# COMP472 - Reports

Thomas Backs[1], Marco Tropiano[2], and Earl Aromin Steven[3]

[1] 27554524 - thomasbacks@gmail.com
[2] 26789331 - tropiano.m@gmail.com
[3] 40004997 - earlaromin@gmail.com

## 1 Introduction

This project is completed within the COMP 472 course. It aims is to create an unsupervised machine learning program that will take some people feedback on various items they purchased and to classify it in two different group: Positive review or negative review. We used 80% of the data collected to train our machine, and the remaining 20% to test our machine. The analysis of each task of the project are shown in different section.

## 2 Task 1 Analysis

For the task 1, we created a function called `def train_nb(documents, labels)`, it takes two parament which are the `documents` and the `labels` to train our machine. We also import the `Counter` from the `collections` modules as well to help us to count the number of occurence of each word. We separate the occurence of word in negative review from the occurence in positive review by using the labels already there to teach

```
def train_nb(documents, labels):
neg_word_count = Counter()
pos_word_count = Counter()
neg_total_word = 0
pos_total_word = 0
# we now create our classification
classifier = list(zip(labels, documents))
for c in classifier:
    if c[0] == 'neg':
        neg_word_count.update(c[1])
    else:
        pos_word_count.update(c[1])

neg_total_word = sum(neg_word_count.values())
pos_total_word = sum(pos_word_count.values())
total_word = sum(neg_total_word, pos_total_word)
return neg_total_word, pos_total_word,
    total_word, neg_word_count, pos_word_count
```

## 3 Task 2 Analysis

The task 2 is seperate in two part, where the first part is to get the score and the second part compares the scores we got and return the appropriate classification (negative or positive review).

The function called `score_doc_label(document, smoothing=0.5)`.
The smoothing paramater is set to default which is 0.5 and can be modified in the function call to suit future need. We use this formula to find our probabilty:

$$score(\text{Negative}) = \log_{10}(P(\text{Negative})) + \sum_{i=1}(P(w_i|\text{Negative}))$$
$$score(\text{Positive}) = \log_{10}(P(\text{Positive})) + \sum_{i=1}(P(w_i|\text{Positive}))$$

Then we take the scores from our formula calcuation above, and we return the negative probability and the positive probability

Now this takes us to the part of of the task 2, to perform our probabilities calculation, we call the function called `classify_nb(document, smoothing=0.5)`, this function will call the function previously mentionned in the first part, and will get the return values of the scores for negative and positive probability. It will compares both scores, pick the highest one and will return the fitting label for our review.

## 4 Accuracy Analysis

With a 0.5 smoothing for our training data and we end up with this result shown below:

```
Training set accuracy (0.5) :
   Overall accuracy : 0.8712621970412339
   Pos accuracy : 0.8429329291398256
   Neg accuracy : 0.906418998354103
```

The data above show the accuracy of our training data with a smoothing of 0.5. Below there is more datas and graphs based on the evaluation (test) data we have. The first graph show are the overall accuracy based on different smoothing values between 0.1 and 1.
`max overall acc at smoothing value 0.93: 0.8120016785564415` or 81.2%.

## 5 Analysis of Misclassified Documents

Below you will find some misclassified documents along with some analysis on why it has been misclassified. Each team member picked 3 misclassified data as random. The accuracy for our misclassified document is based on smoothing value of **0.5**.

### 5.1 Earl's Analysis

**Review 1** First analysis of review is a review that has been labelled as negative but our AI classified it as positive.

> *i live in south africa where the series appeared 20 years ago.thanks to amazon i was able to relive the wonder of archie bunker where each episode provides laughs from one of the family , each of whom is brilliantly cast.i have series 3&4 and now want 1&2.do n't wait any longer , 24 episodes of absolute delight await you*

This is misclassified as training data. It's labeled as negative while speaking only positive words towards the dvd...

**Review 2** This is a positive review that has been clasified as positive.

> *excellent service , and the product worked extremely well . the only criticism i would have is the cost of postage . it was ridiculous , the product came in a big parcel , and inside was a dvd .*

This is a short text. It includes pretty heavy words that would be negative such as ridiculous and criticism. Cost itself is another factor and aside from the obviously positive like excellent and well, the rest can be neutral

**Review 3** This is a negative review that has been classified as positive.

> *i have worked out to denise austin for several years because i think she gives a good workout but i cannot stand her way-too-perky personality . i was very disappointed in this dance workout . the setting is fun and colorful and the music is great , but as far as a workout...this dvd really stinks . i found some of the steps somewhat confusing ( but i am not very coordinated ) . this workout was more of an " indoor walk " style workout ( like walk away the pounds w / leslie sansone , etc. ) . if you are looking for a fun way to get up and move , this may be for you . but , this is not for anyone looking for a cardio intense workout*

Despite the negative review, the person acknowledges the positive aspects of it. In this case, negative just means it doesn't work for them; not necessarily that the product is bad. Biggest offenders would be good,colorful,fun,great

### 5.2 Marco's Analysis

**Review 1** This is a positive review that has been classified as negative.

> *i luv this burberry . pple always comment anytime i wear it . my girlfriend almost made me stop wearing it cos she said anytime i wear it she feels like doing things i ca n't mention here 2 me , and since she feels tht way , other women will certainly feel like tht . i 'm getting another one ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! ! also fast and proper services by amazon*

Misclassified because the word **stop** was used and alot of exclamation points usually means the person is angry

**Review 2** This is a positive review that has been classified as negative.

> *i am very happy with the emjoi compact epilator that i purchased . it does exactly what it claims to do and works wonderfully . until i asked a salon for advice on the best way to remove unwanted facial hair , i did n't even know something like this exhisted . i just wish i would have known about this product years ago . a . davis*

Misclassified because the word **until** was used

**Review 3** This is a positive review that has been classified as negative.

> *i was searching for a two inch curling iron for a while . i have had so -so dealings with gold n hot in the past so i was a little nervous but i purchased this one anyway . so far , so good . it 's surprisingly light ( at least compared to my one inch ceramic tools curling iron ) . i usually let my hair air dry and use my curling irons to straighten my hair as i curl the ends ( i 'm black with a relaxer . ) i have yet to do it with this one . although i might try b / c it gets really hot . ( i usually have it on around 20-25 out of 30 ) the only problem is that the ceramic can chip off ( it happened with the first one i bought , but i returned it and got a new one that has n't chipped so far ) and it 's so big that if you take your thumb off the spring grip , it can be hard to get it back on ( i have huge hands too .*

Misclassified with negative words like **nervous** and **problem** gave it the wrong label

### 5.3 Thomas' Analysis

Here is how I perform calculations to find out the most incriminating word shown below:

$$\begin{aligned} \text{NEG} + 0.5 &= \text{SNEG} \\ \text{POS} + 0.5 &= \text{SPOS} \\ \frac{SNEG}{SPOS} &= x \\ \text{IF } x > 1 &\rightarrow x \\ \text{ELSE } &\rightarrow \frac{1}{x} \end{aligned}$$

Where SNEG and SPOS mean smoothed negative and smoothed positive respectively. The x is how impactful this value is to our formula. This formula is applied for each word in our misclassified reviews. This formula has been re-use for each of my reviews. The actual data-sheet can be found at the end of the file.

**Review 1** First analysis of review is a negative review that been classified as positive.

> *i agree with other reviewers that it feels good and does n't smell too much , however , i 've experimented with it several times to confirm my findings , and it turns out to give me really bad blackheads . i 'm 25 with an oily t-zone and very dry facial skin . on mornings after using this cream , i have nasty blackheads on my forehead and chin . there are better products out there*

After performing the calculation, the culprit are mostly the words **findings**, **mornings** since these word never appear in negative review it gives them a score of 0 (before the 0.5 smoothing), when we perform the neg/pos calculation with smoothing, it give an higher value to it. There is also the word **oily** which is strongly associated with positive review. I have compacted it for ease of view.

**Review 2** This second review is a positive review that been classified as negative.

> *as if this set could be anything but five stars ! ! ! barbra streisand - four of her films on their dvd debut - her own commentary - oy vey ! order it now and plan a vacation day for the day after you get delivery*

The major culprit of this review is the word **streisand** which is associated heavily with negative review.

**Review 3** This third review is a positive review that been classified as negative.

> *$10 bucks and does the job . you do need to adjust it every once in awhile but that 's easy to do .*

This review is a bit more tricky to actually find the culprit, because the highest impacting factor does favour positive classification, however, the next 3 values that are impactful are favouring negative classification, word such as **$10**, **bucks** and **job**.

## 6 What would you expand

Waht would we expand... Obviously looking at each document word by word in isolation limits us because words are contextual. One way to improve this, but still not perfect, is to use an n-gram to look at the sequence of words. Another way to look at it is to review the surrounding words.

On the classification side, we could also split it into more labels. Aside from just positive and negative, we can include the product type being reviewed. For example, we can have dvd-pos, dvd-neg, music-pos, music-neg, health-pos, health-neg. The reasoning for this is that the type of product also gives us

context clues. A quick example is the word "sick", which can be interpreted as more positive in music than in health.

The addition of a k-fold cross validation and evaluating more than just the accuracy(we could do recall, precision, and f1 score ) to measure the performance would also be something to look at.

Review 1 - Thomas

| WORD | NEG | SNEG | POS | SPOS | SNES/POS | X |
|---|---|---|---|---|---|---|
| "i" | 16707 | 16707.5 | 14814 | 14814.5 | 1.127780215 | 1.12778021532958 |
| "agree" | 126 | 126.5 | 68 | 68.5 | 1.846715328 | 1.84671532846715 |
| "with" | 4930 | 4930.5 | 5445 | 5445.5 | 0.905426499 | 1.10445188114796 |
| "other" | 1130 | 1130.5 | 1108 | 1108.5 | 1.01984664 | 1.01984663960307 |
| "reviewers" | 105 | 105.5 | 64 | 64.5 | 1.635658915 | 1.63565891472868 |
| "that" | 7851 | 7851.5 | 6915 | 6915.5 | 1.135348131 | 1.13534813101005 |
| "it" | 12338 | 12338.5 | 11728 | 11728.5 | 1.052010061 | 1.05201006096261 |
| "feels" | 52 | 52.5 | 86 | 86.5 | 0.606936416 | 1.64761904761905 |
| "good" | 1455 | 1455.5 | 1636 | 1636.5 | 0.889398106 | 1.12435589144624 |
| "and" | 15970 | 15970.5 | 18754 | 18754.5 | 0.851555627 | 1.17432140509064 |
| "does" | 1396 | 1396.5 | 1106 | 1106.5 | 1.262087664 | 1.26208766380479 |
| "n't" | 4174 | 4174.5 | 2869 | 2869.5 | 1.454783063 | 1.45478306325144 |
| "smell" | 18 | 18.5 | 23 | 23.5 | 0.787234043 | 1.27027027027027 |
| "too" | 863 | 863.5 | 646 | 646.5 | 1.335653519 | 1.33565351894818 |
| "much" | 1072 | 1072.5 | 1005 | 1005.5 | 1.066633516 | 1.06663351566385 |
| "," | 26333 | 26333.5 | 27532 | 27532.5 | 0.956451466 | 1.04553135739647 |
| "however" | 516 | 516.5 | 458 | 458.5 | 1.126499455 | 1.12649945474373 |
| "," | 26333 | 26333.5 | 27532 | 27532.5 | 0.956451466 | 1.04553135739647 |
| "i" | 16707 | 16707.5 | 14814 | 14814.5 | 1.127780215 | 1.12778021532958 |
| "'ve" | 694 | 694.5 | 855 | 855.5 | 0.811805961 | 1.23182145428366 |
| "experimented" | 1 | 1.5 | 3 | 3.5 | 0.428571429 | 2.33333333333333 |
| "with" | 4930 | 4930.5 | 5445 | 5445.5 | 0.905426499 | 1.10445188114796 |
| "it" | 12338 | 12338.5 | 11728 | 11728.5 | 1.052010061 | 1.05201006096261 |
| "several" | 302 | 302.5 | 285 | 285.5 | 1.059544658 | 1.05954465849387 |
| "times" | 391 | 391.5 | 390 | 390.5 | 1.002560819 | 1.00256081946223 |
| "to" | 16647 | 16647.5 | 15877 | 15877.5 | 1.0484963 | 1.04849629979531 |
| "confirm" | 6 | 6.5 | 1 | 1.5 | 4.333333333 | 4.33333333333333 |
| "my" | 3542 | 3542.5 | 3840 | 3840.5 | 0.922405937 | 1.08412138320395 |
| "findings" | 0 | 0.5 | 2 | 2.5 | 0.2 | 5 |
| "," | 26333 | 26333.5 | 27532 | 27532.5 | 0.956451466 | 1.04553135739647 |
| "and" | 15970 | 15970.5 | 18754 | 18754.5 | 0.851555627 | 1.17432140509064 |
| "it" | 12338 | 12338.5 | 11728 | 11728.5 | 1.052010061 | 1.05201006096261 |
| "turns" | 61 | 61.5 | 74 | 74.5 | 0.825503356 | 1.21138211382114 |
| "out" | 1925 | 1925.5 | 1566 | 1566.5 | 1.229173316 | 1.22917331631025 |
| "to" | 16647 | 16647.5 | 15877 | 15877.5 | 1.0484963 | 1.04849629979531 |
| "give" | 436 | 436.5 | 326 | 326.5 | 1.336906585 | 1.33690658499234 |
| "me" | 1776 | 1776.5 | 1452 | 1452.5 | 1.223063683 | 1.22306368330465 |
| "really" | 964 | 964.5 | 1173 | 1173.5 | 0.821900298 | 1.21669258683256 |
| "bad" | 709 | 709.5 | 261 | 261.5 | 2.713193117 | 2.7131931166348 |
| "blackheads" | 0 | 0.5 | 1 | 1.5 | 0.333333333 | 3 |
| "." | 30682 | 30682.5 | 29981 | 29981.5 | 1.023381085 | 1.02338108500242 |
| "i" | 16707 | 16707.5 | 14814 | 14814.5 | 1.127780215 | 1.12778021532958 |
| "'m" | 749 | 749.5 | 647 | 647.5 | 1.157528958 | 1.15752895752896 |
| "25" | 16 | 16.5 | 22 | 22.5 | 0.733333333 | 1.36363636363636 |
| "with" | 4930 | 4930.5 | 5445 | 5445.5 | 0.905426499 | 1.10445188114796 |
| "an" | 1892 | 1892.5 | 2106 | 2106.5 | 0.898409684 | 1.11307793923382 |
| "oily" | 2 | 2.5 | 11 | 11.5 | 0.217391304 | 4.6 |
| "t-zone" | 0 | 0.5 | 0 | 0.5 | 1 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| "and" | 15970 | 15970.5 | 18754 | 18754.5 | 0.851555627 | 1.17432140509064 |
| "very" | 1596 | 1596.5 | 2050 | 2050.5 | 0.778590588 | 1.28437206388976 |
| "dry" | 44 | 44.5 | 62 | 62.5 | 0.712 | 1.40449438202247 |
| "facial" | 15 | 15.5 | 15 | 15.5 | 1 | 1 |
| "skin" | 85 | 85.5 | 151 | 151.5 | 0.564356436 | 1.7719298245614 |
| "." | 30682 | 30682.5 | 29981 | 29981.5 | 1.023381085 | 1.02338108500242 |
| "on" | 4701 | 4701.5 | 4485 | 4485.5 | 1.048155167 | 1.04815516664809 |
| "mornings" | 0 | 0.5 | 2 | 2.5 | 0.2 | 5 |
| "after" | 1185 | 1185.5 | 840 | 840.5 | 1.410469958 | 1.41046995835812 |
| "using" | 426 | 426.5 | 550 | 550.5 | 0.774750227 | 1.29073856975381 |
| "this" | 9608 | 9608.5 | 9391 | 9391.5 | 1.023106 | 1.02310600010648 |
| "cream" | 20 | 20.5 | 61 | 61.5 | 0.333333333 | 3 |
| "," | 26333 | 26333.5 | 27532 | 27532.5 | 0.956451466 | 1.04553135739647 |
| "i" | 16707 | 16707.5 | 14814 | 14814.5 | 1.127780215 | 1.12778021532958 |
| "have" | 4264 | 4264.5 | 4136 | 4136.5 | 1.030944035 | 1.03094403481204 |
| "nasty" | 11 | 11.5 | 19 | 19.5 | 0.58974359 | 1.69565217391304 |
| "blackheads" | 0 | 0.5 | 1 | 1.5 | 0.333333333 | 3 |
| "on" | 4701 | 4701.5 | 4485 | 4485.5 | 1.048155167 | 1.04815516664809 |
| "my" | 3542 | 3542.5 | 3840 | 3840.5 | 0.922405937 | 1.08412138320395 |
| "forehead" | 2 | 2.5 | 6 | 6.5 | 0.384615385 | 2.6 |
| "and" | 15970 | 15970.5 | 18754 | 18754.5 | 0.851555627 | 1.17432140509064 |
| "chin" | 2 | 2.5 | 5 | 5.5 | 0.454545455 | 2.2 |
| "." | 30682 | 30682.5 | 29981 | 29981.5 | 1.023381085 | 1.02338108500242 |
| "there" | 1773 | 1773.5 | 1543 | 1543.5 | 1.149011986 | 1.14901198574668 |
| "are" | 3469 | 3469.5 | 3641 | 3641.5 | 0.952766717 | 1.04957486669549 |
| "better" | 926 | 926.5 | 743 | 743.5 | 1.246133154 | 1.24613315400135 |
| "products" | 187 | 187.5 | 144 | 144.5 | 1.297577855 | 1.29757785467128 |
| "out" | 1925 | 1925.5 | 1566 | 1566.5 | 1.229173316 | 1.22917331631025 |
| "there" | 1773 | 1773.5 | 1543 | 1543.5 | 1.149011986 | 1.14901198574668 |

| WORD | NEG | | SNEG | POS | SPOS | | SNES/POS | X |
|---|---|---|---|---|---|---|---|---|
| "as" | | 3533 | 3533.5 | 4438 | 4438.5 | 0.796102287 | 1.256119994 | |
| "if" | | 2538 | 2538.5 | 2093 | 2093.5 | 1.212562694 | 1.212562694 | |
| "this" | | 9608 | 9608.5 | 9391 | 9391.5 | 1.023106 | 1.023106 | |
| "set" | | 287 | 287.5 | 373 | 373.5 | 0.769745649 | 1.299130435 | |
| "could" | | 1142 | 1142.5 | 736 | 736.5 | 1.55125594 | 1.55125594 | |
| "be" | | 3082 | 3082.5 | 2698 | 2698.5 | 1.142301278 | 1.142301278 | |
| "anything" | | 376 | 376.5 | 222 | 222.5 | 1.692134831 | 1.692134831 | |
| "but" | | 4498 | 4498.5 | 3853 | 3853.5 | 1.167380304 | 1.167380304 | |
| "five" | | 57 | 57.5 | 148 | 148.5 | 0.387205387 | 2.582608696 | |
| "stars" | | 260 | 260.5 | 216 | 216.5 | 1.203233256 | 1.203233256 | |
| "!" | | 2834 | 2834.5 | 2759 | 2759.5 | 1.027178837 | 1.027178837 | |
| "!" | | 2834 | 2834.5 | 2759 | 2759.5 | 1.027178837 | 1.027178837 | |
| "!" | | 2834 | 2834.5 | 2759 | 2759.5 | 1.027178837 | 1.027178837 | |
| "barbra" | | 0 | 0.5 | 0 | 0.5 | 1 | 1 | |
| "streisand" | | 8 | 8.5 | 0 | 0.5 | 17 | 17 | |
| "-" | | 2025 | 2025.5 | 1998 | 1998.5 | 1.013510133 | 1.013510133 | |
| "four" | | 108 | 108.5 | 161 | 161.5 | 0.671826625 | 1.488479263 | |
| "of" | | 12895 | 12896 | 14207 | 14207.5 | 0.907654408 | 1.101740917 | |
| "her" | | 1067 | 1067.5 | 1044 | 1044.5 | 1.022020105 | 1.022020105 | |
| "films" | | 123 | 123.5 | 142 | 142.5 | 0.866666667 | 1.153846154 | |
| "on" | | 4701 | 4701.5 | 4485 | 4485.5 | 1.048155167 | 1.048155167 | |
| "their" | | 1216 | 1216.5 | 1105 | 1105.5 | 1.100407056 | 1.100407056 | |
| "dvd" | | 425 | 425.5 | 456 | 456.5 | 0.932092004 | 1.072855464 | |
| "debut" | | 18 | 18.5 | 44 | 44.5 | 0.415730337 | 2.405405405 | |
| "-" | | 2025 | 2025.5 | 1998 | 1998.5 | 1.013510133 | 1.013510133 | |
| "her" | | 1067 | 1067.5 | 1044 | 1044.5 | 1.022020105 | 1.022020105 | |
| "own" | | 355 | 355.5 | 443 | 443.5 | 0.801578354 | 1.247538678 | |
| "commentary" | | 29 | 29.5 | 41 | 41.5 | 0.710843373 | 1.406779661 | |
| "-" | | 2025 | 2025.5 | 1998 | 1998.5 | 1.013510133 | 1.013510133 | |
| "oy" | | 0 | 0.5 | 0 | 0.5 | 1 | 1 | |
| "vey" | | 1 | 1.5 | 0 | 0.5 | 3 | 3 | |
| "!" | | 2834 | 2834.5 | 2759 | 2759.5 | 1.027178837 | 1.027178837 | |
| "order" | | 166 | 166.5 | 144 | 144.5 | 1.152249135 | 1.152249135 | |
| "it" | | 12338 | 12339 | 11728 | 11728.5 | 1.052010061 | 1.052010061 | |
| "now" | | 750 | 750.5 | 767 | 767.5 | 0.977850163 | 1.022651566 | |
| "and" | | 15970 | 15971 | 18754 | 18754.5 | 0.851555627 | 1.174321405 | |
| "plan" | | 51 | 51.5 | 76 | 76.5 | 0.673202614 | 1.485436893 | |
| "a" | | 14857 | 14858 | 16184 | 16184.5 | 0.918007971 | 1.089315161 | |
| "vacation" | | 24 | 24.5 | 27 | 27.5 | 0.890909091 | 1.12244898 | |
| "day" | | 322 | 322.5 | 424 | 424.5 | 0.759717314 | 1.31627907 | |
| "for" | | 6033 | 6033.5 | 6707 | 6707.5 | 0.899515468 | 1.111709621 | |
| "the" | | 31914 | 31915 | 33502 | 33502.5 | 0.952600552 | 1.049757947 | |
| "day" | | 322 | 322.5 | 424 | 424.5 | 0.759717314 | 1.31627907 | |
| "after" | | 1185 | 1185.5 | 840 | 840.5 | 1.410469958 | 1.410469958 | |
| "you" | | 5382 | 5382.5 | 5808 | 5808.5 | 0.926659206 | 1.079145379 | |
| "get" | | 1564 | 1564.5 | 1309 | 1309.5 | 1.194730813 | 1.194730813 | |
| "delivery" | | 24 | 24.5 | 33 | 33.5 | 0.731343284 | 1.367346939 | |

Review 3 - Thomas

| WORD | NEG | SNEG | POS | SPOS | SNES/POS | VALUE |
|---|---|---|---|---|---|---|
| "$10" | 18 | 18.5 | 7 | 7.5 | 2.466666667 | 2.466666667 |
| "bucks" | 45 | 45.5 | 18 | 18.5 | 2.459459459 | 2.459459459 |
| "and" | 15970 | 15970.5 | 18754 | 18754.5 | 0.851555627 | 1.174321405 |
| "does" | 1396 | 1396.5 | 1106 | 1106.5 | 1.262087664 | 1.262087664 |
| "the" | 31914 | 31914.5 | 33502 | 33502.5 | 0.952600552 | 1.049757947 |
| "job" | 116 | 116.5 | 291 | 291.5 | 0.399656947 | 2.502145923 |
| "." | 30682 | 30682.5 | 29981 | 29981.5 | 1.023381085 | 1.023381085 |
| "you" | 5382 | 5382.5 | 5808 | 5808.5 | 0.926659206 | 1.079145379 |
| "do" | 2684 | 2684.5 | 1918 | 1918.5 | 1.399270263 | 1.399270263 |
| "need" | 409 | 409.5 | 524 | 524.5 | 0.780743565 | 1.280830281 |
| "to" | 16647 | 16647.5 | 15877 | 15877.5 | 1.0484963 | 1.0484963 |
| "adjust" | 22 | 22.5 | 22 | 22.5 | 1 | 1 |
| "it" | 12338 | 12338.5 | 11728 | 11728.5 | 1.052010061 | 1.052010061 |
| "every" | 445 | 445.5 | 505 | 505.5 | 0.881305638 | 1.134680135 |
| "once" | 245 | 245.5 | 299 | 299.5 | 0.819699499 | 1.219959267 |
| "in" | 7586 | 7586.5 | 8502 | 8502.5 | 0.89226698 | 1.12074079 |
| "awhile" | 14 | 14.5 | 15 | 15.5 | 0.935483871 | 1.068965517 |
| "but" | 4498 | 4498.5 | 3853 | 3853.5 | 1.167380304 | 1.167380304 |
| "that" | 7851 | 7851.5 | 6915 | 6915.5 | 1.135348131 | 1.135348131 |
| "'s" | 4769 | 4769.5 | 5309 | 5309.5 | 0.898295508 | 1.113219415 |
| "easy" | 203 | 203.5 | 704 | 704.5 | 0.288857346 | 3.461916462 |
| "to" | 16647 | 16647.5 | 15877 | 15877.5 | 1.0484963 | 1.0484963 |
| "do" | 2684 | 2684.5 | 1918 | 1918.5 | 1.399270263 | 1.399270263 |
| "." | 30682 | 30682.5 | 29981 | 29981.5 | 1.023381085 | 1.023381085 |