…As I read on my kindle I highlight the passages that I like so that I can re-read them later. These annotations are stored on my Kindle and are backed up at Amazon. And after some time, they started to accumulate and became some kind of data.

I came up with an idea to analyze all those text I highlighted on my kindle, **to figure out what kind of content I was most likely to highlight.**

The plan was to use text mining and sentiment analysis, generate insights and compare them to my real opinions of those books. **So I can have the first-hand test of how useful text mining is. With that knowledge, I can be more convinced when I apply the method to a business problem.**

Now to be objective, I will create an independent character called “the bat”. He is a hacker and he is not the bookworm type but his unfortunate fate gave him a challenge.

His next mission is to hack my data and analyze it to gather insights for selling me more books.
The room was in a mess when he entered. As he plugged the USB drive on his laptop, he barely heard the radio which was still on.

The reporter:

> From the moment we switch on the early morning our cell phones to deal with the flood of information in some form of a whatsapp or facebook message or a tweet until we fall asleep at night overwriting or reading a product review we leave bread crumbs to our personal flavors on internet.
>
> Many businesses use this unstructured data to drive their sales by better marketing through targeted product recommendations or to segregate their customers…

He squeezed his teeth when he saw the data of 21000 lines of text from 28 books. His first encounter was the book “Mindset” by Carol Dweck, where she introduces the concept of growth mindset.

Long lines of text made him tired, he didn’t realize how the time passed. **He decided to learn text mining**. This might sound quite a bit investment but now he is a man of growth.

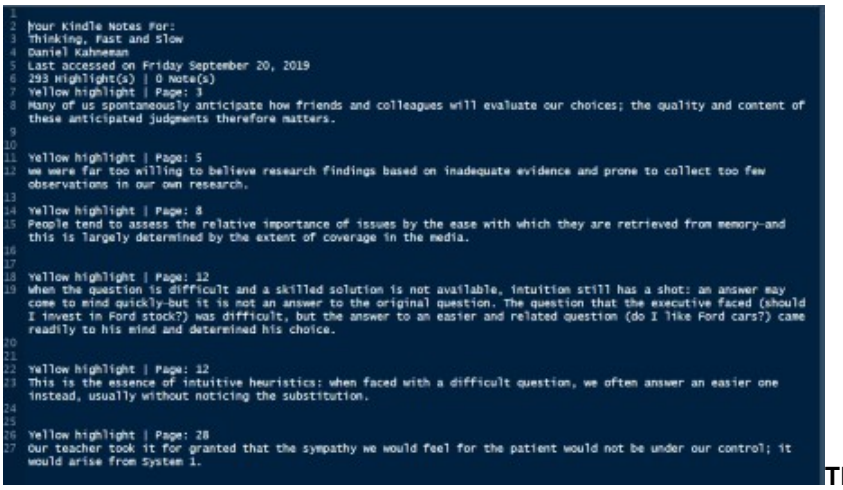
He continued to learn R packages needed for text mining, he didn’t like the package name tidytext but he was slightly losing his prejudices. It was a long night. He fell asleep on his table as the sun slowly rose.
It was lightning my back garden where

…

I could glance from time to time to the trees painted by the snow overnight 🐓. Without an idea about how things went on the another part of the town, I continued to read and highlight my kindle as I zip from a glass of red wine 🍷.

…

This is how the exported kindle highlights look like.


The hacker's notes

He noted down each step of his text mining plan carefully. Let me help you go through them.
Reading and parsing the text file

```
# Use readLines function to parse the text file

highlights <- readLines("posts_data/Kindle_highlights_Serdar.Rmd", encoding =
"UTF-8")

# Create a dataframe where each row is a line from the text

df <- data.frame(highlights)

# Packages

library(tidyverse)   # includes ggplot2, dplyr, tidyr, readr, purrr, tibble,
stringr, forcats
library(tidytext)
library(wordcloud2)
```

In every data science project, there is some sort of data preparation. **Stop words** are generally the most common words in a language and are usually filtered out before processing of text data.

Let's look at the stop_words dataset from the tidytext package. Since it is a long list of words (>1K) **I will print every fifth word as an example.**

```
data(stop_words)

# print every 50th word

stop_words_small <- stop_words[seq(1, nrow(stop_words), 50),]

stop_words_small %>% print(n=50)

## # A tibble: 23 x 2
##    word       lexicon
##
##  1 a          SMART
##  2 at         SMART
##  3 contain    SMART
##  4 few        SMART
##  5 hers       SMART
##  6 last       SMART
##  7 nine       SMART
##  8 presumably SMART
```

```
##  9 some       SMART
## 10 they'd     SMART
## 11 very       SMART
## 12 without    SMART
## 13 what       snowball
## 14 they'll    snowball
## 15 during     snowball
## 16 again      onix
## 17 but        onix
## 18 finds      onix
## 19 if         onix
## 20 much       onix
## 21 parted     onix
## 22 since      onix
## 23 under      onix
```

Sometimes even a small dot can have big influence on your results. **Looking carefully I see that stop_words uses single quotes whereas in the text file used apostrophes (‘).**

e.g. they’ll in stop_words

And how the word they’ll appears in the text:

Yellow highlight | Page: 200
Memories are continually revised, along with the meaning we derive from them, so that in the future they’ll be of even more use.

This incompatibility will prevent some of the stop_words such as they’ll, don’t, can’t e.g. getting filtered. To prevent this we have to replace them.

He quickly spotted that.

str_replace_all() function from Stringr will do that.

```
df$highlights <- str_replace_all(df$highlights, "’", "'")
```

Now, the text is ready for the frequency analysis. Words in a text mining project are called tokens. We can split the text into single words by unnest_tokens() function from tidytext package, filter the stop_words and count.

```
df <- df %>% unnest_tokens(word, highlights) %>%
             anti_join(stop_words) %>%
             filter(!word %in% c("highlights","highlight", "page",
                      "location", "yellow", "pink", "orange", "blue"))
```

He also added here some additional words which frequently appear in kindle highlights output.

dplyr() package functions are very useful for grouping and counting the words from the lists that are created.

```
top_kindle_highlights <- df %>%
 group_by(word) %>%
 count() %>%
 arrange(desc(n))
```

He noted down his first insight. **10 most frequent words from my kindle highlights.**

```
top_kindle_highlights
```

```
## # A tibble: 12,433 x 2
## # Groups:   word [12,433]
##    word       n
##
```

```
##  1 people   592
##  2 story    340
##  3 life     318
##  4 time     309
##  5 mind     213
##  6 change   212
##  7 feel     211
##  8 world    171
##  9 person   170
## 10 habits   157
## # ... with 12,423 more rows
```

Wordclouds are a good alternative to long lists of words for visualizing text data. Wordcloud2 package allows you to use any image as the markup.


```
wordcloud2(top_kindle_highlights, figPath = bat, size = 1, backgroundColor =
"white", color = color_vector(data$freq) )
```

Some ideas started to get shaped in his mind. He thought who made those highlights is someone interested in storytelling, writing and good communication, good habits, and people. **Someone who wants to influence his life in a positive way. He was becoming more and more interested in the books.**

He wanted to dig deeper.

Bigram Analysis

Single words are a good starting point what the books were about. **But they are not very informative without context.** Frequency analysis can also be performed to measure how often pairs of two words **(bigrams)** occur in the text. This allows us to capture finer details in the text.

To do this he combined the unnested single tokens which is isolated above back into a continuous text and then performed bigram analysis. You can use **str_c()** function from stringr package to concatenate the single words.

```
# Recreate the df
df <- data.frame(highlights)
df$highlights <- str_replace_all(df$highlights, "’", "'")

df <- df %>% unnest_tokens(word, highlights) %>%
  anti_join(stop_words) %>%
 filter(!word %in% c("highlights","highlight", "page",
                     "location", "yellow", "pink", "orange", "blue",
                     "export", "hidden", "truncated", "kindle", "note",
"limits"))

df_com <- str_c(df$word, " ")
df_com <- data.frame(df_com)
```

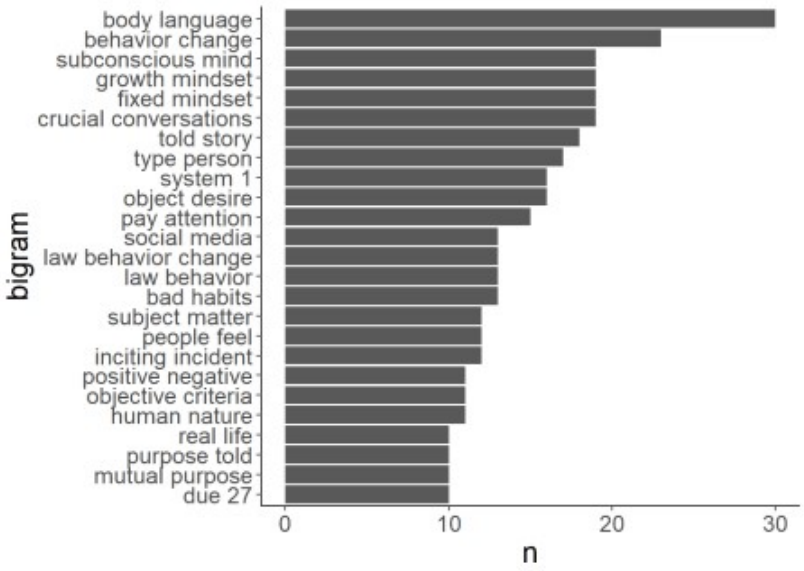
Let's split the text into bigrams and count the most common two word pairs.

```
df_bigram <- df_com %>%
 unnest_tokens(bigram, df_com, token = "ngrams",
 n = 3, n_min = 2)
top_bigrams <- df_bigram %>%
 group_by(bigram) %>%
 count() %>%
 arrange(desc(n))%>%
 print(n=20)
```

```
## # A tibble: 107,317 x 2
## # Groups:   bigram [107,317]
##    bigram                    n
##
##  1 body language            30
##  2 behavior change          23
##  3 crucial conversations    19
##  4 fixed mindset            19
##  5 growth mindset           19
##  6 subconscious mind        19
##  7 told story               18
##  8 type person              17
##  9 object desire            16
## 10 system 1                 16
## 11 pay attention            15
## 12 bad habits               13
## 13 law behavior             13
## 14 law behavior change      13
## 15 social media             13
## 16 inciting incident        12
## 17 people feel              12
## 18 subject matter           12
## 19 human nature             11
## 20 objective criteria       11
## # ... with 1.073e+05 more rows
```

```
# And visualize them on a plot

top <- top_bigrams[1:25,]
```

```
top %>% ungroup() %>% mutate(bigram = fct_reorder(bigram, n)) %>%
 ggplot(aes(x=bigram, y=n)) +
 geom_col() +
 coord_flip() +
 theme_classic() +
 theme(legend.position = "none",
 text = element_text(size=18))
```

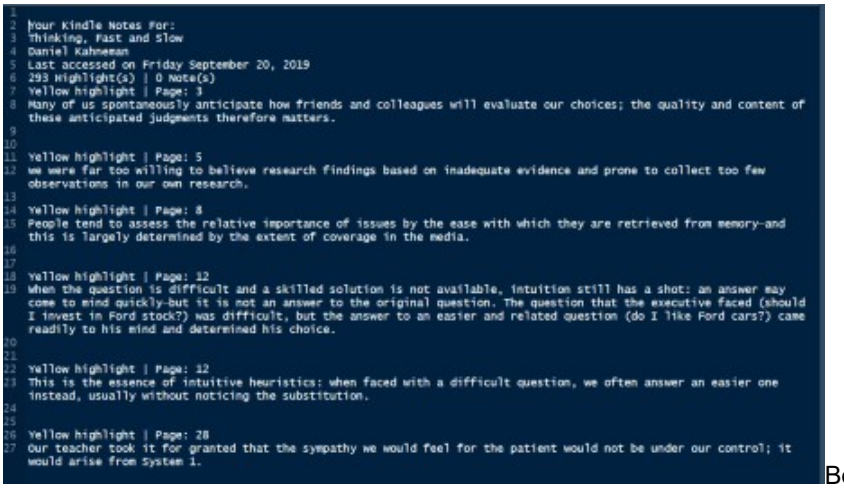


For example, if you go back above in the top 10 most frequent words table 6th word was change. But we didn't know what the change was about. And here we see that one of the most common bigram is behavioral change. It is making more sense. But it can improve to look at each book individually.

We can also do what we did for the whole document for highlights from single books.

But how can we capture them individually?

Let's first look at the text once more.



Before each book **"Your Kindle Notes For:"** appears.

Let's find out the line numbers for the beginning and the end of each book and use those indexes for fishing out each book.

We will reuse the data frame df we created above. **str_which()** function returns index numbers of the lines which contain a given pattern. In the last step, capturing the text between two consecutive indexes will give us the book between them.

```
# Since I modified df above. I will recreate it again.
df <- data.frame(highlights)
df$highlights <- str_replace_all(df$highlights, "’", "'")

# Getting the index number for each book

indexes <- str_which(df$highlights, pattern = fixed("Your Kindle Notes For"))
book_names <- df$highlights[indexes + 1]
indexes <-  c(indexes,nrow(df))
```

```
# Create an empty list

books <- list()

# Now the trick. Capture each 28 book separately in a list.

for(i in 1:(length(indexes)-1)) {
    books[[i]] <- data.frame(df$highlights[(indexes[i]:indexes[i+1]-1)])
    colnames(books[[i]]) <- "word_column"
    books[[i]]$word_column <- as.character(books[[i]]$word_column)
}
```

Let's check whether it worked, for example you can look up the 5th book on our list.

```
head(books[[5]])
```

```
##                                           word_column
## 1
## 2                               Your Kindle Notes For:
## 3 Bird by Bird: Some Instructions on Writing and Life
## 4                                           Anne Lamott
## 5             Last accessed on Saturday July 27, 2019
## 6                          75 Highlight(s) | 4 Note(s)
```

```
head(books[[15]])
```

```
##                                                 word_column
## 1
## 2                                     Your Kindle Notes For:
## 3 Getting to Yes: Negotiating an agreement without giving in
## 4                                   Roger Fisher and William Ury
## 5                 Last accessed on Saturday November 3, 2018
## 6                                   266 Highlight(s) | 3 Note(s)
```

Now, we have the individual books captured. I will repeat the procedure we used to analyse the whole text above to analyze each of the 28 books by using a for loop.

```
top <- list()
for(i in 1:28){
books[[i]] <- books[[i]] %>% unnest_tokens(word, word_column) %>%
             anti_join(stop_words) %>%
             filter(!word %in% c("highlights","highlight", "page",
                      "location", "yellow", "pink", "orange", "blue",
                      "export", "hidden", "truncated", "kindle", "note",
"limits"))

# Find out the top words in each book and capture them in a list (top)

top[[i]] <- books[[i]] %>%
              group_by(word) %>%
              count() %>%
              arrange(desc(n))
}
for(i in 1:28){
  print(book_names[[i]])
  print(top[[i]])
}
```

```
## [1] "Thinking, Fast and Slow"
## # A tibble: 1,619 x 2
```

```
## # Groups:   word [1,619]
##    word          n
##
##  1 people       33
##  2 system       26
##  3 1            18
##  4 mind         18
##  5 effect       17
##  6 bad          15
##  7 cognitive    15
##  8 ease         15
##  9 theory       13
## 10 decision     12
## # ... with 1,609 more rows
## [1] "Influence: The Psychology of Persuasion (Collins Business Essentials)"
## # A tibble: 278 x 2
## # Groups:   word [278]
##    word            n
##
##  1 142             5
##  2 146             5
##  3 131             3
##  4 147             3
##  5 154             3
##  6 179             3
##  7 association     3
##  8 food            3
##  9 information     3
## 10 people          3
## # ... with 268 more rows
## [1] "On Writing Well, 30th Anniversary Edition: An Informal Guide to Writing
Nonfiction"
## # A tibble: 770 x 2
## # Groups:   word [770]
##    word         n
##
##  1 writing     26
##  2 write       18
##  3 sentence    15
##  4 writer      15
##  5 reader      13
##  6 people      10
##  7 words        9
##  8 person       8
##  9 writers      8
## 10 day          7
## # ... with 760 more rows
## [1] "Wired for Story: The Writer's Guide to Using Brain Science to Hook
Readers from the Very First Sentence"
## # A tibble: 1,657 x 2
## # Groups:   word [1,657]
##    word             n
##
##  1 story          104
##  2 goal            41
##  3 protagonist     40
##  4 life            27
```

```
##  5 protagonist's    23
##  6 internal         21
##  7 brain            20
##  8 reader           20
##  9 external         19
## 10 world            19
## # ... with 1,647 more rows
## [1] "Bird by Bird: Some Instructions on Writing and Life"
## # A tibble: 522 x 2
## # Groups:   word [522]
##    word          n
##
##  1 writing      17
##  2 mind          7
##  3 bird          6
##  4 voices        5
##  5 attention     4
##  6 day           4
##  7 hope          4
##  8 life          4
##  9 makes         4
## 10 muscles       4
## # ... with 512 more rows
## [1] "Atomic Habits: An Easy and Proven Way to Build Good Habits and Break Bad
Ones"
## # A tibble: 2,736 x 2
## # Groups:   word [2,736]
##    word         n
##
##  1 habits     140
##  2 habit      110
##  3 behavior    94
##  4 change      73
##  5 people      50
##  6 time        47
##  7 identity    38
##  8 day         36
##  9 brain       32
## 10 person      32
## # ... with 2,726 more rows
## [1] "Storynomics: Story-Driven Marketing in the Post-Advertising World"
## # A tibble: 3,042 x 2
## # Groups:   word [3,042]
##    word          n
##
##  1 story       149
##  2 mind         50
##  3 stories      48
##  4 core         47
##  5 marketing    46
##  6 brand        45
##  7 life         42
##  8 change       41
##  9 audience     35
## 10 due          33
## # ... with 3,032 more rows
## [1] "Crucial Conversations Tools for Talking When Stakes Are High, Second
```

```
Edition"
## # A tibble: 1,828 x 2
## # Groups:   word [1,828]
##    word             n
##
##  1 people           84
##  2 dialogue         40
##  3 stories          40
##  4 due              34
##  5 feel             33
##  6 crucial          31
##  7 conversations    30
##  8 meaning          30
##  9 story            30
## 10 conversation     28
## # ... with 1,818 more rows
## [1] "Pre-Suasion: A Revolutionary Way to Influence and Persuade"
## # A tibble: 524 x 2
## # Groups:   word [524]
##    word             n
##
##  1 attention        6
##  2 influence        5
##  3 mental           5
##  4 trust            5
##  5 visitors         5
##  6 comfort          4
##  7 emotional        4
##  8 experience       4
##  9 message          4
## 10 associations     3
## # ... with 514 more rows
## [1] "Made to Stick: Why some ideas take hold and others come unstuck"
## # A tibble: 1,752 x 2
## # Groups:   word [1,752]
##    word          n
##
##  1 people       64
##  2 knowledge    27
##  3 story        25
##  4 ideas        24
##  5 concrete     18
##  6 surprise     17
##  7 care         16
##  8 time         15
##  9 attention    14
## 10 core         14
## # ... with 1,742 more rows
## [1] "The Charisma Myth: Master the Art of Personal Magnetism"
## # A tibble: 1,802 x 2
## # Groups:   word [1,802]
##    word            n
##
##  1 feel           43
##  2 body           38
##  3 people         35
##  4 language       33
```

```
##  5 charisma       27
##  6 warmth         27
##  7 charismatic    24
##  8 power          22
##  9 person         19
## 10 confidence     18
## # ... with 1,792 more rows
## [1] "The Power of Moments: Why Certain Experiences Have Extraordinary Impact"
## # A tibble: 1,299 x 2
## # Groups:   word [1,299]
##    word             n
##
##  1 moments          29
##  2 moment           21
##  3 people           17
##  4 time             15
##  5 insight          13
##  6 milestones       13
##  7 purpose          11
##  8 relationships    11
##  9 create            9
## 10 goal              9
## # ... with 1,289 more rows
## [1] "Principles: Life and Work"
## # A tibble: 1,131 x 2
## # Groups:   word [1,131]
##    word           n
##
##  1 people        54
##  2 thinking      16
##  3 decision      12
##  4 level         12
##  5 life          12
##  6 pain          12
##  7 habits        11
##  8 understand    11
##  9 change        10
## 10 knowing       10
## # ... with 1,121 more rows
## [1] "Deep Work: Rules for Focused Success in a Distracted World"
## # A tibble: 711 x 2
## # Groups:   word [711]
##    word          n
##
##  1 attention    12
##  2 deep         11
##  3 ability       9
##  4 book          9
##  5 life          9
##  6 time          9
##  7 mind          7
##  8 world         7
##  9 focus         6
## 10 called        5
## # ... with 701 more rows
## [1] "Getting to Yes: Negotiating an agreement without giving in"
## # A tibble: 1,489 x 2
```

```
## # Groups:   word [1,489]
##    word            n
##
##  1 agreement      33
##  2 negotiation    33
##  3 options        23
##  4 people         19
##  5 objective      17
##  6 positions      17
##  7 ideas          16
##  8 position       15
##  9 shared         15
## 10 solution       15
## # ... with 1,479 more rows
## [1] "Who: The A Method for Hiring"
## # A tibble: 920 x 2
## # Groups:   word [920]
##    word           n
##
##  1 people        38
##  2 job           22
##  3 players       16
##  4 person        15
##  5 candidate     14
##  6 candidates    13
##  7 company       13
##  8 hire          11
##  9 hiring        11
## 10 interview     11
## # ... with 910 more rows
## [1] "Mindset: Changing The Way You think To Fulfil Your Potential"
## # A tibble: 910 x 2
## # Groups:   word [910]
##    word         n
##
##  1 mindset     43
##  2 people      33
##  3 growth      27
##  4 fixed       23
##  5 blame       18
##  6 learning    16
##  7 learn       15
##  8 effort      11
##  9 failure     11
## 10 makes       10
## # ... with 900 more rows
## [1] "The 4-Hour Work Week: Escape the 9-5, Live Anywhere and Join the New
Rich"
## # A tibble: 736 x 2
## # Groups:   word [736]
##    word          n
##
##  1 time         11
##  2 life         10
##  3 mail          6
##  4 product       6
##  5 week          6
```

```
##  6 world         6
##  7 baby          5
##  8 celebrity     5
##  9 create        5
## 10 days          5
## # ... with 726 more rows
## [1] "Tools of Titans: The Tactics, Routines, and Habits of Billionaires,
Icons, and World-Class Performers"
## # A tibble: 1,956 x 2
## # Groups:   word [1,956]
##    word       n
##
##  1 people    41
##  2 life      25
##  3 time      24
##  4 write     24
##  5 world     22
##  6 10        17
##  7 ideas     14
##  8 book      13
##  9 times     12
## 10 read      11
## # ... with 1,946 more rows
## [1] "The Elements of Eloquence: How to Turn the Perfect English Phrase"
## # A tibble: 116 x 2
## # Groups:   word [116]
##    word           n
##
##  1 change         5
##  2 english        3
##  3 pattern        3
##  4 poets          3
##  5 19             2
##  6 44             2
##  7 attitude       2
##  8 colour         2
##  9 contradict     2
## 10 fall           2
## # ... with 106 more rows
## [1] "The One Thing: The Surprisingly Simple Truth Behind Extraordinary
Results: Achieve your goals with one of the world's bestselling success books
(Basic Skills)"
## # A tibble: 587 x 2
## # Groups:   word [587]
##    word             n
##
##  1 time            29
##  2 success         15
##  3 results         11
##  4 block            9
##  5 day              9
##  6 extraordinary    9
##  7 life             8
##  8 matters          8
##  9 successful       8
## 10 discipline       7
## # ... with 577 more rows
```

```
## [1] "How to Win Friends and Influence People"
## # A tibble: 140 x 2
## # Groups:   word [140]
##    word          n
##
##  1 people        7
##  2 ability       3
##  3 fears         3
##  4 116           2
##  5 book          2
##  6 human         2
##  7 knowledge     2
##  8 meeting       2
##  9 person's      2
## 10 sell          2
## # ... with 130 more rows
## [1] "The Untethered Soul: The Journey Beyond Yourself"
## # A tibble: 770 x 2
## # Groups:   word [770]
##    word           n
##
##  1 life          73
##  2 feel          34
##  3 events        26
##  4 mind          25
##  5 world         20
##  6 fear          19
##  7 inside        19
##  8 energy        17
##  9 experience    17
## 10 heart         17
## # ... with 760 more rows
## [1] "Man's Search For Meaning: The classic tribute to hope from the
Holocaust"
## # A tibble: 894 x 2
## # Groups:   word [894]
##    word            n
##
##  1 life           29
##  2 suffering      24
##  3 meaning        20
##  4 human          19
##  5 intention      11
##  6 75              9
##  7 logotherapy     9
##  8 patient         9
##  9 world           9
## 10 called          8
## # ... with 884 more rows
## [1] "The Power of your Subconscious Mind and Other Works"
## # A tibble: 600 x 2
## # Groups:   word [600]
##    word             n
##
##  1 mind            34
##  2 subconscious    28
##  3 wealth          13
```

```
##  4 idea            11
##  5 mental          10
##  6 love             9
##  7 life             8
##  8 peace            8
##  9 happiness        7
## 10 desire           6
## # ... with 590 more rows
## [1] "Ego is the Enemy: The Fight to Master Our Greatest Opponent"
## # A tibble: 831 x 2
## # Groups:   word [831]
##    word         n
##
##  1 ego         19
##  2 people      12
##  3 purpose     11
##  4 111          8
##  5 change       7
##  6 147          6
##  7 function     6
##  8 life         6
##  9 passion      6
## 10 path         6
## # ... with 821 more rows
## [1] "Outliers: The Story of Success"
## # A tibble: 105 x 2
## # Groups:   word [105]
##    word            n
##
##  1 ability         3
##  2 knowing         3
##  3 sense           3
##  4 communicate     2
##  5 distance        2
##  6 family          2
##  7 intelligence    2
##  8 power           2
##  9 practical       2
## 10 sternberg       2
## # ... with 95 more rows
## [1] "The Start-up of You: Adapt to the Future, Invest in Yourself, and
Transform Your Career"
## # A tibble: 570 x 2
## # Groups:   word [570]
##    word             n
##
##  1 people          14
##  2 product          8
##  3 opportunities    7
##  4 person           7
##  5 start            7
##  6 assets           6
##  7 job              6
##  8 time             6
##  9 138              5
## 10 create           5
## # ... with 560 more rows
```

Now, looking at the frequent words from each book we can get more insights what they are about.

The bigrams for the same books.

```
df <- data.frame(highlights)
df$highlights <- str_replace_all(df$highlights, "’", "'")

# Getting the index number for each book

indexes <- str_which(df$highlights, pattern = fixed("Your Kindle Notes For"))
book_names <- df$highlights[indexes + 1]
indexes <-  c(indexes,nrow(df))

# Capturing each book individually

books <- list()
for (i in 1:(length(indexes)-1)) {
    books[[i]] <- data.frame(df$highlights[(indexes[i]:indexes[i+1]-1)])
    colnames(books[[i]]) <- "word_column"
    books[[i]]$word_column <- as.character(books[[i]]$word_column)
}

# Next step in the plan was splitting the text into single words by
unnest_tokens function.



for(i in 1:28){
books[[i]] <- books[[i]] %>% unnest_tokens(word, word_column) %>%
             anti_join(stop_words) %>%
             filter(!word %in% c("highlights","highlight", "page",
                      "location", "yellow", "pink", "orange", "blue",
                      "export", "hidden", "truncated", "kindle", "note",
"limits"))
}

# After this preparation step I can combine the single words back into a
continous text

for(i in 1:28){
books[[i]] <- str_c(books[[i]]$word, " ")
books[[i]] <- data.frame(books[[i]])
}


df_bigram <- list()

for(i in 1:28){
df_bigram[[i]] <- books[[i]] %>%
       unnest_tokens(bigram, books..i.., token = "ngrams",
                                     n = 3, n_min = 2)
}

for (i in 1:28){
  print(book_names[i])
df_bigram[[i]] %>%
  group_by(bigram) %>%
  count() %>%
```

```
  arrange(desc(n))%>%
  print(n=10)

}
```

```
## [1] "Thinking, Fast and Slow"
## # A tibble: 5,768 x 2
## # Groups:   bigram [5,768]
##    bigram                   n
##
##  1 system 1                16
##  2 cognitive ease           9
##  3 system 2                 8
##  4 halo effect              4
##  5 loss aversion            4
##  6 possibility effect       4
##  7 affective forecasting    3
##  8 availability bias        3
##  9 cognitive strain         3
## 10 decision weights         3
## # ... with 5,758 more rows
## [1] "Influence: The Psychology of Persuasion (Collins Business Essentials)"
## # A tibble: 673 x 2
## # Groups:   bigram [673]
##    bigram                   n
##
##  1 association principle    2
##  2 click whirr              2
##  3 click whirr response     2
##  4 luncheon technique       2
##  5 reciprocity rule         2
##  6 whirr response           2
##  7 0 13                     1
##  8 0 13 rule                1
##  9 13 rule                  1
## 10 13 rule reciprocation    1
## # ... with 663 more rows
## [1] "On Writing Well, 30th Anniversary Edition: An Informal Guide to Writing
Nonfiction"
## # A tibble: 2,172 x 2
## # Groups:   bigram [2,172]
##    bigram                       n
##
##  1 500th appendix               2
##  2 choice unity                 2
##  3 confronted solved            2
##  4 despair finding              2
##  5 despair finding solution     2
##  6 english language             2
##  7 federal buildings            2
##  8 finally solve                2
##  9 finally solve surgeon        2
## 10 finding solution             2
## # ... with 2,162 more rows
## [1] "Wired for Story: The Writer's Guide to Using Brain Science to Hook
Readers from the Very First Sentence"
## # A tibble: 6,602 x 2
## # Groups:   bigram [6,602]
```

```
##    bigram                   n
##
##  1 external goal            8
##  2 internal goal            6
##  3 cognitive unconscious    5
##  4 internal issue           5
##  5 real life                5
##  6 story question           5
##  7 antonio damasio          4
##  8 effect trajectory        4
##  9 steven pinker            4
## 10 1 story                  3
## # ... with 6,592 more rows
## [1] "Bird by Bird: Some Instructions on Writing and Life"
## # A tibble: 1,304 x 2
## # Groups:   bigram [1,304]
##    bigram                  n
##
##  1 bird bird               3
##  2 muscles cramp           3
##  3 cramp wounds            2
##  4 life view               2
##  5 likable narrator        2
##  6 muscles cramp wounds    2
##  7 pay attention           2
##  8 1,015 read              1
##  9 1,015 read reading      1
## 10 1,048 digress           1
## # ... with 1,294 more rows
## [1] "Atomic Habits: An Easy and Proven Way to Build Good Habits and Break Bad
Ones"
## # A tibble: 12,309 x 2
## # Groups:   bigram [12,309]
##    bigram                 n
##
##  1 behavior change       23
##  2 type person           17
##  3 law behavior          13
##  4 law behavior change   13
##  5 bad habits            11
##  6 social media           9
##  7 habits attractive      6
##  8 3rd law                5
##  9 bad habit              5
## 10 break chain            5
## # ... with 1.23e+04 more rows
## [1] "Storynomics: Story-Driven Marketing in the Post-Advertising World"
## # A tibble: 12,819 x 2
## # Groups:   bigram [12,819]
##    bigram                 n
##
##  1 object desire         16
##  2 told story            16
##  3 inciting incident     12
##  4 positive negative     10
##  5 purpose told          10
##  6 subject matter        10
```

```
##  7 core character         9
##  8 purpose told story     8
##  9 real beauty            7
## 10 change team's          6
## # ... with 1.281e+04 more rows
## [1] "Crucial Conversations Tools for Talking When Stakes Are High, Second
Edition"
## # A tibble: 8,751 x 2
## # Groups:   bigram [8,751]
##    bigram                    n
##
##  1 crucial conversations    19
##  2 due 27                   10
##  3 mutual purpose           10
##  4 shared pool               8
##  5 silence violence          8
##  6 crucial conversation      7
##  7 path action               7
##  8 due 26                    6
##  9 due 43                    6
## 10 fool's choice             6
## # ... with 8,741 more rows
## [1] "Pre-Suasion: A Revolutionary Way to Influence and Persuade"
## # A tibble: 1,261 x 2
## # Groups:   bigram [1,261]
##    bigram                     n
##
##  1 attention goal             2
##  2 concept audience           2
##  3 levels importance          2
##  4 mandel johnson             2
##  5 mental activity            2
##  6 social proof               2
##  7 thousand dollars           2
##  8 twenty thousand            2
##  9 twenty thousand dollars    2
## 10 writing session            2
## # ... with 1,251 more rows
## [1] "Made to Stick: Why some ideas take hold and others come unstuck"
## # A tibble: 6,372 x 2
## # Groups:   bigram [6,372]
##    bigram                     n
##
##  1 curse knowledge            7
##  2 guessing machines          6
##  3 people care                6
##  4 goodyear tires             5
##  5 knowledge gaps             5
##  6 people's attention         5
##  7 popcorn popper             5
##  8 security goodyear          5
##  9 security goodyear tires    5
## 10 sinatra test               5
## # ... with 6,362 more rows
## [1] "The Charisma Myth: Master the Art of Personal Magnetism"
## # A tibble: 6,343 x 2
## # Groups:   bigram [6,343]
```

```
##    bigram                      n
##
##  1 body language               30
##  2 power warmth                 6
##  3 feel bad                     4
##  4 imagination reality          4
##  5 people feel                  4
##  6 responsibility transfer      4
##  7 charismatic body             3
##  8 charismatic body language    3
##  9 confidence ability           3
## 10 distinguish imagination      3
## # ... with 6,333 more rows
## [1] "The Power of Moments: Why Certain Experiences Have Extraordinary Impact"
## # A tibble: 3,967 x 2
## # Groups:   bigram [3,967]
##    bigram                       n
##
##  1 defining moments             5
##  2 backward integrated          3
##  3 backward integrated design   3
##  4 breaking script              3
##  5 connecting meaning           3
##  6 integrated design            3
##  7 moments pride                3
##  8 understanding validation     3
##  9 bad stronger                 2
## 10 bose headphones              2
## # ... with 3,957 more rows
## [1] "Principles: Life and Work"
## # A tibble: 3,960 x 2
## # Groups:   bigram [3,960]
##    bigram                  n
##
##  1 common sense            3
##  2 left brained            3
##  3 responsible parties     3
##  4 134 people              2
##  5 274 remember            2
##  6 407 values              2
##  7 407 values abilities    2
##  8 achieve goals           2
##  9 bad outcomes            2
## 10 blind spots             2
## # ... with 3,950 more rows
## [1] "Deep Work: Rules for Focused Success in a Distracted World"
## # A tibble: 1,981 x 2
## # Groups:   bigram [1,981]
##    bigram                 n
##
##  1 deliberate practice    4
##  2 13 master              2
##  3 14 deep                2
##  4 29 ability             2
##  5 77 gallagher           2
##  6 ability concentrate    2
##  7 anders ericsson        2
```

```
##  8 book shining            2
##  9 choose focus            2
## 10 fixed schedule          2
## # ... with 1,971 more rows
## [1] "Getting to Yes: Negotiating an agreement without giving in"
## # A tibble: 5,363 x 2
## # Groups:   bigram [5,363]
##    bigram                    n
##
##  1 objective criteria       11
##  2 principled negotiation    8
##  3 bottom line               6
##  4 inventing options         6
##  5 mutual gain               6
##  6 reach agreement           6
##  7 reaching agreement        6
##  8 options mutual            5
##  9 options mutual gain       5
## 10 brainstorming session     4
## # ... with 5,353 more rows
## [1] "Who: The A Method for Hiring"
## # A tibble: 3,196 x 2
## # Groups:   bigram [3,196]
##    bigram                    n
##
##  1 talented people           6
##  2 outcomes competencies     4
##  3 96 performance            3
##  4 96 performance compare    3
##  5 fit company               3
##  6 performance compare       3
##  7 2 million                 2
##  8 95 interrupt              2
##  9 career goals              2
## 10 company 31                2
## # ... with 3,186 more rows
## [1] "Mindset: Changing The Way You think To Fulfil Your Potential"
## # A tibble: 3,182 x 2
## # Groups:   bigram [3,182]
##    bigram                   n
##
##  1 fixed mindset           19
##  2 growth mindset          19
##  3 people fixed             4
##  4 people fixed mindset     4
##  5 183 son                  3
##  6 assign blame             3
##  7 social interactions      3
##  8 142 create               2
##  9 157 fixed                2
## 10 157 fixed mindset        2
## # ... with 3,172 more rows
## [1] "The 4-Hour Work Week: Escape the 9-5, Live Anywhere and Join the New
Rich"
## # A tibble: 1,927 x 2
## # Groups:   bigram [1,927]
##    bigram                    n
```

```
##
##  1 http e.ggtimer.com          3
##  2 basic assumptions           2
##  3 car seat                    2
##  4 limit tasks                 2
##  5 offer customer              2
##  6 options offer               2
##  7 options offer customer      2
##  8 parkinson's law             2
##  9 shorten time                2
## 10 suggest days                2
## # ... with 1,917 more rows
## [1] "Tools of Titans: The Tactics, Routines, and Habits of Billionaires,
Icons, and World-Class Performers"
## # A tibble: 6,321 x 2
## # Groups:   bigram [6,321]
##    bigram                  n
##
##  1 10 ideas                4
##  2 bad ideas               4
##  3 keeping track           4
##  4 track times             4
##  5 world war               4
##  6 516 write               3
##  7 extreme ownership       3
##  8 heart head              3
##  9 keeping track times     3
## 10 narrative narrative     3
## # ... with 6,311 more rows
## [1] "The Elements of Eloquence: How to Turn the Perfect English Phrase"
## # A tibble: 272 x 2
## # Groups:   bigram [272]
##    bigram                    n
##
##  1 change attitude           2
##  2 change pattern            2
##  3 change pattern change     2
##  4 fall love                 2
##  5 pattern change            2
##  6 0 19                      1
##  7 0 19 bred                 1
##  8 11 2018                   1
##  9 11 2018 8                 1
## 10 19 bred                   1
## # ... with 262 more rows
## [1] "The One Thing: The Surprisingly Simple Truth Behind Extraordinary
Results: Achieve your goals with one of the world's bestselling success books
(Basic Skills)"
## # A tibble: 1,629 x 2
## # Groups:   bigram [1,629]
##    bigram                    n
##
##  1 extraordinary results     7
##  2 time block                7
##  3 selected discipline       3
##  4 3 time                    2
##  5 3 time block              2
```

```
##  6 achieve extraordinary     2
##  7 block day                 2
##  8 default settings          2
##  9 discipline build          2
## 10 easier unnecessary        2
## # ... with 1,619 more rows
## [1] "How to Win Friends and Influence People"
## # A tibble: 318 x 2
## # Groups:   bigram [318]
##    bigram               n
##
##  1 time meeting         2
##  2 0 72                 1
##  3 0 72 lies            1
##  4 110 people           1
##  5 110 people smile     1
##  6 112 time             1
##  7 112 time meeting     1
##  8 116 116              1
##  9 116 116 bad          1
## 10 116 bad              1
## # ... with 308 more rows
## [1] "The Untethered Soul: The Journey Beyond Yourself"
## # A tibble: 3,195 x 2
## # Groups:   bigram [3,195]
##    bigram                   n
##
##  1 preconceived notions     8
##  2 life avoiding            7
##  3 devote life              6
##  4 empty space              5
##  5 experience life          5
##  6 model reality            5
##  7 rest life                5
##  8 spend life               5
##  9 spend life avoiding      5
## 10 153 events               4
## # ... with 3,185 more rows
## [1] "Man's Search For Meaning: The classic tribute to hope from the
Holocaust"
## # A tibble: 2,917 x 2
## # Groups:   bigram [2,917]
##    bigram                    n
##
##  1 paradoxical intention     6
##  2 hyper intention           4
##  3 anticipatory anxiety      3
##  4 existential vacuum        3
##  5 fall asleep               3
##  6 human existence           3
##  7 intention fall            3
##  8 intention fall asleep     3
##  9 meaning life              3
## 10 potential meaning         3
## # ... with 2,907 more rows
## [1] "The Power of your Subconscious Mind and Other Works"
## # A tibble: 1,750 x 2
```

```
## # Groups:   bigram [1,750]
##    bigram                        n
##
##  1 subconscious mind            17
##  2 dominant idea                 4
##  3 idea subconscious             3
##  4 peace mind                    3
##  5 power subconscious            3
##  6 accept idea                   2
##  7 accepted subconscious         2
##  8 accepted subconscious mind    2
##  9 annoy irritate                2
## 10 annoy irritate permit         2
## # ... with 1,740 more rows
## [1] "Ego is the Enemy: The Fight to Master Our Greatest Opponent"
## # A tibble: 2,290 x 2
## # Groups:   bigram [2,290]
##    bigram                n
##
##  1 112 start             2
##  2 147 deceived          2
##  3 33 purpose            2
##  4 beat people           2
##  5 ego enemy             2
##  6 function function     2
##  7 people beat           2
##  8 people beat people    2
##  9 people beneath        2
## 10 purpose realism       2
## # ... with 2,280 more rows
## [1] "Outliers: The Story of Success"
## # A tibble: 232 x 2
## # Groups:   bigram [232]
##    bigram                          n
##
##  1 knowing knowing                 2
##  2 power distance                  2
##  3 practical intelligence          2
##  4 0 884                           1
##  5 0 884 write                     1
##  6 1,051 robert                    1
##  7 1,051 robert sternberg          1
##  8 1,052 practical                 1
##  9 1,052 practical intelligence    1
## 10 1,063 annette                   1
## # ... with 222 more rows
## [1] "The Start-up of You: Adapt to the Future, Invest in Yourself, and
Transform Your Career"
## # A tibble: 1,611 x 2
## # Groups:   bigram [1,611]
##    bigram                n
##
##  1 product management    3
##  2 faster cheaper        2
##  3 skills experiences    2
##  4 soft assets           2
##  5 weak ties             2
```

```
##  6 0 15                   1
##  7 0 15 paranoid          1
##  8 101 business           1
##  9 101 business crazy     1
## 10 101 inspired           1
## # ... with 1,601 more rows
```

If you want to see another example of this capturing process you can have a look at my recent post here.

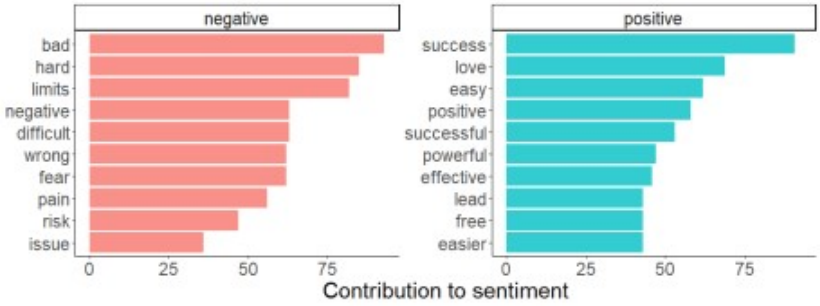
Looking at each book individually, he started to be more and more obsessed about the books in my kindle. He decided to order a couple of them.

Sentiment analysis is used to evaluate emotional charge in a text mining project. Most common uses are social media monitoring, customer experience management, and Voice of Customer, to understand how they feel.

The **bing** lexicon categorizes words into positive and negative categories, in a binary fashion. The **nrc** lexicon uses categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

Using bing lexicon

This gives us the top words contributed to each emotional category. Some examples to note are success, effective, for positive and bad, hard and limits.

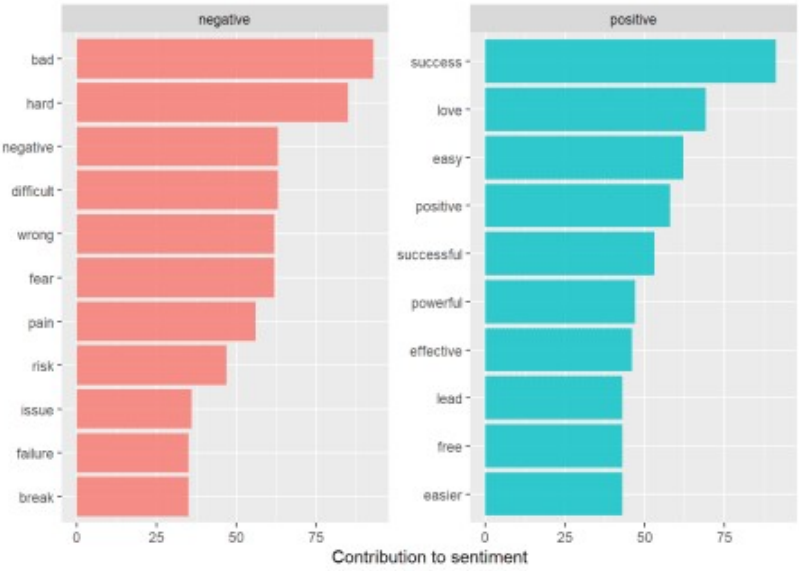


Here is how R produced the above plot:

```
df <- data.frame(highlights)
df$highlights <- str_replace_all(df$highlights, "’", "'")
df <- df %>% unnest_tokens(word, highlights) %>%
  anti_join(stop_words) %>%
 filter(!word %in% c("highlights","highlight", "page",
                     "location", "yellow", "pink", "orange", "blue",
                     "export", "hidden", "truncated", "kindle", "note",
"limits"))

bing_word_counts <- df %>% inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()

bing_word_counts
```

```
## # A tibble: 1,854 x 3
##    word      sentiment     n
##
##  1 bad       negative     93
##  2 success   positive     91
##  3 hard      negative     85
##  4 love      positive     69
##  5 difficult negative     63
##  6 negative  negative     63
```

```
##  7 easy      positive     62
##  8 fear      negative     62
##  9 wrong     negative     62
## 10 positive  positive     58
## # ... with 1,844 more rows

# Top contributors to positive and negative sentiment


bing <- bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ggplot(aes(reorder(word, n), n, fill=sentiment)) +
  geom_bar(alpha=0.8, stat="identity", show.legend = FALSE)+
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y= "Contribution to sentiment", x = NULL) +
  coord_flip()
bing
```



Using nrc lexion

We see that I am more likely to highlight if a text is charged with positive rather than negative sentiment, and individually trust, anticipation and joy rather than fear and sadness.

```
df <- data.frame(highlights)
df$highlights <- str_replace_all(df$highlights, "’", "'")
df <- df %>% unnest_tokens(word, highlights) %>%
  anti_join(stop_words) %>%
 filter(!word %in% c("highlights","highlight", "page",
                     "location", "yellow", "pink", "orange", "blue",
                     "export", "hidden", "truncated", "kindle", "note",
"limits"))

## Joining, by = "word"

sentiment <- df %>%
        left_join(get_sentiments("nrc")) %>%
        filter(!is.na(sentiment)) %>%
        count(sentiment, sort = TRUE)

## Joining, by = "word"
```

```
sentiment
```

```
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive      8326
##  2 trust         4165
##  3 negative      3860
##  4 anticipation  3366
##  5 joy           2642
##  6 fear          2446
##  7 sadness       1844
##  8 anger         1799
##  9 surprise      1339
## 10 disgust       1093
```

Normalized sentiments

One important thing to add, since each emotion category has different number of words in a language. Emotional categories with less words are less likely to appear in a given text. Thus, I would like to normalize them according to their numbers in the lexicon and see how it differs than the above results.

```
# I will add numbers of each categories from the NRC lexicon

lexicon <- c(2317, 3338, 1234, 842, 1483, 691, 1250, 1195, 1060, 535)
polarity <-  c(1,1,1,1,1,0,0,0,0,0)
sentiment <- data.frame(sentiment, lexicon)
norm_sentiment <- sentiment %>% mutate( normalized = n/lexicon) %>%
arrange(desc(normalized))
sentiment <- data.frame(norm_sentiment, polarity)
sentiment
```

```
##       sentiment    n lexicon normalized polarity
## 1  anticipation 3366     842   3.997625        1
## 2      positive 8326    2317   3.593440        1
## 3          fear 2446     691   3.539797        1
## 4      negative 3860    1234   3.128039        1
## 5       disgust 1093     535   2.042991        1
## 6           joy 2642    1483   1.781524        0
## 7         anger 1799    1195   1.505439        0
## 8       sadness 1844    1250   1.475200        0
## 9      surprise 1339    1060   1.263208        0
## 10        trust 4165    3338   1.247753        0
```

```
# General findings

sentiment %>% group_by(polarity) %>% summarize(n2 = sum(lexicon))
```

```
## # A tibble: 2 x 2
##   polarity    n2
##
## 1        0  8326
## 2        1  5619
```

Now, **anticipation** is the highest emotion found in the text that I highlighted. This does not seem a coincidence to me. Since most of the books in our analysis is about productivity and self-development. The productivity tips and tools usually contain words associated with anticipation.

In a similar way, I can look at the sentiment for individual books

```
df <- data.frame(highlights)

# Kindle uses apostrophes (’), but stop_words uses sigle quotes (')
# To be able to use all stop_words I should replace apostrophes with quotes
df$highlights <- str_replace_all(df$highlights, "’", "'")

# Getting the index number for each book

indexes <- str_which(df$highlights, pattern = fixed("Your Kindle Notes For"))
book_names <- df$highlights[indexes + 1]
indexes <-  c(indexes,nrow(df))

# Capturing each book individually

books <- list()
for (i in 1:(length(indexes)-1)) {
    books[[i]] <- data.frame(df$highlights[(indexes[i]:indexes[i+1]-1)])
    colnames(books[[i]]) <- "word_column"
    books[[i]]$word_column <- as.character(books[[i]]$word_column)
}

# Next step in the plan was splitting the text into single words by
unnest_tokens function.


for(i in 1:28){
books[[i]] <- books[[i]] %>% unnest_tokens(word, word_column) %>%
             anti_join(stop_words) %>%
             filter(!word %in% c("highlights","highlight", "page",
                      "location", "yellow", "pink", "orange", "blue"))
}

sentiment <- list()
for (i in 1:28){
sentiment[[i]] <- books[[i]] %>%
        left_join(get_sentiments("nrc")) %>%
        filter(!is.na(sentiment)) %>%
        count(sentiment, sort = TRUE)
        print(book_names[i])
        print(sentiment[[i]])
}
```

```
## [1] "Thinking, Fast and Slow"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       450
##  2 trust          256
##  3 negative       254
##  4 anticipation   163
##  5 fear           153
##  6 sadness        116
##  7 joy            107
##  8 anger          104
##  9 disgust         81
## 10 surprise        75
## [1] "Influence: The Psychology of Persuasion (Collins Business Essentials)"
## # A tibble: 10 x 2
```

```
##    sentiment        n
##
##  1 positive        53
##  2 trust           37
##  3 joy             15
##  4 negative        15
##  5 fear            12
##  6 anticipation    11
##  7 sadness          8
##  8 anger            7
##  9 surprise         3
## 10 disgust          2
## [1] "On Writing Well, 30th Anniversary Edition: An Informal Guide to Writing
Nonfiction"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       172
##  2 negative        98
##  3 trust           81
##  4 anticipation    63
##  5 anger           48
##  6 fear            47
##  7 disgust         42
##  8 sadness         42
##  9 joy             37
## 10 surprise        26
## [1] "Wired for Story: The Writer's Guide to Using Brain Science to Hook
Readers from the Very First Sentence"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       413
##  2 negative       197
##  3 trust          178
##  4 anticipation   168
##  5 fear           152
##  6 joy            116
##  7 sadness        108
##  8 anger           96
##  9 surprise        84
## 10 disgust         41
## [1] "Bird by Bird: Some Instructions on Writing and Life"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive        77
##  2 negative        55
##  3 anticipation    40
##  4 trust           38
##  5 joy             37
##  6 fear            30
##  7 sadness         27
##  8 disgust         17
##  9 surprise        16
## 10 anger           15
## [1] "Atomic Habits: An Easy and Proven Way to Build Good Habits and Break Bad
```

```
Ones"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       835
##  2 trust          455
##  3 anticipation   439
##  4 negative       356
##  5 joy            296
##  6 fear           254
##  7 sadness        180
##  8 anger          147
##  9 surprise       139
## 10 disgust        117
## [1] "Storynomics: Story-Driven Marketing in the Post-Advertising World"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       860
##  2 trust          405
##  3 negative       364
##  4 anticipation   335
##  5 joy            250
##  6 fear           221
##  7 sadness        171
##  8 anger          167
##  9 surprise       166
## 10 disgust         76
## [1] "Crucial Conversations Tools for Talking When Stakes Are High, Second
Edition"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       758
##  2 negative       496
##  3 trust          412
##  4 fear           282
##  5 anticipation   258
##  6 anger          243
##  7 joy            216
##  8 sadness        196
##  9 disgust        142
## 10 surprise       108
## [1] "Pre-Suasion: A Revolutionary Way to Influence and Persuade"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive        84
##  2 trust           51
##  3 negative        31
##  4 anticipation    27
##  5 fear            24
##  6 joy             22
##  7 anger           14
##  8 sadness         12
##  9 surprise         9
## 10 disgust          3
```

```
## [1] "Made to Stick: Why some ideas take hold and others come unstuck"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       499
##  2 trust          236
##  3 anticipation   198
##  4 negative       167
##  5 joy            156
##  6 fear           123
##  7 surprise       107
##  8 sadness         74
##  9 anger           65
## 10 disgust         60
## [1] "The Charisma Myth: Master the Art of Personal Magnetism"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       483
##  2 negative       254
##  3 trust          228
##  4 joy            166
##  5 anticipation   162
##  6 fear           157
##  7 sadness        143
##  8 anger          120
##  9 surprise        65
## 10 disgust         58
## [1] "The Power of Moments: Why Certain Experiences Have Extraordinary Impact"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       294
##  2 trust          132
##  3 anticipation   123
##  4 negative       106
##  5 joy             96
##  6 fear            72
##  7 anger           52
##  8 surprise        50
##  9 sadness         45
## 10 disgust         19
## [1] "Principles: Life and Work"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       313
##  2 trust          178
##  3 negative       129
##  4 anticipation   120
##  5 joy            103
##  6 sadness         80
##  7 fear            78
##  8 anger           53
##  9 surprise        50
## 10 disgust         35
## [1] "Deep Work: Rules for Focused Success in a Distracted World"
```

```
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       176
##  2 trust           69
##  3 anticipation    54
##  4 negative        36
##  5 joy             32
##  6 fear            19
##  7 sadness         14
##  8 surprise        14
##  9 anger           12
## 10 disgust          7
## [1] "Getting to Yes: Negotiating an agreement without giving in"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       444
##  2 trust          234
##  3 negative       180
##  4 anticipation   135
##  5 anger          103
##  6 fear           100
##  7 joy             83
##  8 sadness         68
##  9 surprise        48
## 10 disgust         38
## [1] "Who: The A Method for Hiring"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       259
##  2 trust          125
##  3 anticipation    95
##  4 joy             73
##  5 negative        68
##  6 fear            30
##  7 surprise        29
##  8 anger           25
##  9 sadness         22
## 10 disgust         16
## [1] "Mindset: Changing The Way You think To Fulfil Your Potential"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       317
##  2 trust          160
##  3 negative       134
##  4 joy            117
##  5 anticipation   100
##  6 fear            78
##  7 anger           70
##  8 sadness         65
##  9 disgust         57
## 10 surprise        44
## [1] "The 4-Hour Work Week: Escape the 9-5, Live Anywhere and Join the New
Rich"
```

```
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       131
##  2 anticipation    70
##  3 negative        64
##  4 trust           57
##  5 joy             56
##  6 fear            34
##  7 surprise        27
##  8 anger           24
##  9 sadness         20
## 10 disgust         14
## [1] "Tools of Titans: The Tactics, Routines, and Habits of Billionaires,
Icons, and World-Class Performers"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       406
##  2 negative       251
##  3 trust          199
##  4 anticipation   188
##  5 fear           134
##  6 joy            126
##  7 anger          111
##  8 sadness        108
##  9 surprise        78
## 10 disgust         74
## [1] "The Elements of Eloquence: How to Turn the Perfect English Phrase"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive        18
##  2 negative        13
##  3 fear             9
##  4 trust            9
##  5 joy              6
##  6 sadness          6
##  7 anger            5
##  8 anticipation     3
##  9 disgust          2
## 10 surprise         2
## [1] "The One Thing: The Surprisingly Simple Truth Behind Extraordinary
Results: Achieve your goals with one of the world's bestselling success books
(Basic Skills)"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       139
##  2 anticipation    97
##  3 trust           60
##  4 joy             56
##  5 negative        32
##  6 fear            20
##  7 anger           14
##  8 surprise        14
##  9 disgust          9
```

```
## 10 sadness          9
## [1] "How to Win Friends and Influence People"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive        33
##  2 trust           14
##  3 negative        11
##  4 anticipation    10
##  5 joy              7
##  6 anger            6
##  7 fear             5
##  8 surprise         5
##  9 disgust          4
## 10 sadness          4
## [1] "The Untethered Soul: The Journey Beyond Yourself"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       353
##  2 negative       251
##  3 fear           183
##  4 anticipation   172
##  5 trust          158
##  6 joy            156
##  7 sadness        137
##  8 anger          125
##  9 surprise        65
## 10 disgust         56
## [1] "Man's Search For Meaning: The classic tribute to hope from the
Holocaust"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       199
##  2 negative       172
##  3 fear           108
##  4 sadness        100
##  5 trust           97
##  6 anticipation    93
##  7 joy             77
##  8 anger           62
##  9 disgust         56
## 10 surprise        33
## [1] "The Power of your Subconscious Mind and Other Works"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       183
##  2 joy            110
##  3 trust          110
##  4 anticipation    74
##  5 negative        59
##  6 anger           43
##  7 fear            38
##  8 sadness         29
##  9 surprise        26
```

```
## 10 disgust         22
## [1] "Ego is the Enemy: The Fight to Master Our Greatest Opponent"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       206
##  2 trust          109
##  3 negative        97
##  4 anticipation    85
##  5 joy             79
##  6 fear            59
##  7 anger           54
##  8 sadness         42
##  9 disgust         37
## 10 surprise        31
## [1] "Outliers: The Story of Success"
## # A tibble: 7 x 2
##   sentiment        n
##
## 1 positive        24
## 2 trust           11
## 3 joy              5
## 4 anticipation     4
## 5 fear             3
## 6 surprise         3
## 7 sadness          1
## [1] "The Start-up of You: Adapt to the Future, Invest in Yourself, and
Transform Your Career"
## # A tibble: 10 x 2
##    sentiment        n
##
##  1 positive       145
##  2 anticipation    79
##  3 trust           64
##  4 joy             42
##  5 negative        40
##  6 surprise        22
##  7 fear            21
##  8 sadness         17
##  9 anger           14
## 10 disgust          8
```

```
for (i in 1:28){
sentiment[[i]] %>%
    filter(sentiment %in% c('positive','negative')) %>%
    mutate( n2 = n/sum(n)) %>% print()
  }
```

```
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    450 0.639
## 2 negative    254 0.361
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     53 0.779
## 2 negative     15 0.221
```

```
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    172 0.637
## 2 negative     98 0.363
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    413 0.677
## 2 negative    197 0.323
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     77 0.583
## 2 negative     55 0.417
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    835 0.701
## 2 negative    356 0.299
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    860 0.703
## 2 negative    364 0.297
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    758 0.604
## 2 negative    496 0.396
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     84 0.730
## 2 negative     31 0.270
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    499 0.749
## 2 negative    167 0.251
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    483 0.655
## 2 negative    254 0.345
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    294 0.735
## 2 negative    106 0.265
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    313 0.708
## 2 negative    129 0.292
## # A tibble: 2 x 3
##   sentiment     n    n2
```

```
##
## 1 positive    176 0.830
## 2 negative     36 0.170
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    444 0.712
## 2 negative    180 0.288
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    259 0.792
## 2 negative     68 0.208
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    317 0.703
## 2 negative    134 0.297
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    131 0.672
## 2 negative     64 0.328
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    406 0.618
## 2 negative    251 0.382
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     18 0.581
## 2 negative     13 0.419
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    139 0.813
## 2 negative     32 0.187
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     33  0.75
## 2 negative     11  0.25
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    353 0.584
## 2 negative    251 0.416
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    199 0.536
## 2 negative    172 0.464
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    183 0.756
```

```
## 2 negative     59 0.244
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    206 0.680
## 2 negative     97 0.320
## # A tibble: 1 x 3
##   sentiment     n    n2
##
## 1 positive     24     1
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    145 0.784
## 2 negative     40 0.216
```

```
books <- str_trunc(book_names, width=22)
all <- list()
for (i in 1:28) {
all[[i]] <- sentiment[[i]] %>% filter(sentiment %in% c('positive','negative'))
%>% mutate(n2 = n/sum(n)) %>% print()
}
```

```
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    450 0.639
## 2 negative    254 0.361
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     53 0.779
## 2 negative     15 0.221
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    172 0.637
## 2 negative     98 0.363
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    413 0.677
## 2 negative    197 0.323
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     77 0.583
## 2 negative     55 0.417
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    835 0.701
## 2 negative    356 0.299
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    860 0.703
## 2 negative    364 0.297
```

```
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    758 0.604
## 2 negative    496 0.396
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     84 0.730
## 2 negative     31 0.270
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    499 0.749
## 2 negative    167 0.251
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    483 0.655
## 2 negative    254 0.345
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    294 0.735
## 2 negative    106 0.265
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    313 0.708
## 2 negative    129 0.292
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    176 0.830
## 2 negative     36 0.170
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    444 0.712
## 2 negative    180 0.288
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    259 0.792
## 2 negative     68 0.208
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    317 0.703
## 2 negative    134 0.297
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    131 0.672
## 2 negative     64 0.328
## # A tibble: 2 x 3
##   sentiment     n    n2
```

```
##
## 1 positive    406 0.618
## 2 negative    251 0.382
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     18 0.581
## 2 negative     13 0.419
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    139 0.813
## 2 negative     32 0.187
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive     33  0.75
## 2 negative     11  0.25
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    353 0.584
## 2 negative    251 0.416
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    199 0.536
## 2 negative    172 0.464
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    183 0.756
## 2 negative     59 0.244
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    206 0.680
## 2 negative     97 0.320
## # A tibble: 1 x 3
##   sentiment     n    n2
##
## 1 positive     24     1
## # A tibble: 2 x 3
##   sentiment     n    n2
##
## 1 positive    145 0.784
## 2 negative     40 0.216
```

Positivty Map of the books.

```
all_bound <- do.call("rbind", all) %>% filter(sentiment == "positive")

library(ggrepel)
all_bound %>% ggplot(aes(x= book_names, y=n2)) +
  geom_point() +
  geom_label_repel(aes(label=books, color = ifelse(n2 <0.55, "red", "blue")),
size = 3) +
  theme_classic() +
```

```
  theme(legend.position = "none",
        text = element_text(size=18),
        axis.text.x = element_blank()) +
  xlab("Books") +
  ylab("Positivity score")
```

The lowest positivity score was found in the book **"Man's search for meaning".** This is also kind of expected. Since the book is based on Victor Frankl sufferings during the second world war.

> I am getting more and more convinced text mining is giving good insights.

The book "The Outliers" appeared on the top of the positivity plot was a real outlier here. 😮

No panic.

Let's look at the word count in our Outlier.

```
book_names[[27]]
```

```
## [1] "Outliers: The Story of Success"
```

```
top[[27]]
```

```
## # A tibble: 105 x 2
## # Groups:   word [105]
##    word             n
##
##  1 ability          3
##  2 knowing          3
##  3 sense            3
##  4 communicate      2
##  5 distance         2
##  6 family           2
##  7 intelligence     2
##  8 power            2
##  9 practical        2
## 10 sternberg        2
## # ... with 95 more rows
```

The word count from the book "The Outliers" below is 107. This is really low. So in the next iteration, I would remove it from the analysis since it will not be very informative. It is hard to know everything from the beginning and we will go back and make some additional cleaning.

…

Summary

It is not feasible to read millions of pages to check whether text mining is reliable. But here I got some data that I know the content and I applied text mining approaches and sentiment analysis.
Both the monograms or bigrams pointed to similar ideas what the books were about. And the sentiments made sense with the genres of the books in my kindle.

Let’s come back to our hacker.

He was affected by an unanticipated side effect of the text analyses. As he continued the project, the insights from the frequent ngrams made him more and more interested in the content. He started reading again and the more he read the more world look differently.

He was transformed into a better version of himself.

The world was brighter. ☀

The radio disrupted the silence.

> “brrring…..brrring…..brrring…..”