 | 0966398130 | A2F92M62KN324B | 5 | I liked it so much I bought a copy for one of my buddies. O... |
| 17312 | 0802723527 | A1ZVELOA9LU4MR | 4 | Disclaimers: I received an e-galley of this book in exchange f... |
| 37938 | 061566069X | A25OT2WVEKP6HR | 5 | I am going to keep this review short... I wasn't planning on r... |
| 26588 | 0062316869 | A2IRQY7MU5RTZ8 | 5 | I enjoyed the book very much and have told several of my r... |
| 725 | 0345485920 | A3OS2OTE09QOOX | 2 | I bought this book thinking it would be something that wou... |
| 33265 | 0380804204 | A3GG2QNXWFA3EK | 5 | I just found Ford two weeks ago, and I can say with a big sm... |
| 39482 | 0739458213 | A1IPN9RNZGJ4BP | 5 | This book grabs your attention starting on page one, plus it... |
| 47208 | 1449361323 | A30B0G0CIOOSQ2 | 5 | As the authors discuss in the preface to this text, the conten... |
| 57146 | B009THFEVA | A17DSQRFTCMRQ3 | 5 | I enjoyed puzzles and quizzes, works my brain and is fun! Gr... |
| 21284 | 1494400626 | A1SH3H8CWM8ECA | 4 | Now I see why this was my sister's favorite book when we w... |
| 26988 | 0131576070 | A1K9IW99EFBZ52 | 5 | I recommend this book with two others: Ed Brodow'sNegoti... |
| 46174 | 1416912045 | A24EON2HWZJVN8 | 5 | Neal SHusterman is such an underrated author, UNWIND w... |
| 7345 | 1450590497 | ASJ3RS87GL3VD | 3 | This story rambled about a young Christian woman who lea... |
| 3498 | 0142410705 | A28GMX5NXCV4OT | 3 | I love John Green, and probably have tremendously high ex... |
| 39568 | 0739458213 | A3IKTWWTPQJNOF | 5 | This book was recommended to me by a professor. I was pr... |

| | asin | reviewerID | overall | word |
|---|---|---|---|---|
| 54751 | 1935627880 | A3E2ET9QG1723S | 5 | what |
| 54751.1 | 1935627880 | A3E2ET9QG1723S | 5 | i |
| 54751.2 | 1935627880 | A3E2ET9QG1723S | 5 | loved |
| 54751.3 | 1935627880 | A3E2ET9QG1723S | 5 | about |
| 54751.4 | 1935627880 | A3E2ET9QG1723S | 5 | this |
| 54751.5 | 1935627880 | A3E2ET9QG1723S | 5 | book |
| 54751.6 | 1935627880 | A3E2ET9QG1723S | 5 | was |
| 54751.7 | 1935627880 | A3E2ET9QG1723S | 5 | the |
| 54751.8 | 1935627880 | A3E2ET9QG1723S | 5 | way |
| 54751.9 | 1935627880 | A3E2ET9QG1723S | 5 | that |
| 54751.10 | 1935627880 | A3E2ET9QG1723S | 5 | chris |
| 54751.11 | 1935627880 | A3E2ET9QG1723S | 5 | has |
| 54751.12 | 1935627880 | A3E2ET9QG1723S | 5 | bound |
| 54751.13 | 1935627880 | A3E2ET9QG1723S | 5 | together |
| 54751.14 | 1935627880 | A3E2ET9QG1723S | 5 | two |
| 54751.15 | 1935627880 | A3E2ET9QG1723S | 5 | love |

# Step 2: Stem Words and Remove Stop Words

| | asin | reviewerID | overall | word |
|---|---|---|---|---|
| 1 | 1935627880 | A3E2ET9QG1723S | 5 | love |
| 2 | 1935627880 | A3E2ET9QG1723S | 5 | book |
| 3 | 1935627880 | A3E2ET9QG1723S | 5 | chris |
| 4 | 1935627880 | A3E2ET9QG1723S | 5 | bind |
| 5 | 1935627880 | A3E2ET9QG1723S | 5 | love |
| 6 | 1935627880 | A3E2ET9QG1723S | 5 | story |
| 7 | 1935627880 | A3E2ET9QG1723S | 5 | separate |
| 8 | 1935627880 | A3E2ET9QG1723S | 5 | generation |
| 9 | 1935627880 | A3E2ET9QG1723S | 5 | bring |
| 10 | 1935627880 | A3E2ET9QG1723S | 5 | overlay |
| 11 | 1935627880 | A3E2ET9QG1723S | 5 | ellie |
| 12 | 1935627880 | A3E2ET9QG1723S | 5 | central |
| 13 | 1935627880 | A3E2ET9QG1723S | 5 | character.ellie |
| 14 | 1935627880 | A3E2ET9QG1723S | 5 | raconteur |
| 15 | 1935627880 | A3E2ET9QG1723S | 5 | diary |
| 16 | 1935627880 | A3E2ET9QG1723S | 5 | agony |

| | word | lexicon |
|---|---|---|
| 1 | a | SMART |
| 2 | a's | SMART |
| 3 | able | SMART |
| 4 | about | SMART |
| 5 | above | SMART |
| 6 | according | SMART |
| 7 | accordingly | SMART |
| 8 | across | SMART |
| 9 | actually | SMART |
| 10 | after | SMART |
| 11 | afterwards | SMART |
| 12 | again | SMART |
| 13 | against | SMART |
| 14 | ain't | SMART |
| 15 | all | SMART |
| 16 | allow | SMART |

Showing 1 to 17 of 1,149 entries

| | stem | term |
|---|---|---|
| 182 | abolish | abolishes |
| 183 | abolish | abolishing |
| 184 | abolition | abolitions |
| 185 | abolitionist | abolitionists |
| 186 | A-bomb | A-bombs |
| 187 | abominate | abominated |
| 188 | abominate | abominates |
| 189 | abominate | abominating |
| 190 | abomination | abominations |
| 191 | aboriginal | aboriginals |
| 192 | Aborigine | Aborigines |
| 193 | abort | aborted |
| 194 | abort | aborting |
| 195 | abort | aborts |
| 196 | abortifacient | abortifacients |
| 197 | abortion | abortions |

Showing 183 to 199 of 41,760 entries

# Step 3: Join a Dictionary (lexicon) with Sentiment Score

| | asin | reviewerID | overall | word | sentiment | lexicon | score |
|---|---|---|---|---|---|---|---|
| 1 | 1935627880 | A3E2ET9QG1723S | 5 | love | NA | AFINN | 3 |
| 2 | 1935627880 | A3E2ET9QG1723S | 5 | book | NA | NA | NA |
| 3 | 1935627880 | A3E2ET9QG1723S | 5 | chris | NA | NA | NA |
| 4 | 1935627880 | A3E2ET9QG1723S | 5 | bind | NA | NA | NA |
| 5 | 1935627880 | A3E2ET9QG1723S | 5 | love | NA | AFINN | 3 |
| 6 | 1935627880 | A3E2ET9QG1723S | 5 | story | NA | NA | NA |
| 7 | 1935627880 | A3E2ET9QG1723S | 5 | separate | NA | NA | NA |
| 8 | 1935627880 | A3E2ET9QG1723S | 5 | generation | NA | NA | NA |
| 9 | 1935627880 | A3E2ET9QG1723S | 5 | bring | NA | NA | NA |
| 10 | 1935627880 | A3E2ET9QG1723S | 5 | overlay | NA | NA | NA |
| 11 | 1935627880 | A3E2ET9QG1723S | 5 | ellie | NA | NA | NA |
| 12 | 1935627880 | A3E2ET9QG1723S | 5 | central | NA | NA | NA |
| 13 | 1935627880 | A3E2ET9QG1723S | 5 | character.ellie | NA | NA | NA |
| 14 | 1935627880 | A3E2ET9QG1723S | 5 | raconteur | NA | NA | NA |
| 15 | 1935627880 | A3E2ET9QG1723S | 5 | diary | NA | NA | NA |
| 16 | 1935627880 | A3E2ET9QG1723S | 5 | agony | NA | NA | NA |

| | word | sentiment | lexicon | score |
|---|---|---|---|---|
| 1 | abandon | NA | AFINN | -2 |
| 2 | abandoned | NA | AFINN | -2 |
| 3 | abandons | NA | AFINN | -2 |
| 4 | abducted | NA | AFINN | -2 |
| 5 | abduction | NA | AFINN | -2 |
| 6 | abductions | NA | AFINN | -2 |
| 7 | abhor | NA | AFINN | -3 |
| 8 | abhorred | NA | AFINN | -3 |
| 9 | abhorrent | NA | AFINN | -3 |
| 10 | abhors | NA | AFINN | -3 |
| 11 | abilities | NA | AFINN | 2 |
| 12 | ability | NA | AFINN | 2 |
| 13 | aboard | NA | AFINN | 1 |
| 14 | absentee | NA | AFINN | -1 |
| 15 | absentees | NA | AFINN | -1 |
| 16 | absolve | NA | AFINN | 2 |

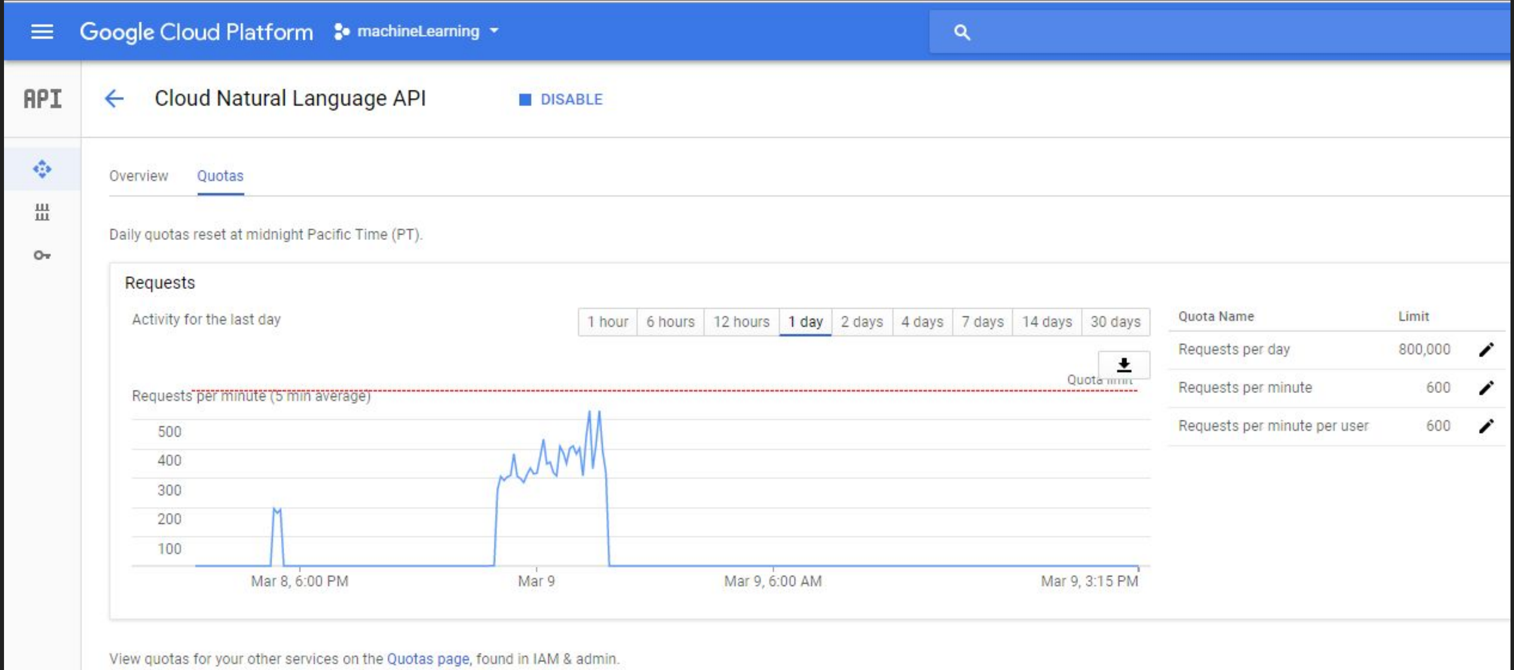
Showing 1 to 17 of 2,476 entries

# Step 4: Calculate Sentiment Score for Each Review

| | asin | reviewerID | sentiment_score | overall |
|---|---|---|---|---|
| 1 | 0026009102 | A2AYSFGUP5VTY3 | 1.6666667 | 4 |
| 2 | 0061730327 | A3IFEXK0M52J2L | 0.4166667 | 4 |
| 3 | 0061994316 | A2FJ5NWS5LQ9LN | 1.3076923 | 4 |
| 4 | 0061998974 | A1AQDSTEBI8BB2 | -1.2500000 | 5 |
| 5 | 0062026879 | AHUT55E980RDR | 2.3333333 | 5 |
| 6 | 006203619X | A3S0XEPOFOCB5W | 1.4166667 | 5 |
| 7 | 0062316869 | A2IRQY7MU5RTZ8 | 2.0000000 | 5 |
| 8 | 0131576070 | A1K9IW99EFBZ52 | 0.1428571 | 5 |
| 9 | 0142410705 | A28GMX5NXCV4OT | 1.0000000 | 3 |
| 10 | 0143114808 | A2ZQG435OSH2GZ | -0.2826087 | 4 |
| 11 | 0152010661 | A1XUD5NNV5LOSH | 2.5000000 | 5 |
| 12 | 0307594017 | A19JA54QH6N041 | 2.0000000 | 4 |
| 13 | 0307712842 | A310BJW6IJQNPK | 1.6000000 | 5 |
| 14 | 0310330513 | A34RH1RQBHK1P1 | 1.0000000 | 4 |
| 15 | 0312614578 | A2MF4TISBBQT5A | 1.1153846 | 4 |
| 16 | 0312890796 | A3EX36SNRYD5VL | 2.5000000 | 1 |

# Google Natural Language and BigQuery API

- BigQuery: The dataset is too big to be analyzed on a PC
- Google Cloud NLP provides an automated way to calculate sentiment scores



# Word Clouds

Negative single word

# Word Clouds

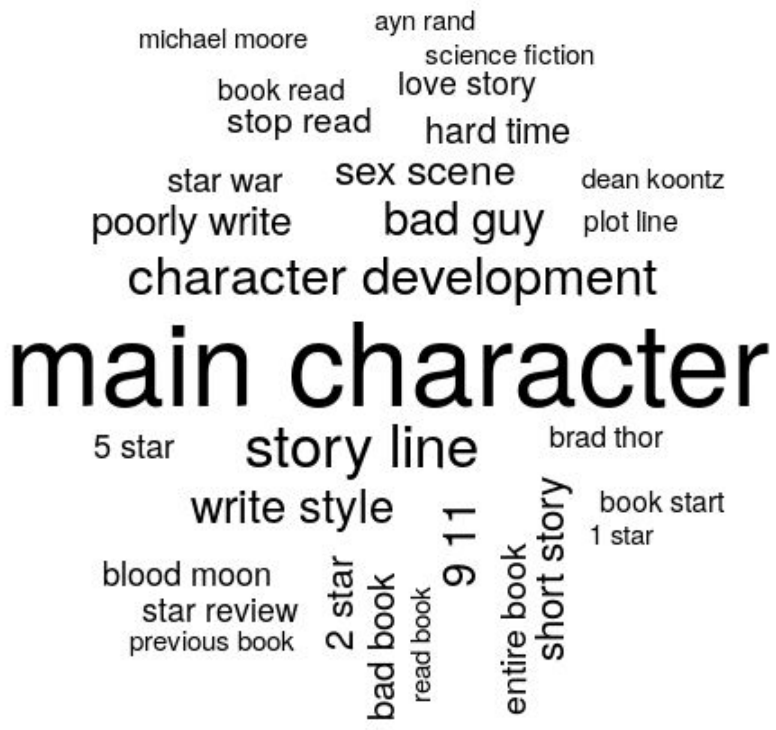
Neutral single words

# Word Clouds

Positive single words:

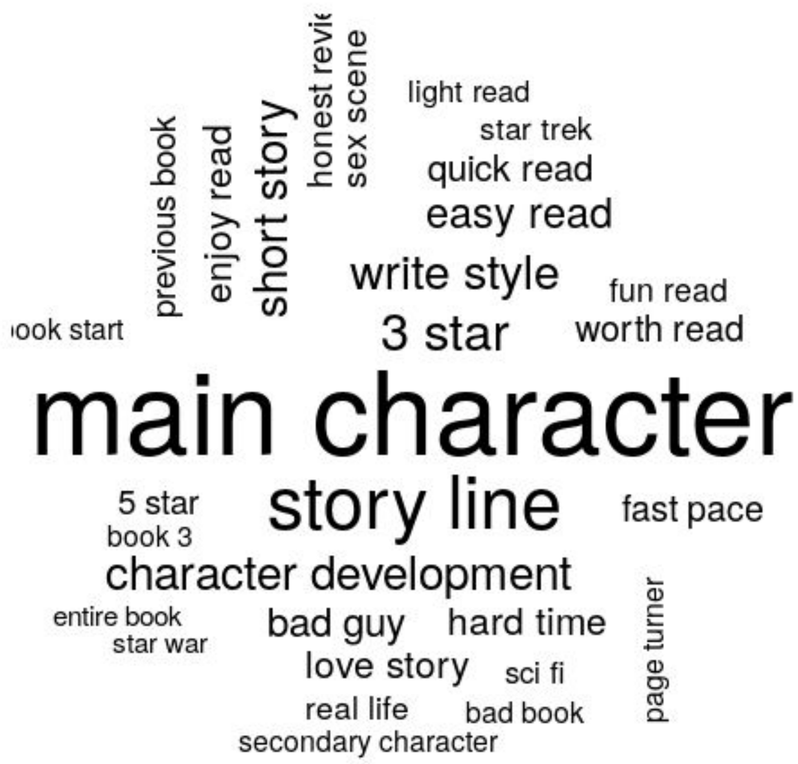
# Word Clouds

Negative bigram:



# Word Clouds

Neutral bigram:

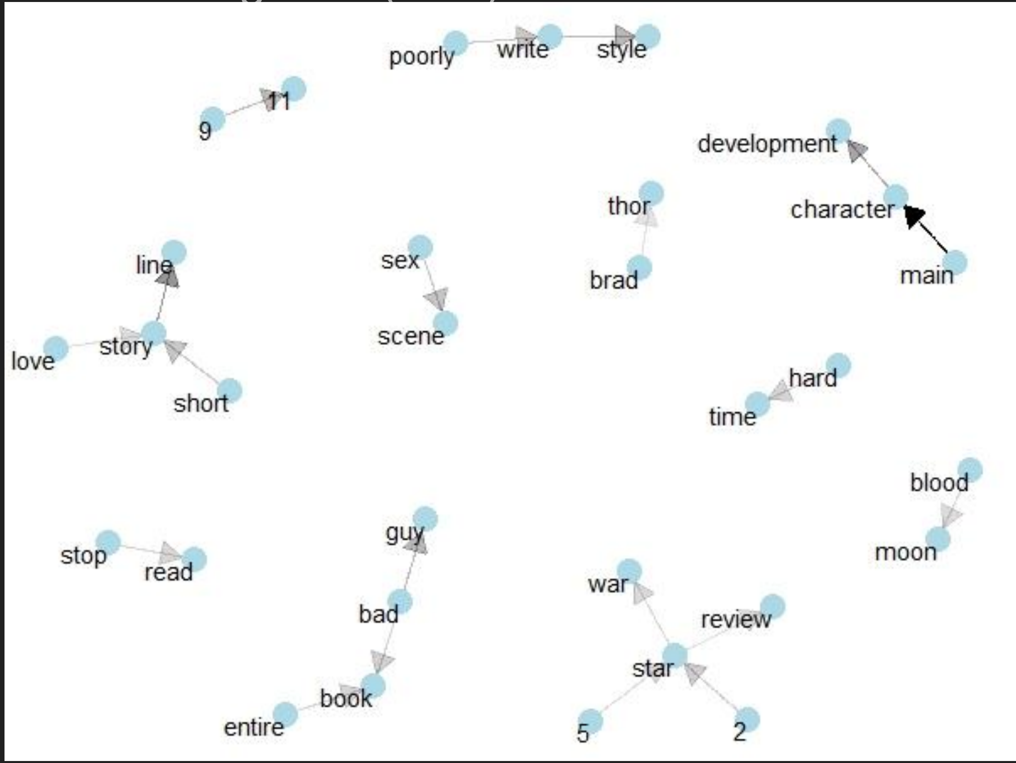


# Word Clouds

Positive bigram:

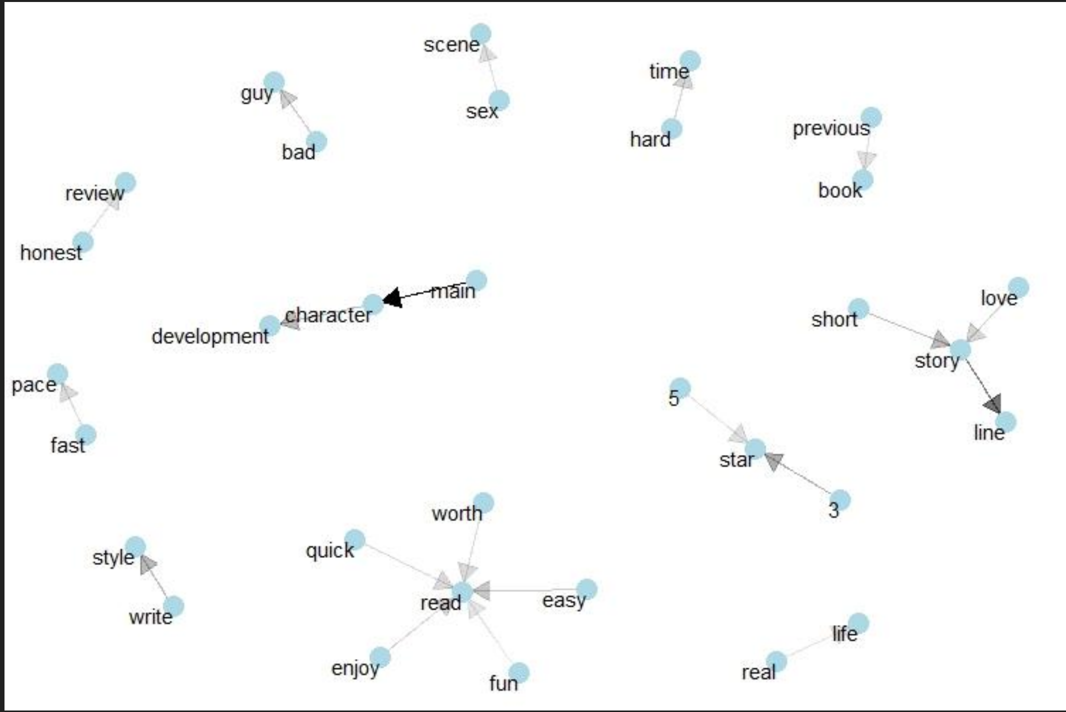
# Word Network

Negative (1-2*) reviews



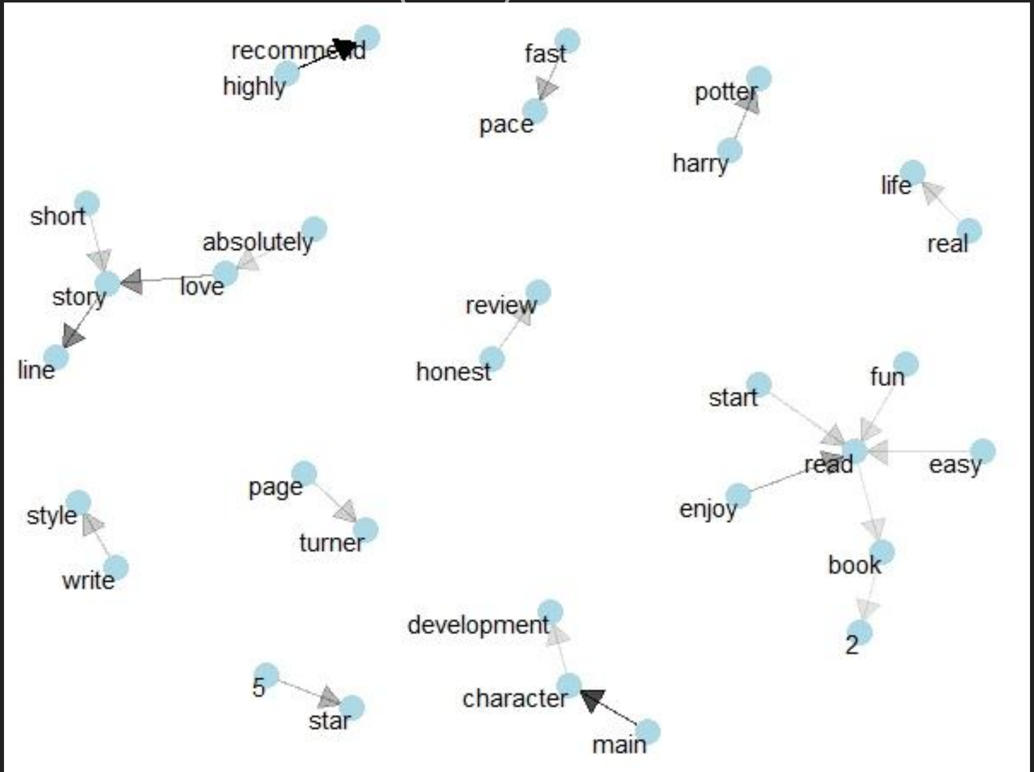
# Word Network

Neutral (3*) reviews



# Word Network

Positive (4-5*) reviews



# Web scraping Amazon for prices

We want check if we can use price of books in our model, so we scrapes books' price from Amazon.

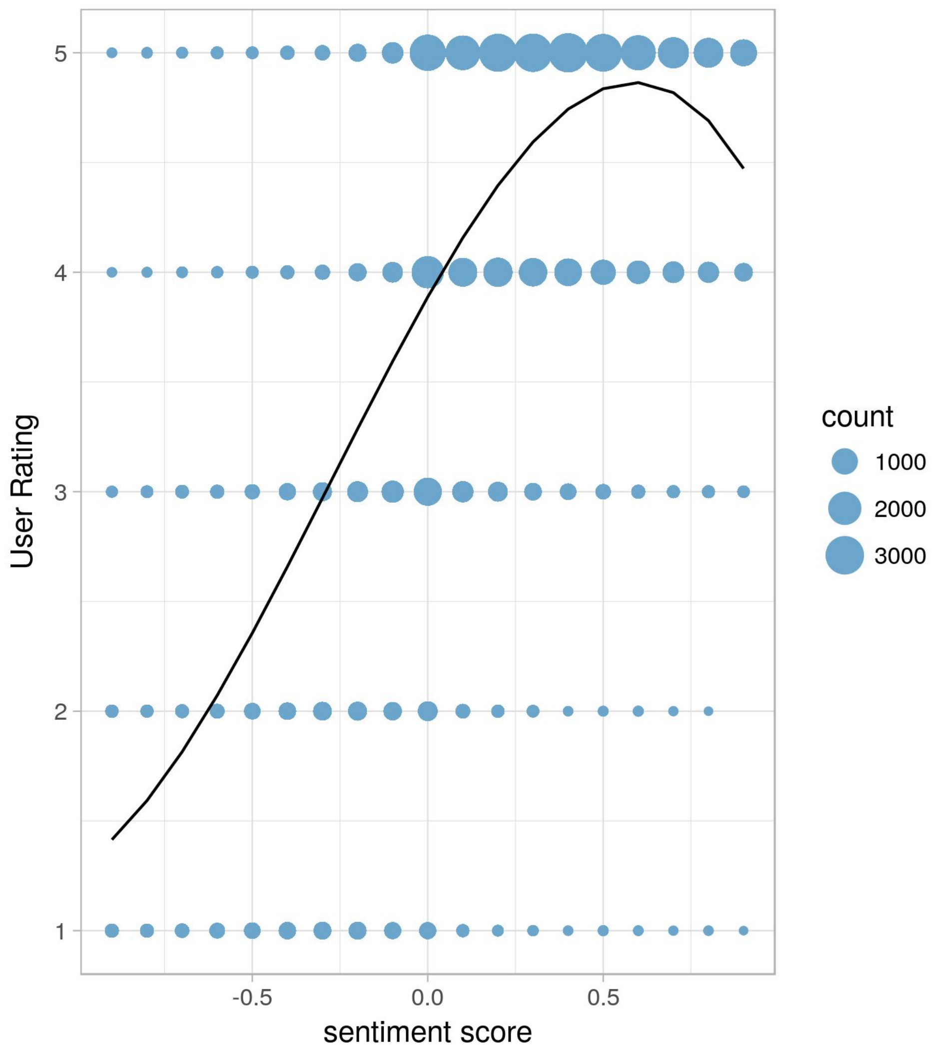
Filter

| | title | price | asin |
|---|---|---|---|
| 1 | Red-Headed Stepchild (Sabina Kane, Book 1) | 7.852500 | 0316037761 |
| 2 | Conservative Comebacks to Liberal Lies: Issue by Issue Resp... | 12.147241 | 0977227901 |
| 3 | Jordyn and the Caverns of Gloom: A Daemon Hunter Novel ... | 12.288000 | 1491291125 |
| 4 | Hello Kitty Must Die | 6.688571 | 1935562029 |
| 5 | Song of the Summer King | 14.104211 | 0985805803 |
| 6 | Lake Caerwych | 6.660000 | 061554228X |
| 7 | The Tombs: A Fargo Adventure (A Sam and Remi Fargo Adv... | 38.985152 | 0399159266 |
| 8 | Love's A Bitch (LSDV Productions) | 3.628182 | B00H0BVTVG |
| 9 | Night Prey | 7.492500 | 0821736612 |
| 10 | Return to Sullivan's Island (Lowcountry Tales) | 9.225312 | 0061438456 |
| 11 | Race With the Devil: My Journey from Racial Hatred to Ratio... | 12.541333 | 161890065X |
| 12 | Reunited for the Holidays (Love Inspired, Texas Twins) | 5.431364 | 037387782X |
| 13 | Rest Not in Peace (Chronicles of Hugh de Singleton, Surgeon) | 10.557586 | 1782640088 |
| 14 | The Journals of Kara and Jason | 2.414231 | B00BP9GRCK |
| 15 | Miss Pettigrew Lives for a Day (Persephone Classics) | 9.563793 | 190646202X |
| 16 | 7: An Experimental Mutiny Against Excess | 12.355000 | 1433672960 |

Showing 1 to 17 of 966 entries

# A polynomial model

- 3rd degree polynomial model including to predict the rating a user will give a book
- Dependent variables tested: Sentiment score, Sentiment*Magnitude, Price
- Sentiment only gave us a better fit, so we dropped the other two.



5
4
3
2
1
User Rating
-0.5
0.0
0.5
sentiment score
count
1000
2000
3000

# Mean user and book rating “pull” factors

Inspired by the Netflix recommendation system prize winners
We apply two “pulling” factors on our prediction from the polynomial model:

- One pulls towards the existing mean rating for the book being reviewed
- One pulls towards the mean rating the reviewer has given to other books
- The more reviews we have for the book or the user in question, the greater the weight of the pull

y2 = y1 + bookPullWeight * (y1 - bookMeanRating) + userPullWeight * (y1 - userMeanRating)

## Results on test data

- Mean residual: 0.625
- Our predicted rating is within 1 point from the actual rating 90.8% of the time

## Applicability

- Gauge a rating score from text on book discussion forums which do not include star ratings

# Thank You