
GE3104

Data is Beautiful: Visualization in the Humanities

Term paper

# Common vocabularies and contrasting adjectives on Amazon Book Review dataset with correlation to positive and negative reviews

LI Wang Yau

wangyauli2

*2495 words*

# Contents

# 1. Introduction

In the era of big data, various online text materials can be viewed as corpora which provide important information for studying language usages and customer behavior. In this study, five key findings of word usages across the high-scored and low-scored portions of the Amazon Book Review Dataset (McAuley, 2015) was examined. Wordlists and visualization graphs were constructed to reflect the similarities and differences in language usage.

By locating contrasting adjectives, common vocabularies and purchasing patterns, it is hoped that this research can provide useful for other researchers in linguistics, business, and marketing aspects. Strengths and weaknesses of the visualization methods, including Antconc, Gephi and an extra visualization tool “Rawsgraph” are also discussed. Results will be released online for other researchers to carry out further studies.

# 2. Data

## 2.1. Source of data

All the data was selected form an open source Amazon Book Review Dataset (McAuley, 2015). The original dataset contains around 8.9 million book reviews with approximately 1.5 billion words from Amazon spanning May 1996 - July 2014.

## 2.2. Data preparation

Two similar-sized contracting subsets corresponding to positive and negative reviews were selected and further processed by using Linux commands, such as “tr”, “grap”, and “sed” etc.


*Figure 1. Illustration of selected subsets*

Dataset 1(Positive reviews) is a collection of 300,000 reviews(approximately 46 million tokens) scoring 5.0 in overall rating. Dataset 2(Negative reviews) is a collection of 300,000 reviews(approximately 43 million tokens) scoring 1.0-2.0 in overall rating. Reviews with higher helpfulness score were selected first so as to ensure selected entities are most representative of this set.

A filter list[1] based on the default setting in worditout was used as stop list in AntConc processing. Part-of-Speech tagging was performed by TagAnt(Anthony, 2015).

[1] It consists of pronouns, function words except for negation markers “not” and conjunctions “but”

# 3. Methodology

## 3.1. Research questions

The research questions for this project are as follow: 1) What is the exact list of **adjectives** that associated with high and low scored reviews respectively? 2) What are the **common vocabularies** that usually mentioned in reviews? 3) Is there any particular **purchasing pattern** that can be revealed by the time-stamped reviews? A graphic illustration of research questions is shown in Figure 2.

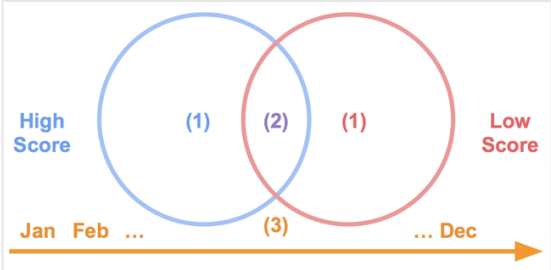



*Figure 2. Illustration of research questions with datasets*

## 3.2. Visualization methods

| Functions \ Software | Voyant Tools | Excel | Ant-Conc | Sketch Engine | Google Ngram Viewer | Gephi |
|---|---|---|---|---|---|---|
| **Available dataset(s)** | | | | | | |
| External Corpus / Dataset | ✓ | ✓ | ✓ | - | - | ✓ |
| **Analysis** | | | | | | |
| Word frequency / Wordlist | ✓ | ✓ | ✓ | ✓ | - | - |
| Keyword list (from 2 dataset) | - | - | ✓ | ✓ | - | - |
| Adjectives in contracting environments | ✓ | - | ✓ | ✓ | - | - |
| Common words in both environments | - | - | ✓ | ✓ | - | - |
| **Visualization** | | | | | | |
| Contrasting / Co-occurring relations | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Focus / Filtering | ✓ | ✓ | - | - | ✓ | ✓ |
| Temporal changes | ✓ | ✓ | - | - | ✓ | ✓ |

*Figure 3. Comparison of software*

AntConc was the most suitable software for analysis involving external datasets and contrasting relationships(see Figure 3). By using AntConc, word lists and keyword lists can be constructed for the two contrasting datasets respectively. The keyword lists could show

contrasting adjectives while common vocabularies for book reviews will occur frequently in both word lists. Further analysis of co-occurring relations between adjective and their modified nouns were done by using collocate function. These answer research questions (1) and (2).

Gephi was also selected to show internal relations and communities. The top 60 keywords from AntConc analysis would then be used as a dictionary in a python script to generate a .csv file for Gephi illustration. In this part, adjectives and common vocabularies would be nodes, whenever they co-occur in the same review would increase their edge weight by 1. By using colors and filters to illustrate word communities, Gephi graphs can support and clearly illustrate the above-mentioned results.

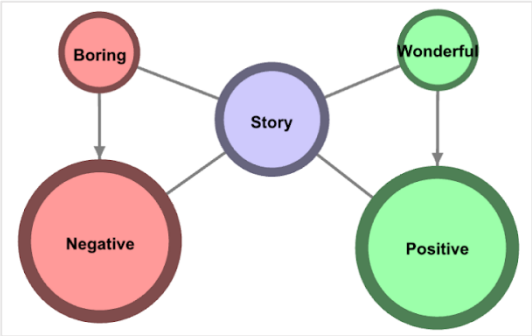



*Figure 4. Gephi network*

### 3.3. Extra tools

Some extra tools were used to compensate for the limitations of the previous tools. Part of speech(POS) and lemma tagging was done by TagAnt(Anthony, 2015). This helps separating different word classes of the types on a worklist and allow us carry out analysis by word class. The outputs generated by AntConc and TagAnt can then be taken as parameters and identify their internal relations.

By using Rawgraphs, we can clearly visualize their relations to audience by a various types of graphs, including “sunburst”, “circular dendrogram hierarchy”, and “treemap”. Bar chart will also be used to overcome Gephi’s limitations and reveal the purchasing behaviour of readers. And thus answers the last research question.

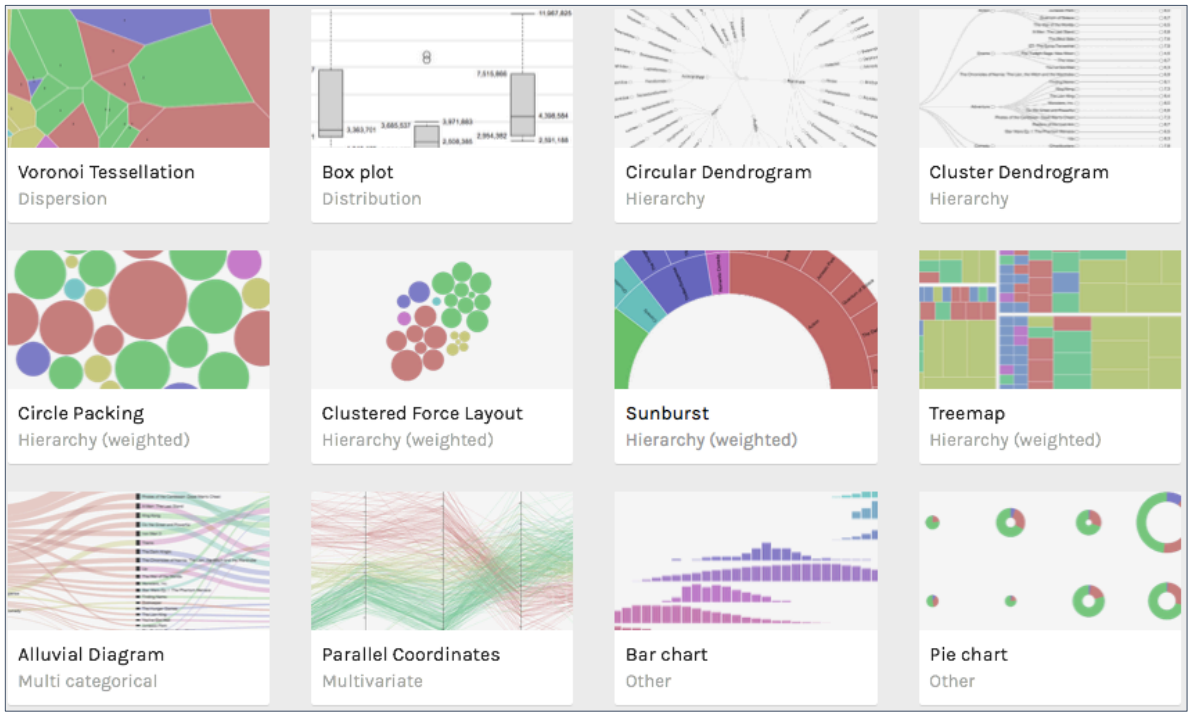



*Figure 5. Tools on Rawgraphs*

# 4. Discussion

## 4.1. Visualizations results

This section will cover five key findings of the dataset, including similarities and differences across the two contrasting datasets. The research questions will be answered by these findings. Some unexpected extra findings regarding other word classes are also mentioned.

### 4.1.1. Common vocabularies

Wordlists were generated from the positive and negative datasets respectively. It is found that common vocabularies exist.

| Positive - Rank | Frequency | % | Type |
|---|---|---|---|
| 1 | 531719 | 2.0% | book |
| 2 | 266401 | 1.0% | but |
| 3 | 228293 | 0.9% | not |
| 4 | 227556 | 0.9% | read |
| 5 | 216467 | 0.8% | t |
| 6 | 202656 | 0.8% | one |
| 7 | 180951 | 0.7% | all |
| 8 | 173359 | 0.7% | story |
| 9 | 160487 | 0.6% | about |
| 10 | 142677 | 0.5% | more |
| 11 | 120757 | 0.5% | love |
| 12 | 116936 | 0.4% | like |
| 13 | 113458 | 0.4% | just |
| 14 | 111134 | 0.4% | very |
| 15 | 102905 | 0.4% | well |
| 16 | 100740 | 0.4% | great |
| 17 | 98615 | 0.4% | life |
| 18 | 95465 | 0.4% | books |
| 19 | 92269 | 0.4% | some |
| 20 | 92101 | 0.4% | time |
| 21 | 84590 | 0.3% | characters |
| 22 | 84497 | 0.3% | good |
| 23 | 83063 | 0.3% | first |
| 24 | 81367 | 0.3% | reading |
| 25 | 79647 | 0.3% | really |
| 26 | 78995 | 0.3% | other |
| 27 | 74944 | 0.3% | author |
| 28 | 73814 | 0.3% | many |
| 29 | 73438 | 0.3% | much |
| 30 | 72992 | 0.3% | way |

| Negative - Rank | Frequency | % | Type |
|---|---|---|---|
| 1 | 566027 | 2.2% | book |
| 2 | 357474 | 1.4% | t |
| 3 | 355256 | 1.4% | not |
| 4 | 351164 | 1.3% | but |
| 5 | 191842 | 0.7% | one |
| 6 | 184322 | 0.7% | read |
| 7 | 182656 | 0.7% | just |
| 8 | 181060 | 0.7% | about |
| 9 | 175708 | 0.7% | like |
| 10 | 175605 | 0.7% | all |
| 11 | 155471 | 0.6% | story |
| 12 | 132861 | 0.5% | more |
| 13 | 125221 | 0.5% | no |
| 14 | 109962 | 0.4% | author |
| 15 | 107559 | 0.4% | really |
| 16 | 102021 | 0.4% | some |
| 17 | 101963 | 0.4% | very |
| 18 | 99404 | 0.4% | books |
| 19 | 98657 | 0.4% | characters |
| 20 | 97369 | 0.4% | time |
| 21 | 95515 | 0.4% | good |
| 22 | 95340 | 0.4% | much |
| 23 | 93771 | 0.4% | even |
| 24 | 90579 | 0.3% | because |
| 25 | 89645 | 0.3% | other |
| 26 | 88068 | 0.3% | don |
| 27 | 83945 | 0.3% | first |
| 28 | 80391 | 0.3% | only |
| 29 | 79054 | 0.3% | reading |
| 30 | 72779 | 0.3% | didn |

*Figure 5. Common vocabularies across reviews*

In figure 5, some words were shown very frequently in both corpora with similar percentages. These words include "*book(s), read-ing, story, author, time, character(s)*". Based on the similar percentage, these vocabularies are relatively sentimental neutral. They probably imply a list of important features which customers often comment on when they evaluate the products.

| Positive - Rank | Frequency | % | Type |
|---|---|---|---|
| 1 | 531719 | 2.0% | book |
| **2** | **266401** | **1.0%** | **but** |
| **3** | **228293** | **0.9%** | **not** |
| 4 | 227556 | 0.9% | read |
| **5** | **216467** | **0.8%** | **t** |
| 6 | 202656 | 0.8% | one |
| 7 | 180951 | 0.7% | all |
| 8 | 173359 | 0.7% | story |
| 9 | 160487 | 0.6% | about |
| 10 | 142677 | 0.5% | more |
| 11 | 120757 | 0.5% | love |
| **12** | **116936** | **0.4%** | **like** |
| 13 | 113458 | 0.4% | just |
| 14 | 111134 | 0.4% | very |
| 15 | 102905 | 0.4% | well |
| 16 | 100740 | 0.4% | great |
| 17 | 98615 | 0.4% | life |
| 18 | 95465 | 0.4% | books |
| 19 | 92269 | 0.4% | some |
| 20 | 92101 | 0.4% | time |
| 21 | 84590 | 0.3% | characters |
| **22** | **84497** | **0.3%** | **good** |
| 23 | 83063 | 0.3% | first |
| 24 | 81367 | 0.3% | reading |
| 25 | 79647 | 0.3% | really |
| 26 | 78995 | 0.3% | other |
| 27 | 74944 | 0.3% | author |
| 28 | 73814 | 0.3% | many |
| 29 | 73438 | 0.3% | much |
| 30 | 72992 | 0.3% | way |

| Negative - Rank | Frequency | % | Type |
|---|---|---|---|
| 1 | 566027 | 2.2% | book |
| **2** | **357474** | **1.4%** | **t** |
| **3** | **355256** | **1.4%** | **not** |
| **4** | **351164** | **1.3%** | **but** |
| 5 | 191842 | 0.7% | one |
| 6 | 184322 | 0.7% | read |
| 7 | 182656 | 0.7% | just |
| 8 | 181060 | 0.7% | about |
| **9** | **175708** | **0.7%** | **like** |
| 10 | 175605 | 0.7% | all |
| 11 | 155471 | 0.6% | story |
| 12 | 132861 | 0.5% | more |
| 13 | 125221 | 0.5% | no |
| 14 | 109962 | 0.4% | author |
| 15 | 107559 | 0.4% | really |
| 16 | 102021 | 0.4% | some |
| 17 | 101963 | 0.4% | very |
| 18 | 99404 | 0.4% | books |
| 19 | 98657 | 0.4% | characters |
| 20 | 97369 | 0.4% | time |
| **21** | **95515** | **0.4%** | **good** |
| 22 | 95340 | 0.4% | much |
| 23 | 93771 | 0.4% | even |
| 24 | 90579 | 0.3% | because |
| 25 | 89645 | 0.3% | other |
| 26 | 88068 | 0.3% | don |
| 27 | 83945 | 0.3% | first |
| 28 | 80391 | 0.3% | only |
| 29 | 79054 | 0.3% | reading |
| 30 | 72779 | 0.3% | didn |

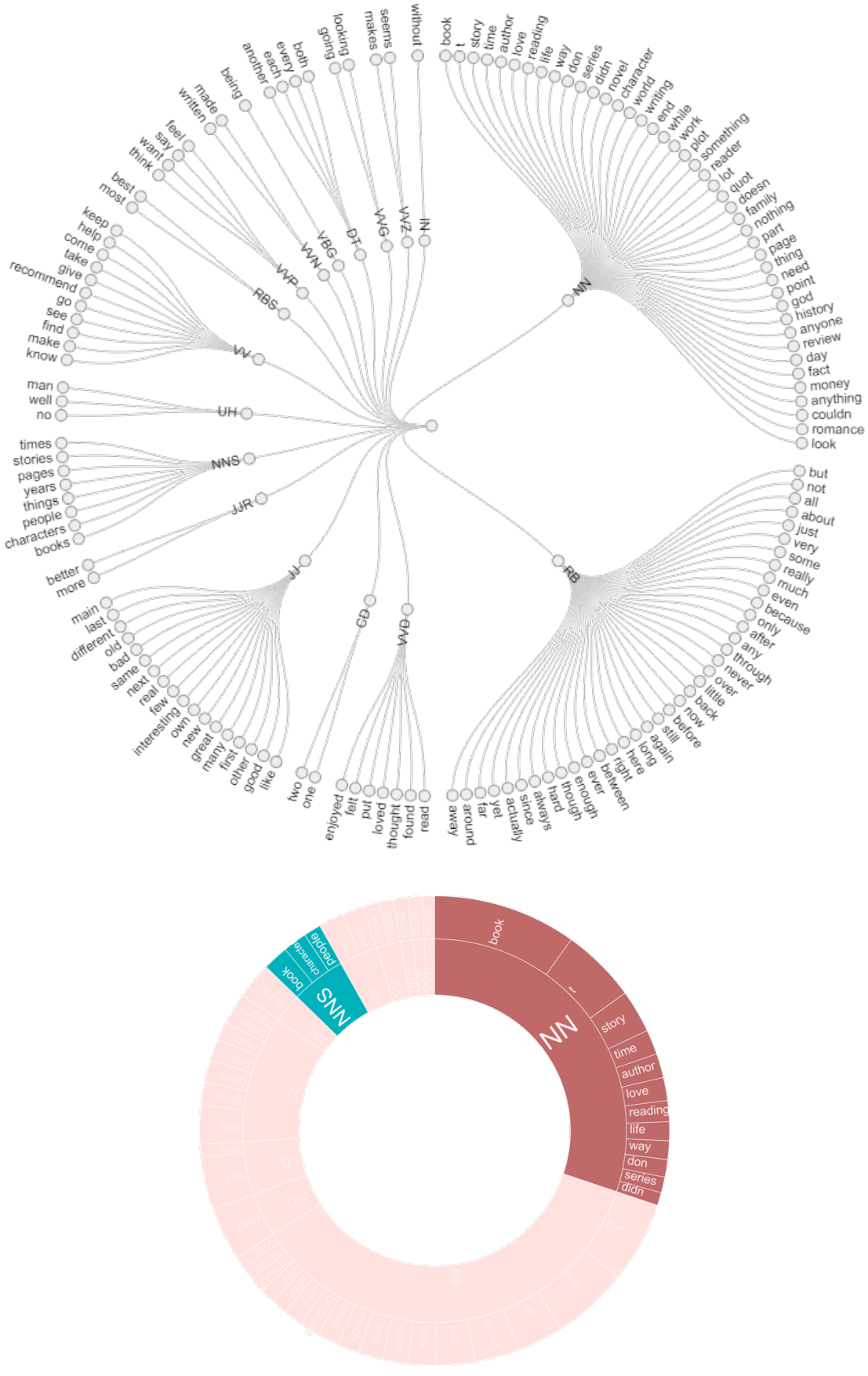
*Figure 6. Common vocabularies across reviews*

In figure 6, some words were also shown very frequently in both corpora but differ slightly in percentages. These words include "*but, not, t*[2]". The percentage in positive wordlist indicates the baseline which people normally use contacting conjunctions and negation when commenting. While review text with negative sentiment has these three items in reverse order, and it is correlated with significantly higher percentage of contraction "t".

But surprisingly, it is found that some words which intuitively connected with "positive" comments were shown in higher percentages in negative reviews. One possible explanation is that people tends to use stronger adjectives like "*great*" or "*amazing*" when they would like to comment positively. With these taken away some portion of the word usage, "*good*" in review text may just mean "fair" which referring to the satisfactory degree below average.

[2] From words such as "don't, wasn't, were't" due to segmentation error.

*Figure 7. Common vocabularies in all reviews (group by POS)*



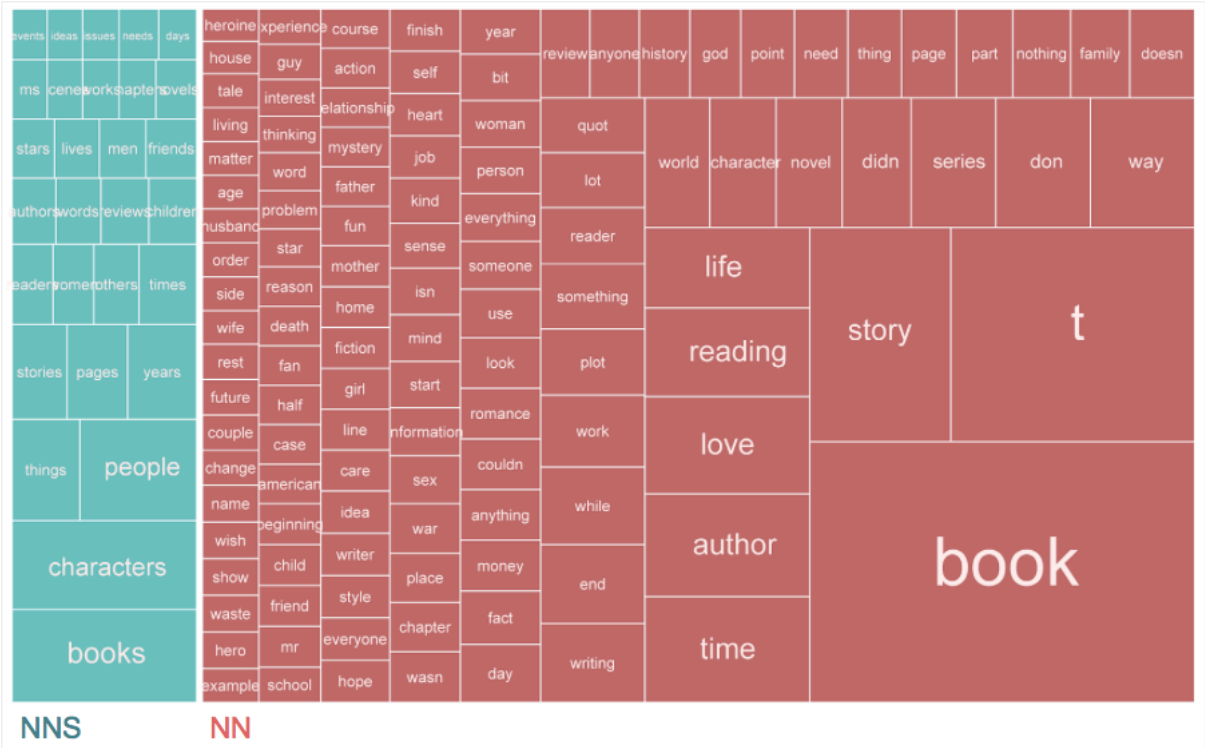



*Figure 8. Common vocabularies in all reviews*

A wordlist was generated from all review text. Figure 7 was generated by using the output from AntConc. By focusing only on the highlighted part, we obtained a treemap in figure 8 which showing top 150 nouns with size corresponding to its frequency. This tree map again proofed that some features are important to customer's opinion on books. Two specific groups of words were identified in figure 9.

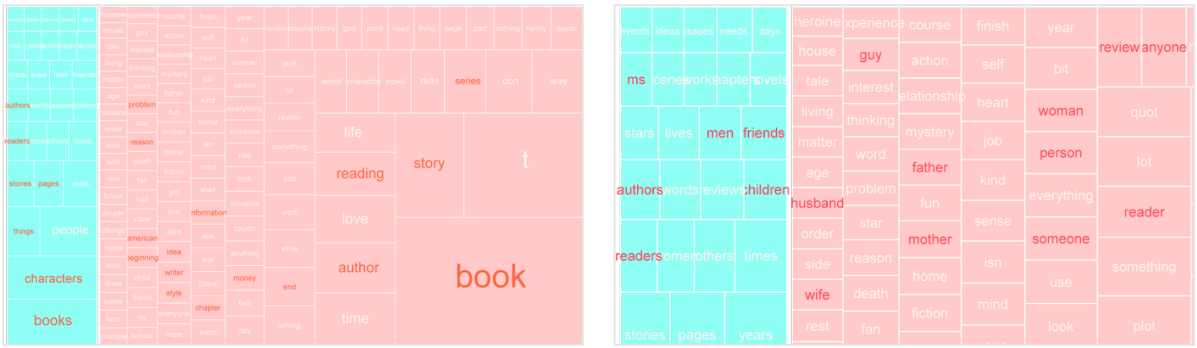



*Figure 9. Features(left) and personal expressions(right)*

One group of words includes features such as "*book, story, style…*" which again confirmed the previous findings. Another group of words are non-context related personal expressions. Only personal recommendations will account for the overall high frequency across different genre, and thus this data shows recommendations from related groups will highly affect reader's choice.

### 4.1.2. Keyword lists comparison

| POS - Rank | % | Keyness | POS | Lemma |
|---|---|---|---|---|
| 1 | 0.90% | 23704.39 | JJ | great |
| 2 | 0.47% | 23105.97 | VVD | love |
| 3 | 1.07% | 21560.81 | NN | love |
| 4 | 0.28% | 18489.03 | RB | highly |
| 5 | 0.88% | 16961.83 | NN | life |
| 6 | 0.24% | 14691.25 | JJ | wonderful |
| 7 | 0.25% | 12347.2 | VV | wait |
| 8 | 0.19% | 11749.65 | JJ | amazing |
| 9 | 0.20% | 10942.53 | JJ | excellent |
| 10 | 0.23% | 10650.33 | NN | heart |
| 11 | 0.24% | 8490.86 | JJ | easy |
| 12 | 0.92% | 8351.11 | UH | well |
| 13 | 0.42% | 6811.14 | DT | each |
| 14 | 0.34% | 6344.97 | NN | family |
| 15 | 0.52% | 5716.27 | NN | world |
| 16 | 0.35% | 5633.97 | DT | both |
| 17 | 0.27% | 5613.8 | VVD | enjoy |
| 18 | 0.10% | 5444.76 | NN | journey |
| 19 | 0.41% | 5238.22 | RBS | well |
| 20 | 0.16% | 5209.68 | JJ | favorite |
| 21 | 0.15% | 5130.83 | JJ | beautiful |
| 22 | 0.19% | 5093.51 | NN | fun |
| 23 | 0.36% | 4937.7 | VV | recommend |
| 24 | 0.11% | 4866.3 | VVD | recommend |
| 25 | 0.15% | 4829.86 | JJ | perfect |
| 26 | 0.08% | 4598.1 | JJ | fantastic |
| 27 | 0.06% | 4486.69 | RB | beautifully |
| 28 | 0.07% | 4095.38 | JJ | awesome |
| 29 | 0.35% | 4042.45 | JJ | next |
| 30 | 0.20% | 4028.55 | NN | job |
| 31 | 0.18% | 4001.33 | NNS | life |
| 32 | 0.17% | 3866.44 | RB | definitely |
| 33 | 0.09% | 3861 | JJ | fascinating |
| 34 | 0.35% | 3506.4 | VVD | put |
| 35 | 0.28% | 3401.22 | RB | always |
| 36 | 0.07% | 3277.19 | VVZ | provide |
| 37 | 0.09% | 3209.27 | JJ | unique |
| 38 | 0.54% | 3126.46 | JJ | new |
| 39 | 0.07% | 3085.73 | VVZ | bring |
| 40 | 0.08% | 3070.72 | VV | thank |
| 41 | 0.08% | 3047.87 | NN | adventure |
| 42 | 0.41% | 3008.78 | JJ | own |
| 43 | 0.14% | 3008.58 | VVZ | give |
| 44 | 0.27% | 3003.64 | VV | help |
| 45 | 0.22% | 2951.07 | NN | war |
| 46 | 0.13% | 2904.72 | NN | tale |
| 47 | 0.03% | 2891.94 | JJ | delightful |
| 48 | 0.05% | 2816.33 | JJ | hooked |
| 49 | 0.18% | 2797.48 | NN | home |
| 50 | 0.03% | 2637.37 | VVG | refresh |

| NEG. Rank | % | Keyness | POS | Lemma |
|---|---|---|---|---|
| 1 | 3.01% | 43413.92 | NN | t |
| 2 | 2.99% | 35435.24 | RB | not |
| 3 | 0.24% | 25790.49 | JJ | boring |
| 4 | 0.61% | 25599.81 | NN | didn |
| 5 | 0.18% | 22580.06 | NN | waste |
| 6 | 1.54% | 20339.11 | RB | just |
| 7 | 1.06% | 20119.09 | UH | no |
| 8 | 2.96% | 16833.84 | RB | but |
| 9 | 0.74% | 15397.15 | NN | don |
| 10 | 1.48% | 15313.47 | JJ | like |
| 11 | 0.35% | 14620.4 | NN | nothing |
| 12 | 0.37% | 14347.21 | JJ | bad |
| 13 | 0.44% | 14060.72 | NN | plot |
| 14 | 0.22% | 13220.42 | NN | finish |
| 15 | 0.28% | 11955.45 | NN | money |
| 16 | 0.09% | 11287.93 | RB | poorly |
| 17 | 0.18% | 10996.92 | VVD | seem |
| 18 | 0.08% | 9901.03 | JJ | disappointing |
| 19 | 0.13% | 9391.5 | RB | unfortunately |
| 20 | 0.13% | 9227.59 | RB | sorry |
| 21 | 0.36% | 9180.39 | NNS | page |
| 22 | 0.93% | 8702.15 | NN | author |
| 23 | 0.22% | 8600.43 | NNS | review |
| 24 | 0.21% | 8489.62 | RB | maybe |
| 25 | 0.09% | 8414.9 | JJ | stupid |
| 26 | 0.08% | 8410.44 | VVG | annoy |
| 27 | 0.19% | 8028.79 | VVN | disappoint |
| 28 | 0.07% | 7961 | JJ | ridiculous |
| 29 | 0.61% | 7850.47 | RB | any |
| 30 | 0.22% | 7709.55 | NN | sex |
| 31 | 0.49% | 7431.86 | NN | character |
| 32 | 0.23% | 7286.2 | NN | wasn |
| 33 | 0.76% | 7177.66 | RB | because |
| 34 | 0.10% | 7163.89 | VVN | suppose |
| 35 | 0.21% | 7123.09 | RB | instead |
| 36 | 0.07% | 6641.21 | NN | disappointment |
| 37 | 0.10% | 6492.74 | JJS | bad |
| 38 | 0.42% | 6221.43 | JJR | good |
| 39 | 0.09% | 6036.02 | VV | ok |
| 40 | 0.09% | 5957.76 | NN | premise |
| 41 | 0.07% | 5952.75 | JJ | predictable |
| 42 | 0.07% | 5827.07 | JJ | flat |
| 43 | 0.91% | 5819.88 | RB | really |
| 44 | 0.06% | 5752.76 | JJ | awful |
| 45 | 0.25% | 5275.21 | VVZ | seem |
| 46 | 0.10% | 5182.52 | NN | guess |
| 47 | 0.11% | 4969.14 | JJ | poor |
| 48 | 0.06% | 4784.73 | JJ | unbelievable |
| 49 | 0.16% | 4706.24 | NN | half |
| 50 | 0.05% | 4677.48 | VV | bother |

*Figure 10. Keyword list in positive/negative(left/right) reviews*

Keyword list(figure 10) was generated by AntConc. Although we may sort the list by its defining items, it is difficult for comparing their frequency, keyness, and POS all together at one time.

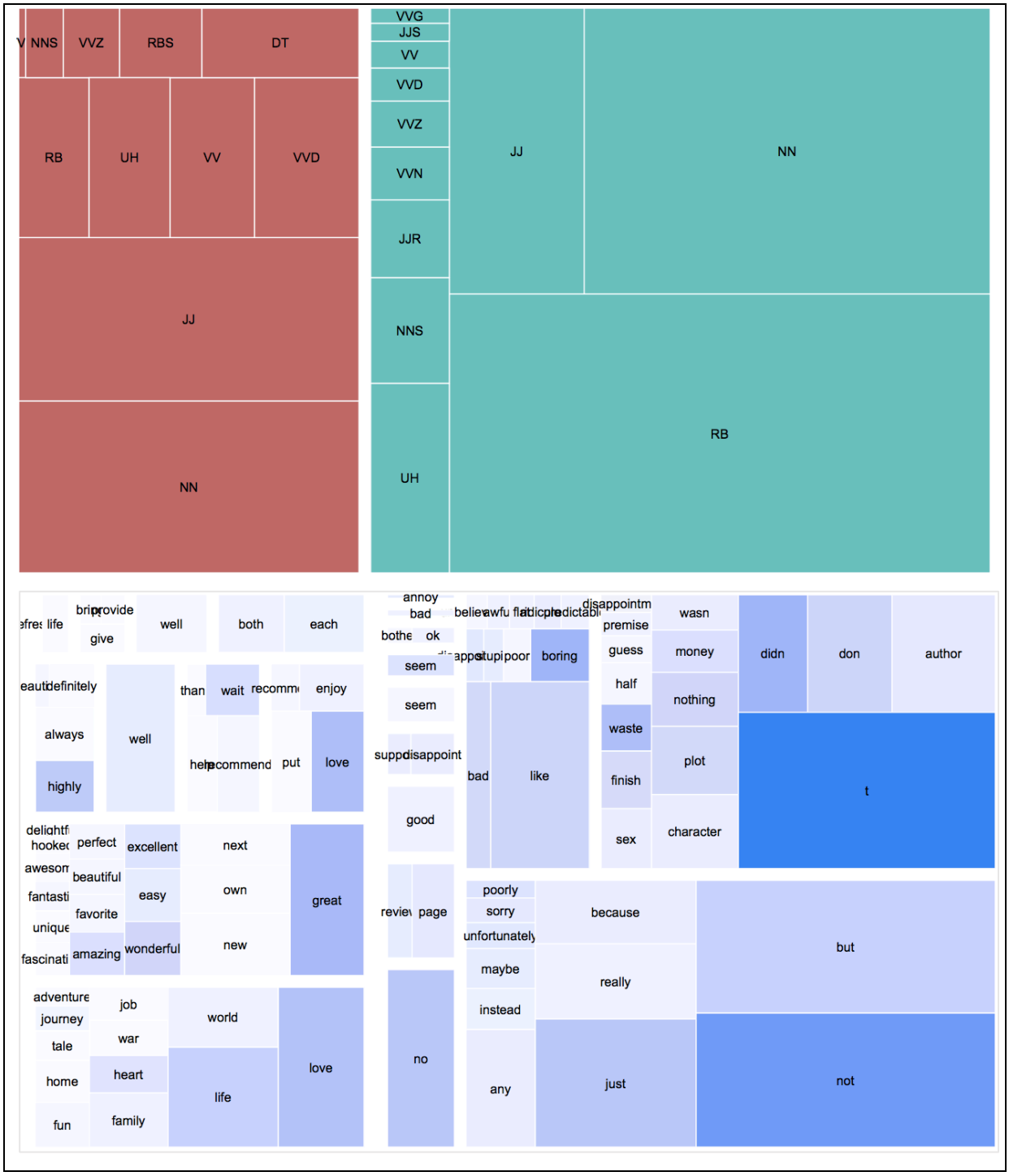



*Figure 11. Hierarchy(POS), Dark blue(High keyness), Size(Frequency), Positive/Negative(Left/Right)*

By using Rawgraphs, it is possible to merge all information together and thus locate three patterns of word usages across positive and negative reviews (figure 11). **(1)** For adjectives, the boundary between two corpora is clear and normal, for example "*great, wonderful*" versus "*boring, bad*". Supporting evident was also found from Gephi by applying degree filter.

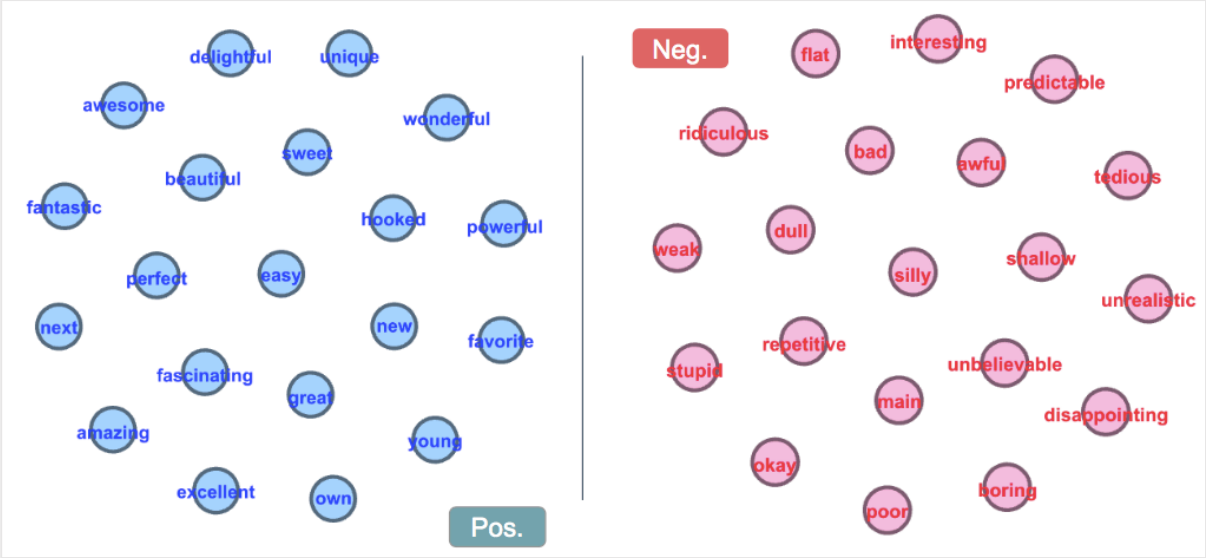



*Figure 12. Filter by degree (Adjective class)*

**(2)** For adverbs and conjunctions, "*but, not, just, because*" stands out in the negative portion (figure 11). It implies that people tends to use more reasoning and contrasting structures in making negative comments**. (3)** For nouns (figure 11), the set of words differ from each other in the sense that positive comments mention more on personal aspects, for instance, "*love, life, heart, family*". On the contrary, negative reviews tends to use more negation markets and nouns regarding specific features of books, such as "*author, plot, character, money*".

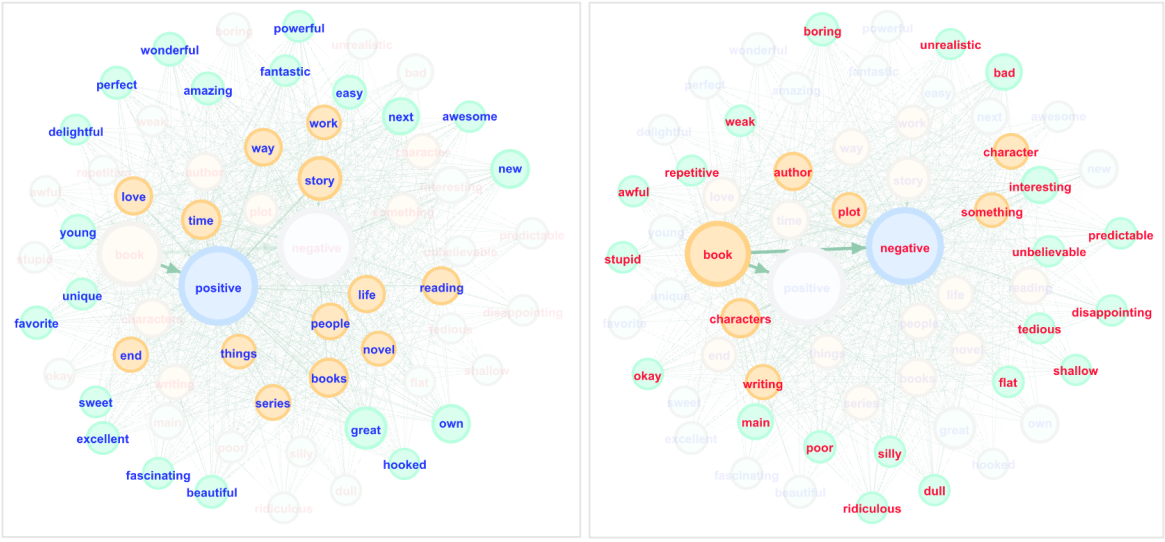



*Figure 13. Filter by modularity (Positive/Negative related words)*

This pattern of noun usage is also found in Gephi(figure 13) by simply applying the modularity filter. The graphs support that negative reviews have higher keyness on noun regarding "specific feature of books", while nouns in the positive portion tends to be more general, such as "*time, things, people*. This information is important for publishers to adjust their marketing strategies in editing and book selling.

### 4.1.3. Gradual scale of adjectives

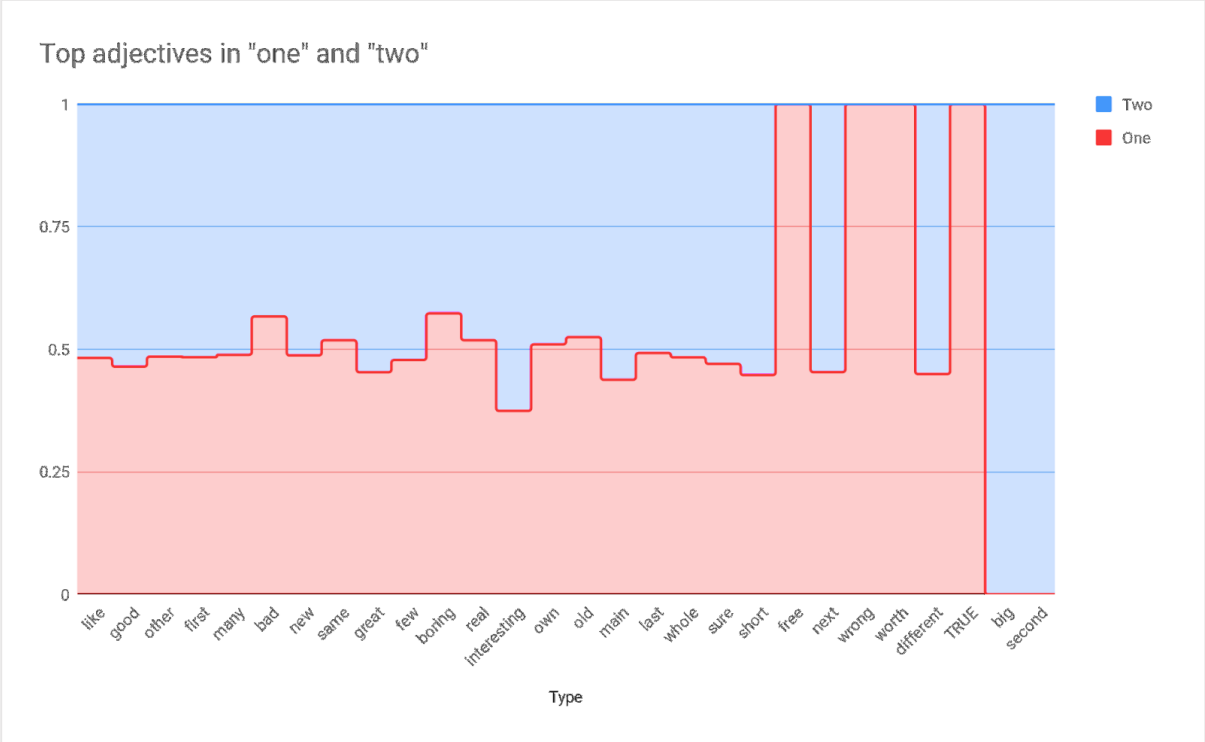



*Figure 14. Adjectives in reviews scored "1.0" and "2.0"*

From percentage of wordlist frequency, top adjectives has similar percentage in both negative sections "1.0" and "2.0" However, "*bad*" and "*boring*" have higher percentage in "1.0", while "*interesting*" has higher percentage in "2.0", which is a higher score.

This might suggest that appraisal adjectives are not straight polarize, but instead, rather in a gradual scale. That means people gradually change their use of adjectives. Further investigation is needed to proof this hypothesis.

### 4.1.4. Collocate

| Author_Positive_ByFreq2L2R | | |
|---|---|---|
| Rank | POS | Type |
| **28** | **JJ** | **new** |
| 41 | JJ | great |
| 51 | JJ | favorite |
| **56** | **JJ** | **amazing** |
| 59 | JJ | like |
| **73** | **JJ** | **first** |
| 75 | JJ | other |
| 77 | JJ | talented |
| 86 | JJ | own |
| 102 | JJ | such |
| 104 | JJ | same |
| **139** | **JJ** | **next** |
| 142 | JJ | excellent |
| 144 | JJ | able |
| **145** | **JJ** | **wonderful** |

| Book_Positive_ByFreq2L2R | | |
|---|---|---|
| Rank | POS | Type |
| 19 | JJ | great |
| **23** | **JJ** | **first** |
| **29** | **JJ** | **next** |
| 44 | JJ | good |
| 51 | JJ | like |
| 62 | JJ | second |
| 63 | JJ | excellent |
| **65** | **JJ** | **wonderful** |
| **91** | **JJ** | **amazing** |
| **102** | **JJ** | **new** |
| 122 | JJ | other |
| 127 | JJ | many |
| 133 | JJ | third |
| 136 | JJ | full |
| 137 | JJ | last |

*Figure 15. Adjectives collocate with "author" and "book"*

| Author_Positive_ByFreq2L2R | | |
|---|---|---|
| Rank | POS | Type |
| 28 | JJ | new |
| 41 | JJ | great |
| **51** | **JJ** | **favorite** |
| 56 | JJ | amazing |
| 59 | JJ | like |
| 73 | JJ | first |
| 75 | JJ | other |
| **77** | **JJ** | **talented** |
| **86** | **JJ** | **own** |
| 102 | JJ | such |
| **104** | **JJ** | **same** |
| 139 | JJ | next |
| 142 | JJ | excellent |
| 144 | JJ | able |
| 145 | JJ | wonderful |

| Book_Positive_ByFreq2L2R | | |
|---|---|---|
| Rank | POS | Type |
| 19 | JJ | great |
| 23 | JJ | first |
| 29 | JJ | next |
| **44** | **JJ** | **good** |
| 51 | JJ | like |
| **62** | **JJ** | **second** |
| 63 | JJ | excellent |
| 65 | JJ | wonderful |
| 91 | JJ | amazing |
| 102 | JJ | new |
| 122 | JJ | other |
| **127** | **JJ** | **many** |
| **133** | **JJ** | **third** |
| **136** | **JJ** | **full** |
| **137** | **JJ** | **last** |

*Figure 16. Distinctive adjectives collocate with "author" and "book"*

In figure 15, it is found that some adjectives can be used to describe different nouns in the common vocabulary set. For example, "*amazing author/book*". While some of the adjectives are dominantly collocated with certain nouns. For example, "*talented*" author versus "*second/third/last*" book. This shows the customer's preferences on various factures.

### 4.1.5. Word Community

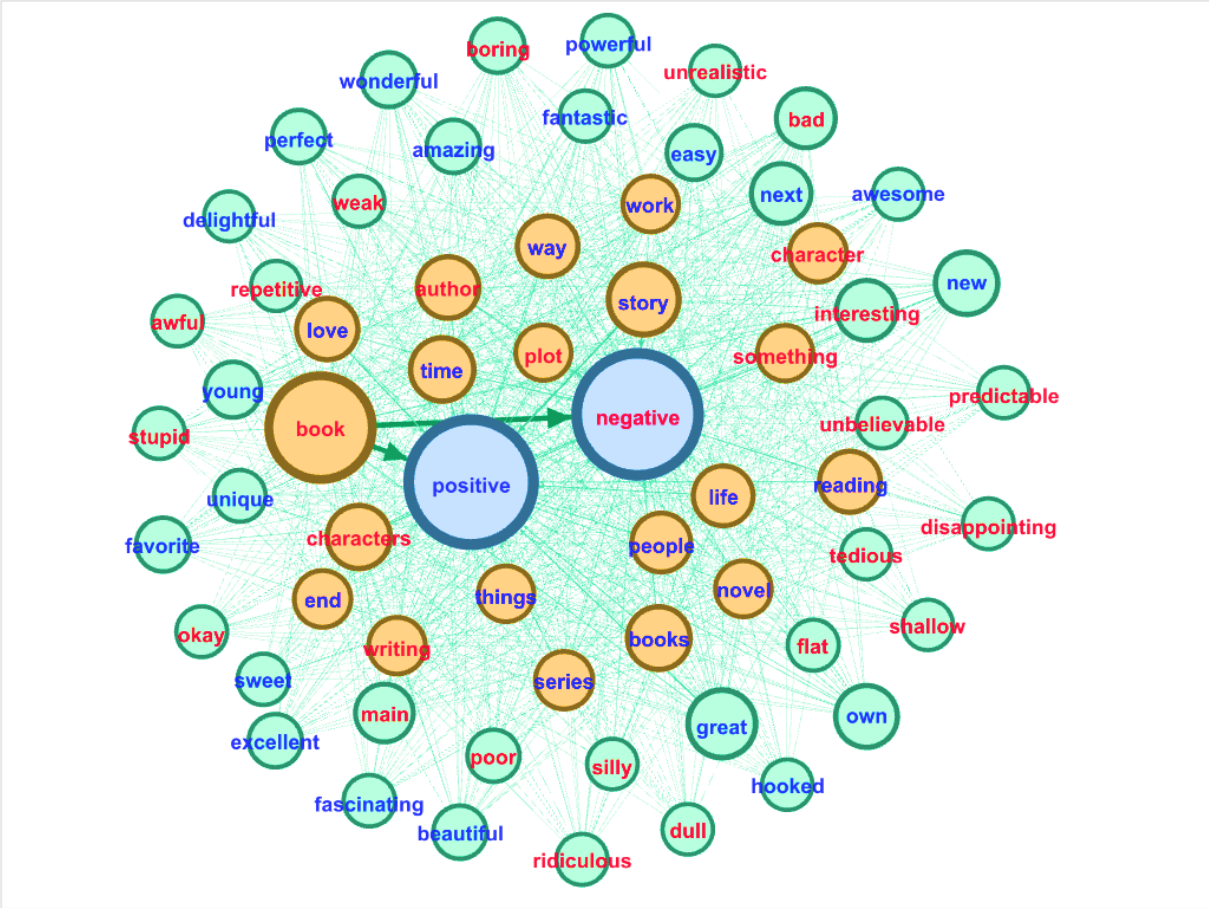



*Figure 14. Gephi graph of top-ranked adjectives and nouns*

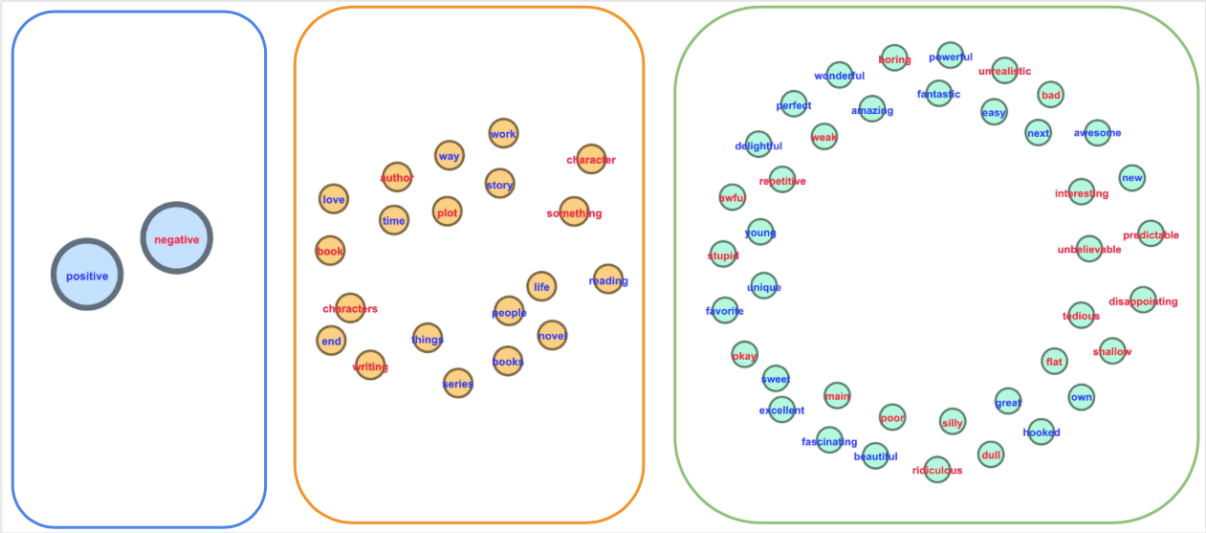



*Figure 15. Degree filter*

In Gephi, degree filter can be applied to differentiate their word classes. On the left is the categories of reviews. The nouns and adjectives can be clearly separated to the middle and right graphs by using degree filters. This demonstrated that words can be separated to different word classes by using only co-occurring frequency.

### 4.1.6. Review peak

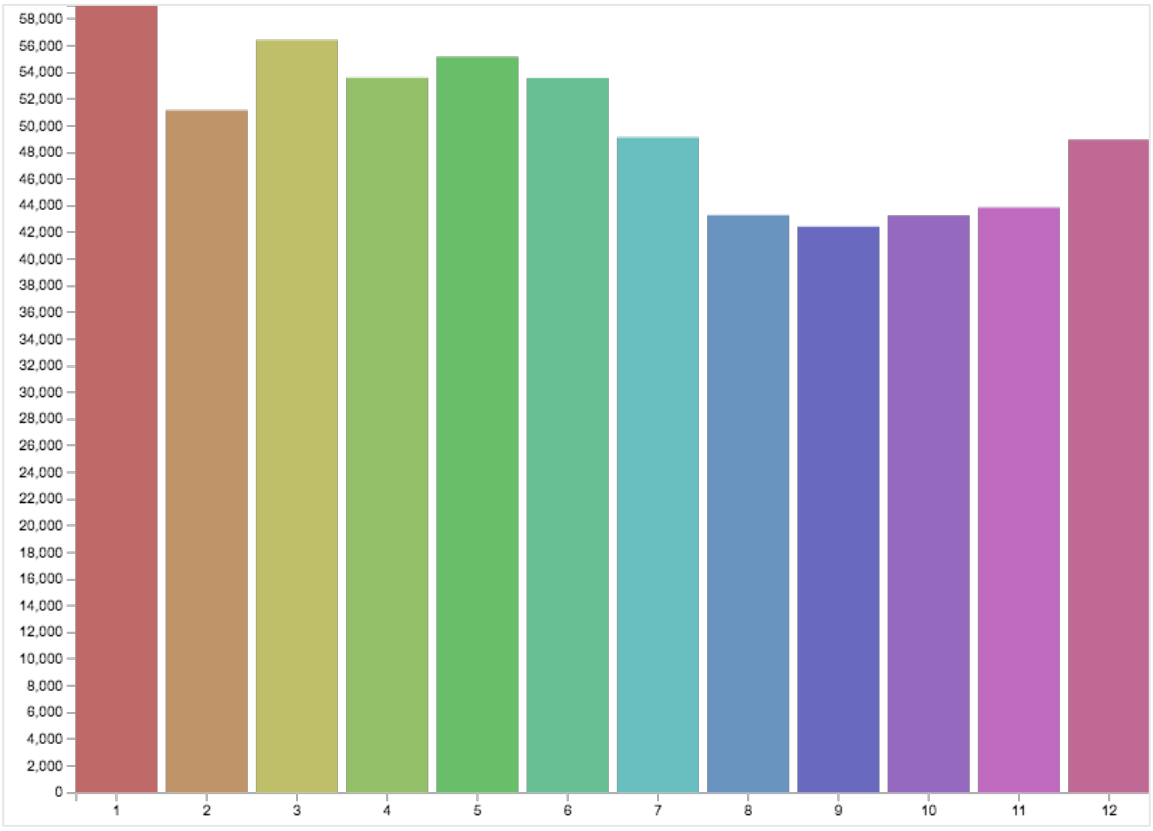



*Figure 16. Bar chart of review numbers*

Originally Gephi was used to plot the trend of the time-stamped reviews. However, Gephi can only show the timeline from 1996-2014 in linear order. It cannot illustrate the pattern by months. Figure 16 shows that peak of reviews occurs at January, March – May, and December. With a possible delay in purchasing time to review posting time, it is possible that Christmas is the peak of purchasing books.

## 4.2. Visualization methods

The two selected visualization methods have been compared in the methodology in terms of their available functions (figure 17).

| Functions \ Software | Voyant Tools | Excel | Ant-Conc | Sketch Engine | Google Ngram Viewer | Gephi |
|---|---|---|---|---|---|---|
| **Available dataset(s)** | | | | | | |
| External Corpus / Dataset | ✓ | ✓ | ✓ | - | - | ✓ |
| **Analysis** | | | | | | |
| Word frequency / Wordlist | ✓ | ✓ | ✓ | ✓ | - | - |
| Keyword list (from 2 dataset) | - | - | ✓ | ✓ | - | - |
| Adjectives in contracting environments | ✓ | - | ✓ | ✓ | - | - |
| Common words in both environments | - | - | ✓ | ✓ | - | - |
| **Visualization** | | | | | | |
| Contrasting / Co-occurring relations | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Focus / Filtering | ✓ | ✓ | - | - | ✓ | ✓ |
| Temporal changes | ✓ | ✓ | - | - | ✓ | ✓ |

*Figure 17. Comparison of software*

### 4.2.1. AntConc

One strength for AntConc is that it is corpus-independent, which user can analysis their own corpus(Kilgarriff & Kosem, 2012). Also, it is good at generating lists for direct comparison of pre-defined items: frequency, keyness, ranked items, or collocates.

While AntConc is good for showing frequency distributions, the format of a list may not be a clear way to highlight or link lots of items which separated in a range of different rank. One clear example of this constrain is shown earlier in figure 10 and 11. Not to mention internal relations between types cannot be reflected by the wordlist. In order to make a graphic representation comparing multi-parameter together, tools like Gephi or Rawgraphs are needed.

Also, it is limited to direct string comparisons, thus not only segmentation problems may occur, one of its weakest areas is handling data in HTML/XML format. Further investigation on lemma form or POS relations would require extra wordlist or manual annotation. If user would

like to perform grammatical analysis, either annotation by TagAnt is needed, or they can only be done by other advanced corpus tools like Sketch Engine, XAIRA, and KorpusDK(Kilgarriff & Kosem, 2012).

### 4.2.2. Gephi

Gephi is good for illustrating internal relations. For example, it can show the in-degree and out-degree relations of nodes. Also, it can show the different communities of nodes by using both the statistical analysis and filtering functions. In this project, nouns could then be categorizing into positive or negative by analysing its modality.

Limitations also occur in Gephi's visualization. The timeline filter could only show the trend of review in linear order. But there is no way to merge all reviews by months instead of years. Showing a few pictures in a series may not the best way to illustrate the pattern. Bar chart was used to compensate for Gephi's limitations and reveal the purchasing behaviour of readers.

# 5. Conclusion and Impact

## 5.1. Findings

The three research questions have all been answered. The results provided concrete evidence and wordlists that suggest:

i) two sets of contrasting adjectives were obtained from keyword lists,

ii) common vocabularies do exist across reviews,

iii) pattern of purchasing period can be revealed by review peaks

Besides the research questions, a few extra findings were located, including "gradual scale of adjectives", "collocates of nouns and adjectives", comparison of other word classes in the two contrasting datasets.

It is also found that both visualization methods have their weakness and limitations. By using various methods together, they can compensate for each other's weaknesses and turn numerical and/or textual data to intuitively understandable patterns.

## 5.2. Impacts

This project demonstrated an approach of using part (around 7%) of the Amazon Book review dataset to produce key figures on word usage. Further research can be carried out by using a larger portion of the dataset or even other categories including music reviews, CDs, and clothes.

Also, it is hoped that the detailed wordlists can be used for constructing sentiment analysis model in computational linguistics. The actual usage of adjectives may also provide empirical insight to appraisal theory. Not only for academic purposes, the result revealing the customer's preference could also help publishers better manage their marketing strategies.

For real applications, full wordlists would be released on github after final submission. It will be combined with other findings from another course and used in my sentiment analysis project[3].

[3] https://github.com/kennethli319/Sentiment-analysis-tool

# 6. References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *proceedings of IWLeL*, 7-13.

Anthony, L. (2015). TagAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University.

Kilgarriff, A., & Kosem, I. (2012). Corpus tools for lexicographers. na.

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information* Retrieval (pp. 43-52). ACM.