Adam Smith - A Comparative Text Analytics Approach

Artem Urlapov

2023-03-07

Being an Economist, and particularly one interested in History of Economic Thought, I am naturally drawn to examine some of the cannonical works in my field.

As such, even though Adam Smith drew much of the inspiration from the Physiocrats (these were a group of French economists in the 18th century known for their 'produit net' doctrine i.e., who believed that agriculture was the source of wealth for nations, and also advocated for laissez-faire economics, free trade, and protection of private property rights), he has nevertheless long been regarded as the "father of modern economics", and his contribution, made both in "The Theory of Moral Sentiments" (1759) and in "An Inquiry into the Nature and Causes of the Wealth of Nations" (first published in 1776, and modified on several occasions throughout the remains of his life), remains as relevant today as ever.

Although Adam Smith is also the author of some other less known treatises (particularly in the field of law, which he practised -being himself a barrister-), "An Inquiry into the Nature and Causes of the Wealth of Nations" and "The Theory of Moral Sentiments" are two of his most famous works, and there are both similarities and differences between them.

One of the main similarities between the two works is that they both deal with human behaviour and motivation.

At the same time, there are also differences between the two works. One of the main distinctions is their subject matter: "The Theory of Moral Sentiments" focuses on moral philosophy and the nature of human ethics, while "An Inquiry into the Nature and Causes of the Wealth of Nations" focuses on economics and political economy. As Smith wrote in the former, "How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it." Whereas in the latter, Smith argues that "It is not the benevolence of the butcher, the brewer, or the baker that we expect our dinner from, but from their regard to their own interest."

In terms of style, "The Theory of Moral Sentiments" is more focused on philosophy and contemplation, while "An Inquiry into the Nature and Causes of the Wealth of Nations" is more empirical and fact-based. As Smith wrote in the former, "We can never survey our own sentiments and motives, we can never form any judgment concerning them; unless we remove ourselves, as it were, from our own natural station, and endeavour to view them as at a certain distance from us." Whereas in the latter, Smith argues that "The annual labour of every nation is the fund which originally supplies it with all the necessaries and conveniences of life which it annually consumes, and which consist always either in the immediate produce of that labour, or in what is purchased with that produce from other nations."

Overall, while "An Inquiry into the Nature and Causes of the Wealth of Nations" and "The Theory of Moral Sentiments" share some similarities in their focus on human behaviour and motivation, there are also notable differences in their subject matter, style, and approach.

Through the tools learnt in the course of 'Text Mining' (Master in Computation Social Science at Universidad Carlos III de Madrid), I will attempt to show some of the aforementioned

similarities and differences found in these two works of Adam Smith, as well as other relevant points that can be drawn from a thorough analysis of these 2 'oeuvres'.

Required libraries:

```
library(gutenbergr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidytext)
library(tidyverse)
library(stringr)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

Fist, I will proceed to retrieve the corpus:

```
TheWealthOfNations <- gutenberg_download(c(3300), meta_fields = "title")
```

```
## Determining mirror for Project Gutenberg from https://www.gutenberg.org/robot/harvest

## Using mirror http://aleph.gutenberg.org
```

```
TheTheoryOfMoralSentiments <- gutenberg_download(c(67363), meta_fields = "title")
```

Next, I will tokenize them.

Let us have a look at "The Theory of Moral Sentiments" first. I will save it into another variable so that the interested reader can see it in full.

```
TTOMSBook <- TheTheoryOfMoralSentiments %>%
  unnest_tokens(word, text, drop = FALSE, strip_punct = FALSE, to_lower = FALSE)

head(TTOMSBook)
```

```
## # A tibble: 6 x 4
##   gutenberg_id text                                    title        word
##          <int> <chr>                                   <chr>        <chr>
## 1        67363 "                                THE"  "The Theory~ THE
## 2        67363 "                               THEORY" "The Theory~ THEO~
## 3        67363 "                                 OF"  "The Theory~ OF
## 4        67363 "                      MORAL SENTIMENTS;" "The Theory~ MORAL
## 5        67363 "                      MORAL SENTIMENTS;" "The Theory~ SENT~
## 6        67363 "                      MORAL SENTIMENTS;" "The Theory~ ;
```

Now, I will perform the same action concerning "An Inquiry into the Nature and Causes of the Wealth of Nations", an 'oeuvre' which, for simplicity purposes, I will call simply "The Wealth of Nations" from now on.

```
TWONBook <- TheWealthOfNations %>%
  unnest_tokens(word, text, drop = FALSE, strip_punct = FALSE, to_lower = FALSE)

head(TWONBook)
```

```
## # A tibble: 6 x 4
##   gutenberg_id text                                                  title word
##          <int> <chr>                                                 <chr> <chr>
## 1         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ An
## 2         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ Inqu~
## 3         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ into
## 4         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ the
## 5         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ Natu~
## 6         3300 An Inquiry into the Nature and Causes of the Wealth ~ An I~ and
```

Since to have the variable 'gutenberg_id' only makes things messier, while not adding additional information to us, I will drop it.

```
TTOMSBook <- TTOMSBook %>%
  select(-c(gutenberg_id))

TWONBook <- TWONBook %>%
  select(-c(gutenberg_id))
```

Now, we have only 3 variables in each case: "text", "title", and "word".

Needless to say, both "The Theory of Moral Sentiments" and "The Wealth of Nations" have their specific internal structures, through which we will navigate.

To be more precise, **"The Theory of Moral Sentiments" is divided into 5 main parts**, which contain multiple sections and chapters:

Part I: Of the Propriety of Action

Part II: Of Merit and Demerit; or, of the Objects of Reward and Punishment

Part III: Of the Foundation of our Judgments Concerning our Own Sentiments and Conduct, and of the Sense of Duty

Part IV: Of the Effect of Utility Upon the Sentiment of Approbation in the Individual

Part V: Of the Influence of Custom and Fashion Upon Moral Sentiments

As to what to "The Wealth of Nations" it pertains, it is divided into 4 main "books" (each of them containing multiple sections and chapters):

Book I: Of the Causes of Improvement in the Productive Powers of Labour

Chapter I:Introduction

Chapter II: Of the Principle which gives occasion to the Division of Labour

Chapter III:That the Division of Labour is limited by the extent of the Market

Chapter IV: Of the Origin and Use of Money

```
Original_Sentiments_Book <- TTOMSBook %>%
  mutate(part = cumsum(
    str_detect(text, regex("^\\s*PART\\s+[IVXLCDM]+\\s*\\.", ignore_case = TRUE)) &
    !str_detect(lag(text), regex("^\\s*PART\\s+[IVXLCDM]+\\s*\\.", ignore_case = TRUE))
  )) %>%
  filter(part <= 5) %>%
  group_by(part) %>%
  mutate(linenumber = row_number()) %>%
  ungroup()

#Now we can check that there are indeed 5 parts in the book (plus the text before the 1st Part).

unique(Original_Sentiments_Book$part)
```

```
## [1] 0 1 2 3 4 5
```

Consequently, we have "The Theory of Moral Sentiments" nicely divided into its corresponding 5 parts.

Let's proceed to perform the same process with the 4 books of "The Wealth of Nations".

```
Original_Nations_Book <- TWONBook %>%
  mutate(chapter = cumsum(
    str_detect(text, regex("^\\s*CHAPTER\\s+[IVXLCDM]+\\s*\\.", ignore_case = TRUE)) &
    !str_detect(lag(text), regex("^\\s*CHAPTER\\s+[IVXLCDM]+\\s*\\.", ignore_case = TRUE))
  )) %>%
  filter(chapter <= 43) %>%
  group_by(chapter) %>%
  mutate(linenumber = row_number()) %>%
  ungroup()

#Now we can check that there are 43 chapters in the book.

unique(Original_Nations_Book$chapter)
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
## [26] 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
```

Equally, "The Wealth of Nations" is also nicely divided into its corresponding 43 chapters.

Next, I will proceed to tokenise "The Theory of Moral Sentiments" and "The Wealth of Nations".

```
Tidy_Original_Sentiments_Book <- Original_Sentiments_Book %>%
  unnest_tokens(word, text)

head(Tidy_Original_Sentiments_Book)
```

```
## # A tibble: 6 x 4
##   title                                                word   part linenumber
##   <chr>                                                <chr> <int>      <int>
## 1 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ the       0          1
## 2 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ theo~     0          2
## 3 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ of        0          3
```

```
## 4 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ moral     0          4
## 5 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ sent~     0          4
## 6 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ moral     0          5
```

```
Tidy_Original_Nations_Book <- Original_Nations_Book %>%
  unnest_tokens(word, text)

head(Tidy_Original_Nations_Book)
```

```
## # A tibble: 6 x 4
##   title                                                word  chapter linenumber
##   <chr>                                                <chr>   <int>      <int>
## 1 An Inquiry into the Nature and Causes of the Wealth ~ an          0          1
## 2 An Inquiry into the Nature and Causes of the Wealth ~ inqu~       0          1
## 3 An Inquiry into the Nature and Causes of the Wealth ~ into        0          1
## 4 An Inquiry into the Nature and Causes of the Wealth ~ the         0          1
## 5 An Inquiry into the Nature and Causes of the Wealth ~ natu~       0          1
## 6 An Inquiry into the Nature and Causes of the Wealth ~ and         0          1
```

As we remember from the definition, "A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing." Consequently, not all words are useful for analysis. Let's remove them!

```
Tidy_Original_Sentiments_Book <- Tidy_Original_Sentiments_Book %>%
  filter(!word %in% c("chap", "_of", "ii", "iii", "iv", "_that", "_in", "ibid")) %>%
  anti_join(stop_words, by = "word")

head(Tidy_Original_Sentiments_Book)
```

```
## # A tibble: 6 x 4
##   title                                                  word   part linenumber
##   <chr>                                                  <chr> <int>      <int>
## 1 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ theo~     0          2
## 2 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ moral     0          4
## 3 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ sent~     0          4
## 4 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ moral     0          5
## 5 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ sent~     0          5
## 6 "The Theory of Moral Sentiments\r\nOr, an Essay Toward~ moral     0          6
```

```
Tidy_Original_Nations_Book <- Tidy_Original_Nations_Book %>%
  anti_join(stop_words)
```

```
## Joining with `by = join_by(word)`
```

```
head(Tidy_Original_Nations_Book)
```

```
## # A tibble: 6 x 4
##   title                                                word  chapter linenumber
##   <chr>                                                <chr>   <int>      <int>
## 1 An Inquiry into the Nature and Causes of the Wealth ~ inqu~       0          1
## 2 An Inquiry into the Nature and Causes of the Wealth ~ natu~       0          1
```

```
## 3 An Inquiry into the Nature and Causes of the Wealth ~ weal~       0          1
## 4 An Inquiry into the Nature and Causes of the Wealth ~ nati~       0          1
## 5 An Inquiry into the Nature and Causes of the Wealth ~ inqu~       0          2
## 6 An Inquiry into the Nature and Causes of the Wealth ~ natu~       0          2
```

Another interesting operation that we can perform is to count how many times each word is repeated in "The Theory of Moral Sentiments" and "The Wealth of Nations", respectively.

```
Tidy_Original_Sentiments_Book %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 210 x 2
##    word           n
##    <chr>      <int>
##  1 influence    102
##  2 sense         69
##  3 origin        60
##  4 utility       50
##  5 conduct       49
##  6 sentiments    49
##  7 sympathy      48
##  8 merit         45
##  9 appears       43
## 10 propriety     43
## # i 200 more rows
```

```
Tidy_Original_Nations_Book %>%
  count(word, sort = TRUE)
```

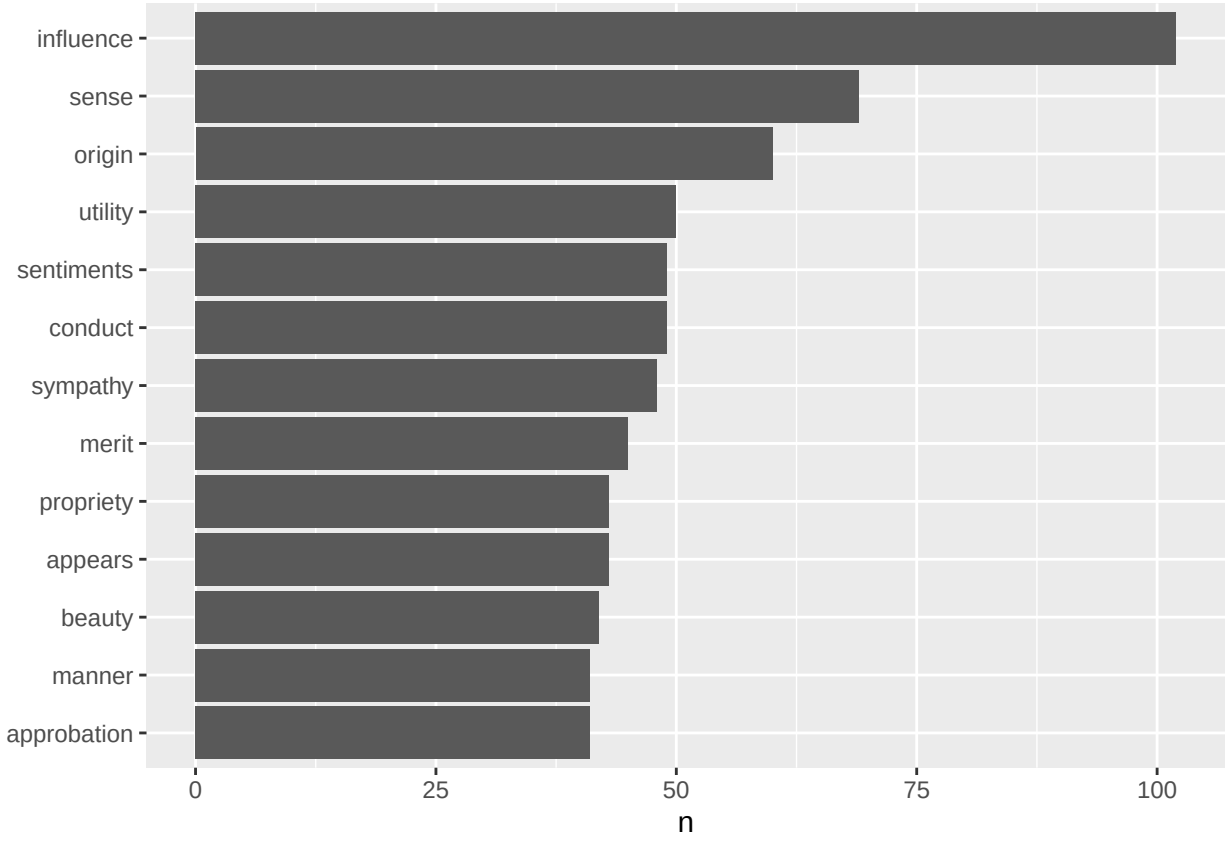
```
## # A tibble: 6,935 x 2
##    word         n
##    <chr>    <int>
##  1 price    15296
##  2 labour   11253
##  3 country  11156
##  4 quantity  9057
##  5 money     9032
##  6 silver    8568
##  7 produce   8011
##  8 trade     7134
##  9 gold      6578
## 10 corn      6228
## # i 6,925 more rows
```

Two conclusions can be drawn: as been mentioned in the beginning, "The Theory of Moral Sentiments" focuses much more on moral philosophy and the nature of human ethics; hence the frequent appearance of words such as "influence", sense", "origin", "utility", "conduct", "sentiments", and so on.

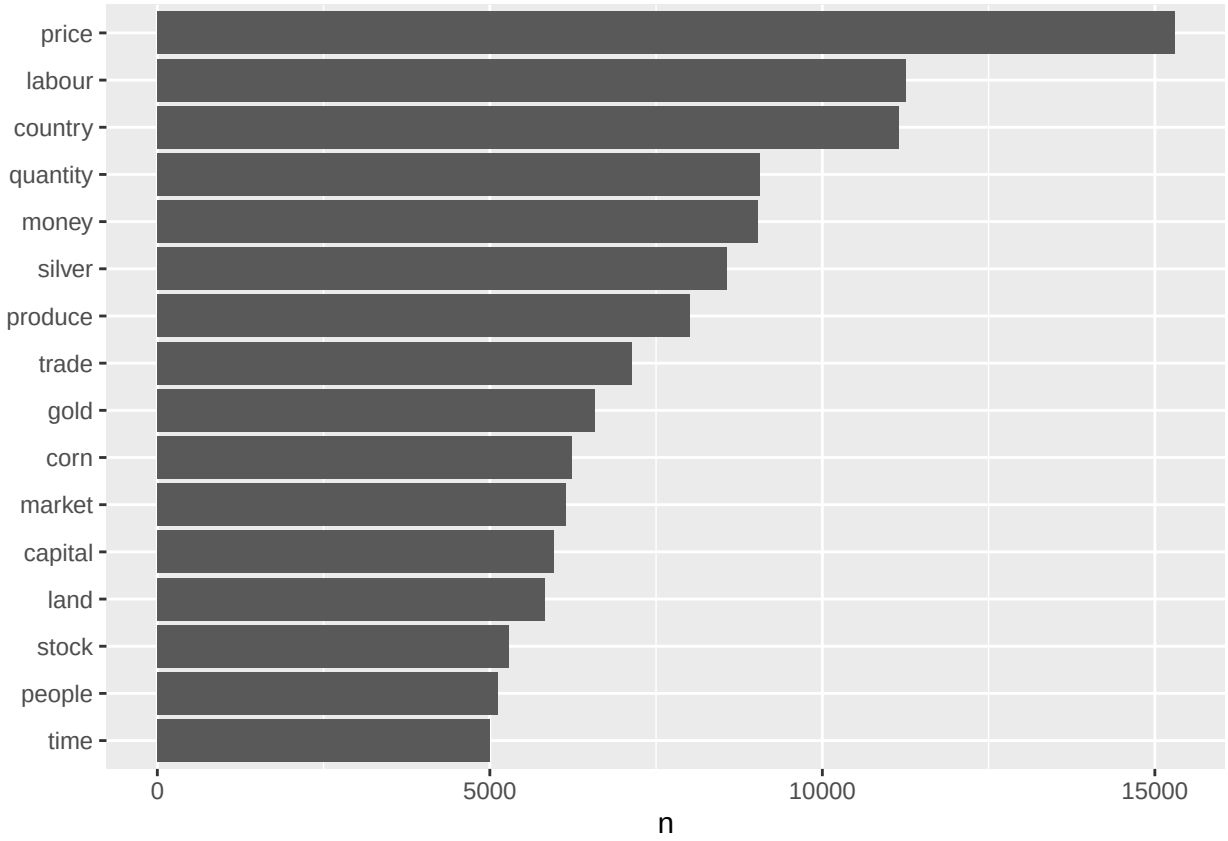
At the same time, it is a much shorter book than "The Wealth of Nations", and because of this "n" is smaller. Regarding "The Wealth of Nations", its "mercantile" character can be easily seen: words such as "price", "labour", "country", "quantity", "money" and / or "silver" are highly prominent in it.

To make things even easier, let us visualise what has been mentioned previously!

```
Tidy_Original_Sentiments_Book %>%
  count(word, sort = TRUE) %>%
  filter(n > 40) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL)
```




```
Tidy_Original_Nations_Book %>%
  count(word, sort = TRUE) %>%
  filter(n > 5000) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL)
```



It seems clear enough, by now, that an actual thorough human reading is required in order to find similarities between these two works by Adam Smith, since -at first glance; and having read myself both books I can definitely assure the interested reader that these similarities exist between them- what is shown in this analysis is that two different -if not opposite- directions are followed in these two 'oeuvres'. But again, this why A.I. is by no means a substitute for human intelligence!

Let us proceed now with some sentiment analysis!

As is relatively well-known, in the 1980s, the American psychologist Robert Plutchik created a Wheel of Emotions, which shows 8 basic emotions: "joy", "trust", "fear", "surprise", "sadness", "anticipation", "anger", and "disgust".

It would be interesting to see the prominance of, for instance, "sadness" and "disgust" (negative emotions, yes, but why not? Industrial Revolution was not an entirely idyllic process, after all.) in the two selected works of Adam Smith.

```
NRC_Sadness <- get_sentiments("nrc") %>%
  filter(sentiment == "sadness")


Sentiments_Sadness <- Tidy_Original_Sentiments_Book %>%
    inner_join(NRC_Sadness) %>%
    count(word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
head(Sentiments_Sadness)
```

```
## # A tibble: 6 x 2
##   word           n
##   <chr>      <int>
## 1 sympathy      48
## 2 resentment    21
## 3 remorse       16
## 4 sorrow        13
## 5 art           12
## 6 violence      12
```

```
Nations_Sadness <- Tidy_Original_Nations_Book %>%
  inner_join(NRC_Sadness) %>%
  count(word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
head(Nations_Sadness)
```

```
## # A tibble: 6 x 2
##   word         n
##   <chr>    <int>
## 1 scarcity  1313
## 2 weight    1270
## 3 scarce    1257
## 4 tax       1198
## 5 lower     1192
## 6 fall      1142
```

Thus, concerning the feeling of "sadness", we can clearly see the appearance of words such as "sympathy", "resentment", "remorse", "sorrow", and others in the case of "The Theory of Moral Sentiments". On the other hand, in the case of "The Wealth of Nations", "sadness" is associated mostly with the words "scarcity", "weight", "tax", "debt" i.e., something seen as negative for economic prosperity of a mercantile point of view (it is clear that "debt" or "tax" have different associations depending on each concrete school of economic thought).

Nevertheless, what is interesting, is that there is a slight overlapping here: for instance, the word "scarse" appears both in "The Theory of Moral Sentiments" and in "The Wealth of Nations". There are other similarities: words such as "loss", "sorrow" and / or "pain" can be considered as synonyms, and they are present in both works.

Let us see now what happens with the emotion of "disgust"!

```
NRC_Disgust <- get_sentiments("nrc") %>%
  filter(sentiment == "disgust")


Sentiments_Disgust <- Tidy_Original_Sentiments_Book %>%
    inner_join(NRC_Disgust) %>%
    count(word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
head(Sentiments_Disgust)
```

```
## # A tibble: 3 x 2
##   word           n
##   <chr>      <int>
## 1 resentment    21
## 2 punishment    14
## 3 selfish        9
```

```
Nations_Disgust <- Tidy_Original_Nations_Book %>%
  inner_join(NRC_Disgust) %>%
  count(word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
head(Nations_Disgust)
```

```
## # A tibble: 6 x 2
##   word           n
##   <chr>      <int>
## 1 weight      1270
## 2 owing        477
## 3 remains      438
## 4 poverty      413
## 5 prohibited   407
## 6 abundance    401
```

Here, we can observe some very clear moral philosophy and ethics overtones in "The Theory of Moral Sentiments": words most associated with disgust are: "resentment", "punishment", and "selfish".

In what to "The Wealth of Nations" it refers, with the exception of a few words like "prohibited", "confined" and / or "poverty", most other words seem to be context-dependent and age-dependent: for instance, it would be hard to conceive "weight" to be associated with the emotion of "disgust" nowadays. And yet, it is the most frequent word associated with disgust in "The Wealth of Nations". Again, "bad" seems to have a much stronger negative connotation in the XVIII Century than it does today.

Because "NRC" is not the only tool for "sentiment analysis" that we have at our disposal, I will proceed to continue the analysis by making use of "Bing et al.".

```
Sentiments_Bing <- Tidy_Original_Sentiments_Book %>%
  inner_join(get_sentiments("bing")) %>%
  count(part, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining with `by = join_by(word)`
```

```
head(Sentiments_Bing)
```

```
## # A tibble: 6 x 5
##    part index negative positive sentiment
##   <int> <dbl>    <int>    <int>     <int>
## 1     1     0       17       11        -6
## 2     1     1        0       32        32
## 3     1     2       36       69        33
## 4     1     3        9       15         6
## 5     2     0       37      103        66
## 6     2     1       42       35        -7
```

```
library(tidyverse)

Nations_Bing <- Tidy_Original_Nations_Book %>%
  inner_join(get_sentiments("bing")) %>%
  count(chapter, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining with `by = join_by(word)`
```

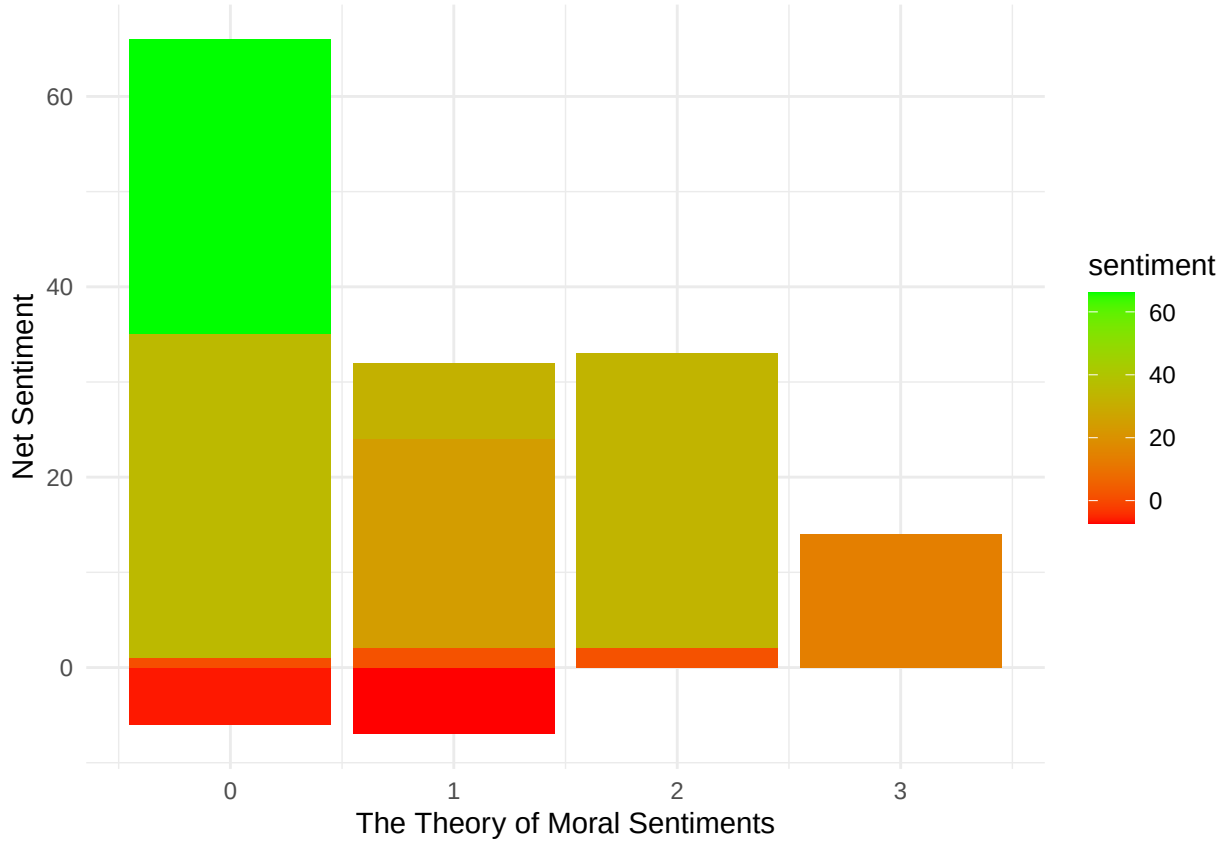
```
head(Nations_Bing)
```

```
## # A tibble: 6 x 5
##   chapter index positive negative sentiment
##     <int> <dbl>    <int>    <int>     <int>
## 1       0     0       22        0        22
## 2       2     0        0       11       -11
## 3       7     0        5        5         0
## 4       8     0       10        0        10
## 5       9     0       10       23       -13
## 6      10     0       11       10         1
```

In order to see it all more clearly, a graphical representation would be of much help!

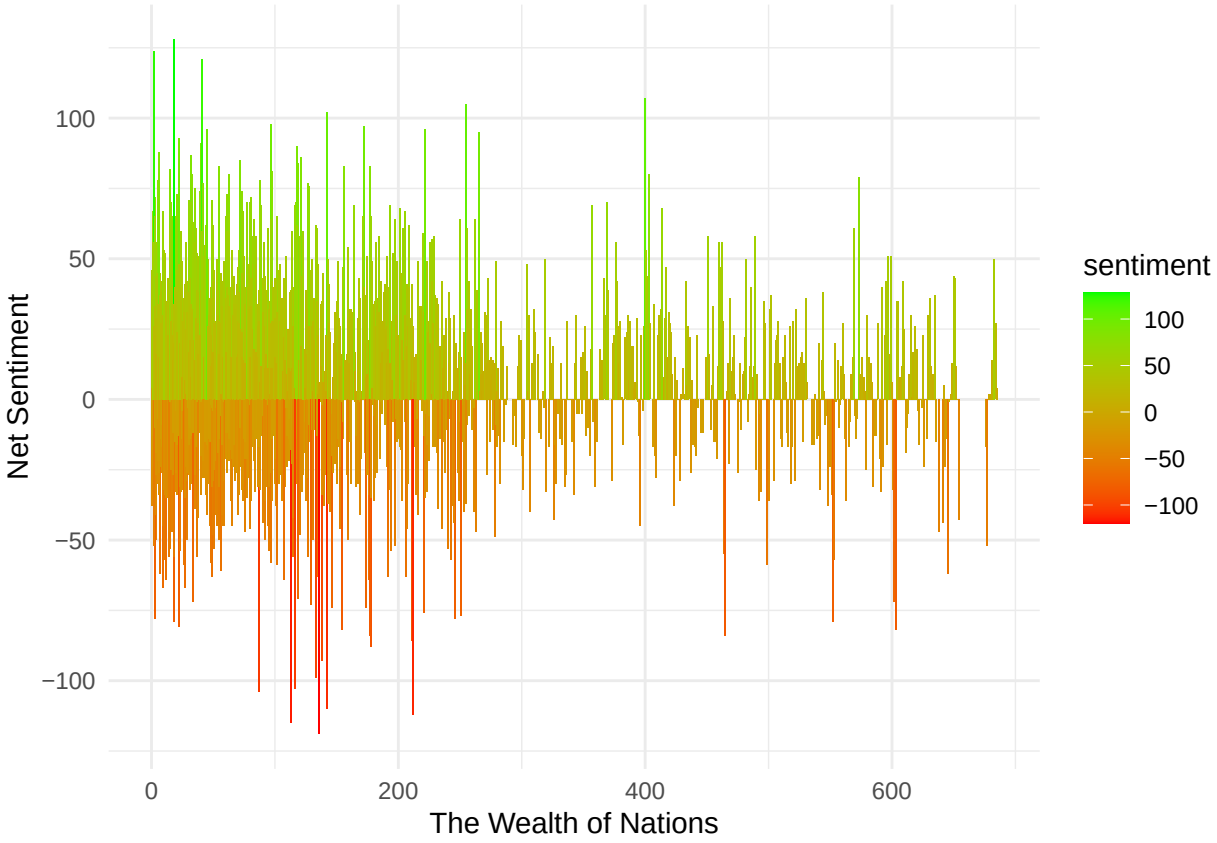
```
library(ggplot2)

ggplot(Sentiments_Bing, aes(x = index, y = sentiment, fill = sentiment)) +
  geom_col(position = "dodge") +
  scale_fill_gradient(low = "red", high = "green") +
  labs(x = "The Theory of Moral Sentiments", y = "Net Sentiment") +
  theme_minimal()
```



Overall, we can observe a huge predominance of the "greens", which stand for positive moral sentiments.

```
ggplot(Nations_Bing, aes(x = index, y = sentiment, fill = sentiment)) +
  geom_col(position = "dodge") +
  scale_fill_gradient(low = "red", high = "green") +
  labs(x = "The Wealth of Nations", y = "Net Sentiment") +
  theme_minimal()
```



In what to "The Wealth of Nations" it refers, there is a mix of both positive and negative sentiments, though we can say that the positive ones are slightly more prevalent. At the same time, it is worth noticing that this 'oeuvre' follows a rather "explosive" pattern: sentiments are very strong in the beginning, and they decay in intensity afterwards. This is because Adam Smith focuses on more technical -hence, language-neutral- details in the latter parts of "The Wealth of Nations".

Concerning the style of both 'oeuvres', it can be observed that even though these are two academic works, the language used in both of them is a highly expressive one; which stands out in comparison to the rather dull -neutral- language employed by economists nowadays.

Even though what has been done so far can be regarded as a good aproximation, there must surely be a better metric in order to calculate how many positive and negative sentiments there are in each 'oeuvre'.

Let's do this first with "The Theory of Moral Sentiments"!

```
#AFINN

ASSentiments_afinn <- Tidy_Original_Sentiments_Book %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(index = linenumber %/% 80) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(method = "AFINN")
```

```
## Joining with `by = join_by(word)`
```

```
#Bing and NRC
ASSentiments_bing_and_nrc <- bind_rows(
```

```
  #Bing
  Tidy_Original_Sentiments_Book %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  #NRC
  Tidy_Original_Sentiments_Book %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))
    ) %>%
    mutate(method = "NRC")) %>%
  count(method, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment,
              values_from = n,
              values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
## Warning in inner_join(., get_sentiments("nrc") %>% filter(sentiment %in% : Detected an unexpected ma
## i Row 2062 of `x` matches multiple rows in `y`.
## i Row 3373 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

Let us have a look at the results:

```
head(ASSentiments_afinn)
```

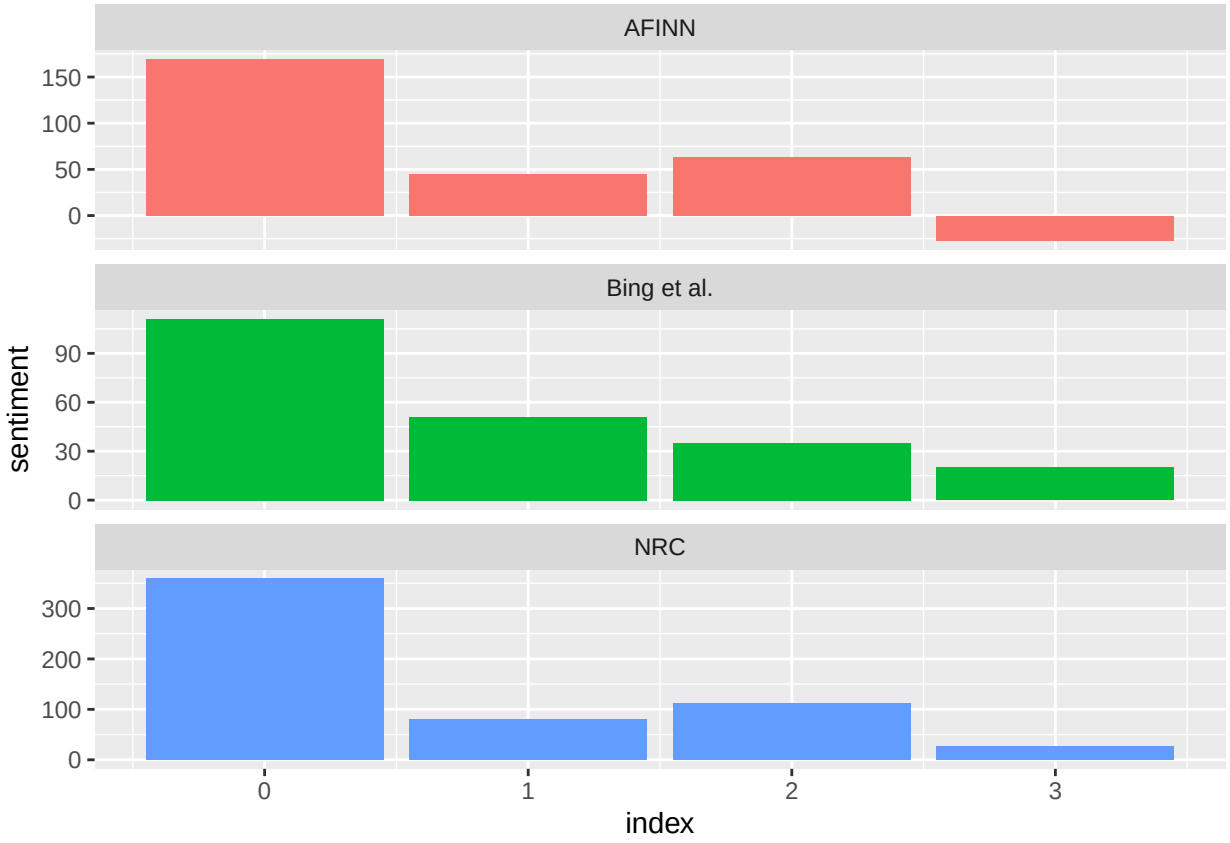
```
## # A tibble: 4 x 3
##   index sentiment method
##   <dbl>     <dbl> <chr>
## 1     0       169 AFINN
## 2     1        45 AFINN
## 3     2        63 AFINN
## 4     3       -27 AFINN
```

```
head(ASSentiments_bing_and_nrc)
```

```
## # A tibble: 6 x 5
##   method      index negative positive sentiment
##   <chr>       <dbl>    <int>    <int>     <int>
## 1 Bing et al.     0       58      169       111
## 2 Bing et al.     1       42       93        51
## 3 Bing et al.     2       52       87        35
## 4 Bing et al.     3       21       41        20
## 5 NRC             0      115      474       359
## 6 NRC             1       52      132        80
```

For further clarity, let us plot the results in "ggplot2":

```
bind_rows(ASSentiments_afinn,
          ASSentiments_bing_and_nrc) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```



As we can see in the graph above, there are some substantial differences among the 3 methods.

This further proceedings clearly relate to the following point made by Prof. Carmen Torrijos Caruda:

"We can not trust the lexicon 100%. Generally, there are three **possible biases** in sentiment lexicons:

- The lexicon is biased towards negative sentiment (more negative than positive words)
- The lexicon is biased towards positive sentiment (more positive than negative words)
- The lexicon is very different in style or context from the text we are analyzing."

Finally, let us obtain a very nice graphical representation of the most prominent words in "The Theory of Moral Sentiments":

```
Tidy_Original_Sentiments_Book %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 30))
```

```
## Joining with `by = join_by(word)`
```

Now, let us repeat the process with "The Wealth of Nations"!

```
#AFINN

ASNations_afinn <- Tidy_Original_Nations_Book %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(index = linenumber %/% 80) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(method = "AFINN")
```

```
## Joining with `by = join_by(word)`
```

```
#Bing and NRC
ASNations_bing_and_nrc <- bind_rows(
  #Bing
  Tidy_Original_Nations_Book %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  #NRC
  Tidy_Original_Nations_Book %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))
    ) %>%
    mutate(method = "NRC")) %>%
  count(method, index = linenumber %/% 80, sentiment) %>%
```

```
  pivot_wider(names_from = sentiment,
              values_from = n,
              values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
## Warning in inner_join(., get_sentiments("nrc") %>% filter(sentiment %in% : Detected an unexpected ma
## i Row 2841 of `x` matches multiple rows in `y`.
## i Row 2810 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

Let us see what we obtained:

```
head(ASNations_afinn)
```

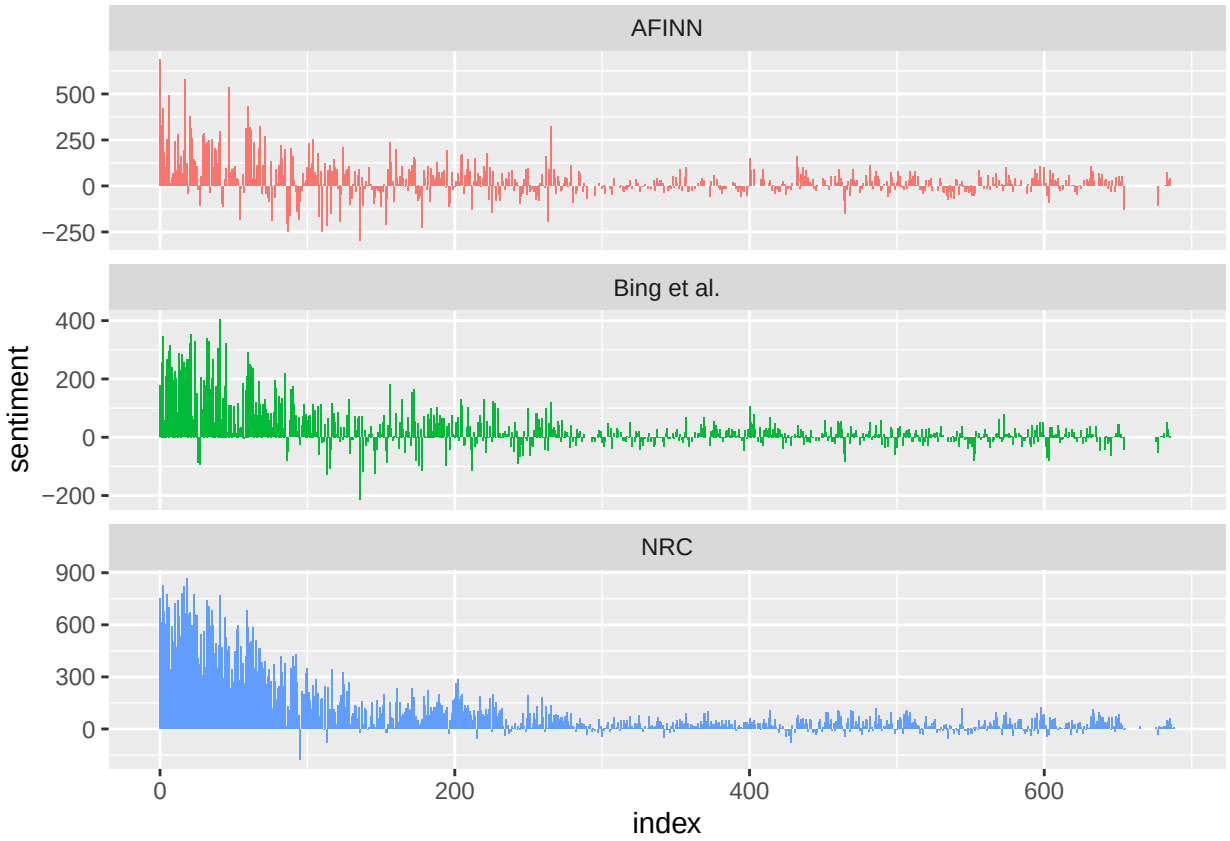
```
## # A tibble: 6 x 3
##   index sentiment method
##   <dbl>     <dbl> <chr>
## 1     0       687 AFINN
## 2     1       330 AFINN
## 3     2       420 AFINN
## 4     3       181 AFINN
## 5     4       101 AFINN
## 6     5       255 AFINN
```

```
head(ASNations_bing_and_nrc)
```

```
## # A tibble: 6 x 5
##   method      index negative positive sentiment
##   <chr>       <dbl>    <int>    <int>     <int>
## 1 Bing et al.     0      370      547       177
## 2 Bing et al.     1      328      587       259
## 3 Bing et al.     2      467      813       346
## 4 Bing et al.     3      349      409        60
## 5 Bing et al.     4      428      638       210
## 6 Bing et al.     5      385      651       266
```

For further clarity, let us plot the results in "ggplot2":

```
bind_rows(ASNations_afinn,
          ASNations_bing_and_nrc) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```



In the case of "The Wealth of Nations", the 3 different metrics are much more similar. With the slight exception of "AFINN", negative sentiments are practically non-existent. As been commented before, there is a very strong narrative / stylistic intensity in the beginning of the work, one that slowly decays as "The Wealth of Nations" proceeds to analyse more technical aspects, hence employing a more academic, neutral language.

Finally, let us plot the most frequent words:

```
Tidy_Original_Nations_Book %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 30))
```

```
## Joining with `by = join_by(word)`
```

```
## Warning in wordcloud(word, n, max.words = 30): price could not be fit on page.
## It will not be plotted.
```

A comparative analysis regarding the most frequent words in both "The Theory of Sentiments" and "The Wealth of Nations" has already been made previously, so I see no need for being repetitive.

So far so good, I think that I have covered all the essential aspects of "working with multiple texts" and sentiment analysis", and have explained the similarities and differences between the two selected works of Adam Smith.

Let us proceed now with "term frequency" analysis!

As usual, we will go with "The Theory of Moral Sentiments" first:

```
Sentiments_Book_Words <- Tidy_Original_Sentiments_Book %>%
#I have already tokenised the book, so I will just count the word frequency here.
  count(part, word, sort = TRUE)

head(Sentiments_Book_Words)
```

```
## # A tibble: 6 x 3
##    part word          n
##   <int> <chr>     <int>
## 1     2 merit        45
## 2     1 origin       43
## 3     1 propriety    43
## 4     2 appears      43
## 5     2 gratitude    40
## 6     5 custom       39
```

Next, I will sum up the total number of words contained in each Part.

```
Sentiments_Total_Words <- Sentiments_Book_Words %>%
  group_by(part) %>%
  summarize(total = sum(n))

Sentiments_Total_Words
```

```
## # A tibble: 6 x 2
##    part total
##   <int> <int>
## 1     0   363
## 2     1   857
## 3     2  1053
## 4     3   389
## 5     4   291
## 6     5   195
```

As we can see, Part I and Part II are by far the lengthiest ones in "The Theory of Moral Sentiments".

To be able to continue the work, it is necessary to save the results in a DataFrame.

```
Sentiments_Words <- left_join(Sentiments_Book_Words, Sentiments_Total_Words)
```

```
## Joining with `by = join_by(part)`
```

```
head(Sentiments_Words)
```

```
## # A tibble: 6 x 4
##    part word          n total
##   <int> <chr>     <int> <int>
## 1     2 merit        45  1053
## 2     1 origin       43   857
## 3     1 propriety    43   857
## 4     2 appears      43  1053
## 5     2 gratitude    40  1053
## 6     5 custom       39   195
```

Perfect! Now we will be able to see the "term frequency" in "The Theory of Moral Sentiments":

```
Sentiments_Words <- Sentiments_Words %>%
  mutate(Sentiments_Term_Frequency = n/total)

head(Sentiments_Words)
```

```
## # A tibble: 6 x 5
##    part word          n total Sentiments_Term_Frequency
##   <int> <chr>     <int> <int>                      <dbl>
## 1     2 merit        45  1053                     0.0427
## 2     1 origin       43   857                     0.0502
## 3     1 propriety    43   857                     0.0502
## 4     2 appears      43  1053                     0.0408
## 5     2 gratitude    40  1053                     0.0380
## 6     5 custom       39   195                     0.2
```

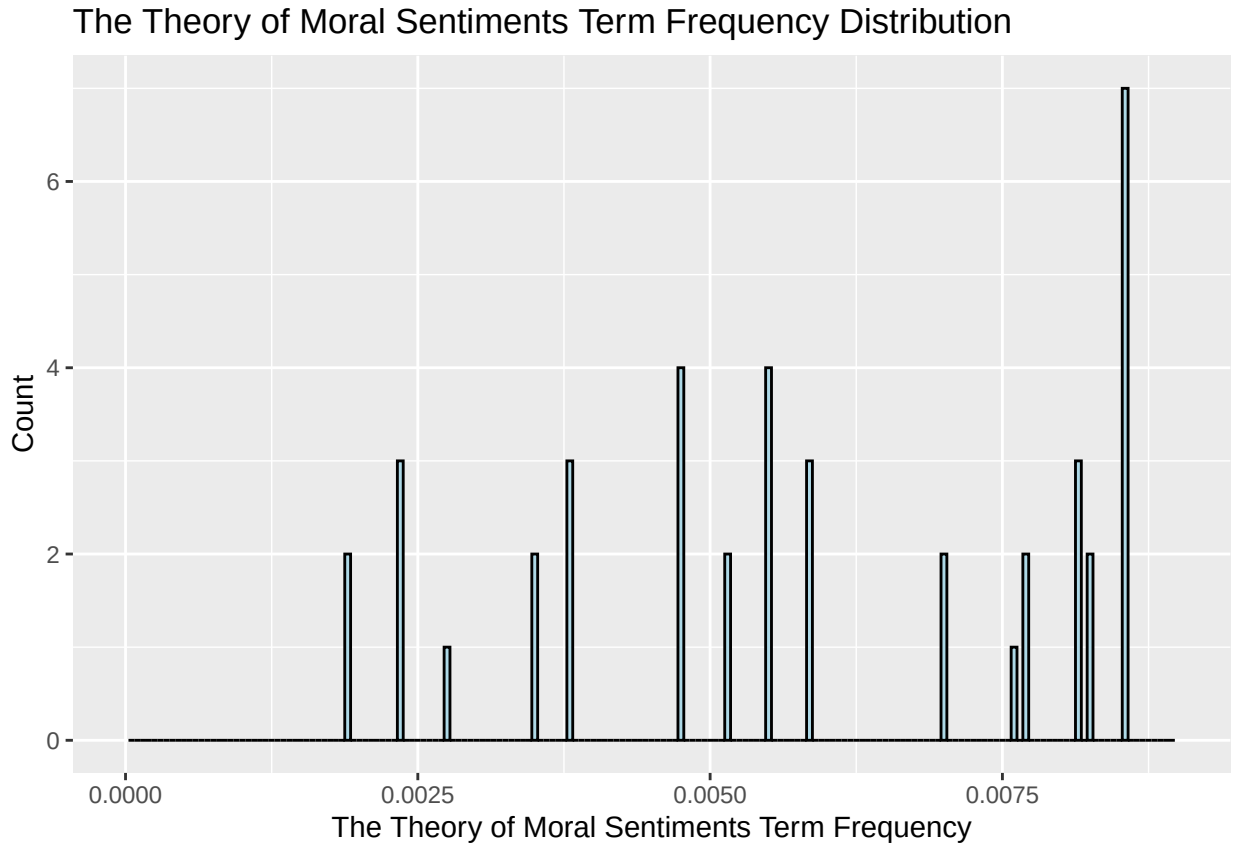
As can be seen, there are many relevant words, such as “merit”, “origin”, “propriety”, “gratitude”, “custom”, and others; that appear in this Adam Smith’s ‘oeuvre’.

Let us make a plot of the word distribution in “The Theory of Moral Sentiments”:

```
ggplot(Sentiments_Words, aes(x = Sentiments_Term_Frequency)) +
  geom_histogram(binwidth = 0.00005, color = "black", fill = "lightblue") +
  scale_x_continuous(limits = c(0, 0.009)) +
  labs(title = "The Theory of Moral Sentiments Term Frequency Distribution",
       x = "The Theory of Moral Sentiments Term Frequency",
       y = "Count")
```

```
## Warning: Removed 206 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



Overall, a fat-tailed distribution can be observed.

As per Prof. Carmen Torrijos Caruda notes, “Zipf’s law was claimed by George Zipf, a 20th century linguist from the United States. It states that the *frequency of a word appearance in a text is inversely proportional to its rank.*

“The lower the frequency, the higher the rank.”

Let us see what happens with “The Theory of Moral Sentiments”!

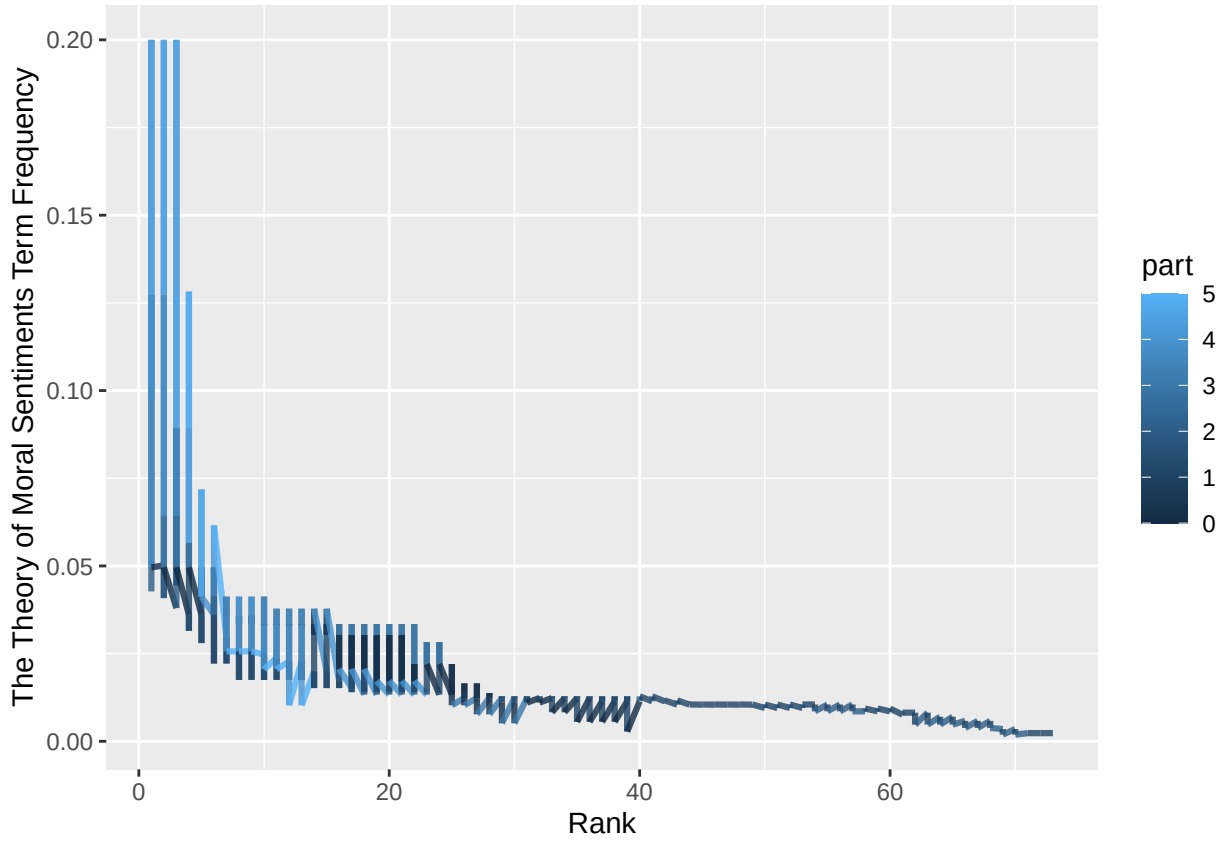
```
Sentiments_Frequency_By_Rank <- Sentiments_Words %>%
  group_by(part) %>%
  mutate(rank = row_number()) %>%
  ungroup()

head(Sentiments_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##    part word          n total Sentiments_Term_Frequency  rank
##   <int> <chr>     <int> <int>                      <dbl> <int>
## 1     2 merit        45  1053                     0.0427     1
## 2     1 origin       43   857                     0.0502     1
## 3     1 propriety    43   857                     0.0502     2
## 4     2 appears      43  1053                     0.0408     2
## 5     2 gratitude    40  1053                     0.0380     3
## 6     5 custom       39   195                     0.2        1
```

Let's visualise the results of Zipf's Law for "The Theory of Moral Sentiments":

```
Sentiments_Frequency_By_Rank %>%
  ggplot(aes(rank, Sentiments_Term_Frequency, colour = part)) +
  geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
  xlab("Rank") +
  ylab("The Theory of Moral Sentiments Term Frequency")
```



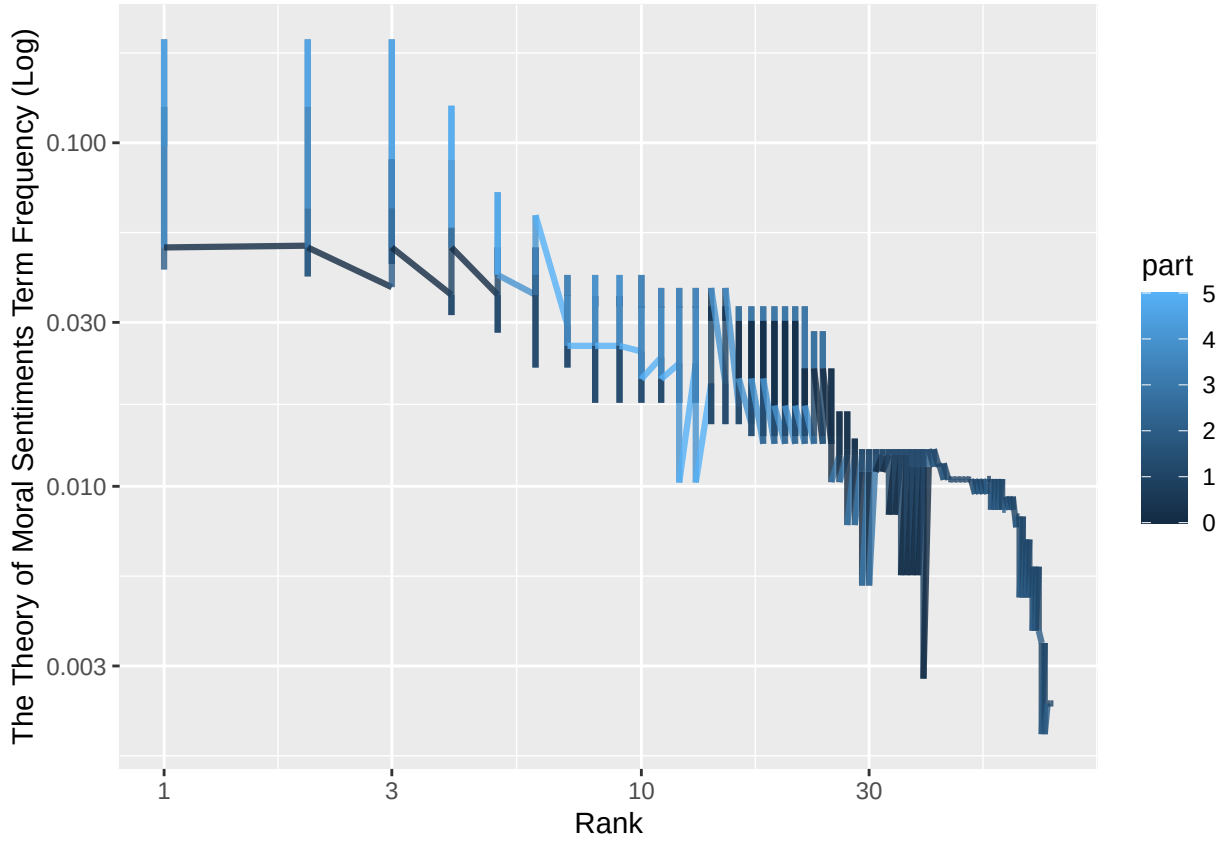
```
head(Sentiments_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##    part word          n total Sentiments_Term_Frequency  rank
##   <int> <chr>     <int> <int>                      <dbl> <int>
## 1     2 merit        45  1053                     0.0427     1
## 2     1 origin       43   857                     0.0502     1
## 3     1 propriety    43   857                     0.0502     2
## 4     2 appears      43  1053                     0.0408     2
## 5     2 gratitude    40  1053                     0.0380     3
## 6     5 custom       39   195                     0.2        1
```

We can indeed observe what we would call in mathematics a monotonically decreasing exponential function. In linguistic terms, what we can take out of this graph is that those words with higher term frequency are lower in the rank.

Let us see how all this plays out on a logarithmic scale!

```
Sentiments_Frequency_By_Rank %>%
  ggplot(aes(rank, Sentiments_Term_Frequency, colour = part)) +
  geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
  scale_x_log10() +
  scale_y_log10() +
  xlab("Rank") +
  ylab("The Theory of Moral Sentiments Term Frequency (Log)")
```



```
head(Sentiments_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##    part word          n total Sentiments_Term_Frequency  rank
##   <int> <chr>     <int> <int>                      <dbl> <int>
## 1     2 merit        45  1053                     0.0427     1
## 2     1 origin       43   857                     0.0502     1
## 3     1 propriety    43   857                     0.0502     2
## 4     2 appears      43  1053                     0.0408     2
## 5     2 gratitude    40  1053                     0.0380     3
## 6     5 custom       39   195                     0.2        1
```

Here, we can see "inversely proportional relationship will have a constant, negative slope from right to left".

Let us proceed with TF*IDF analysis of "The Theory of Moral Sentiments"!

```
Sentiments_TF_IDF <- Sentiments_Words %>%
  filter(!word %in% c("32")) %>%
  bind_tf_idf(word, part, n)

head(Sentiments_TF_IDF)
```

```
## # A tibble: 6 x 8
##    part word          n total Sentiments_Term_Frequency     tf   idf tf_idf
##   <int> <chr>     <int> <int>                      <dbl>  <dbl> <dbl>  <dbl>
## 1     2 merit        45  1053                     0.0427 0.0427 1.79  0.0766
## 2     1 origin       43   857                     0.0502 0.0502 0.693 0.0348
## 3     1 propriety    43   857                     0.0502 0.0502 1.79  0.0899
## 4     2 appears      43  1053                     0.0408 0.0408 1.79  0.0732
## 5     2 gratitude    40  1053                     0.0380 0.0380 1.79  0.0681
## 6     5 custom       39   195                     0.2    0.2    1.79  0.358
```

Let us see now what are the most distinctive words in this major Adam Smith 'oeuvre'!

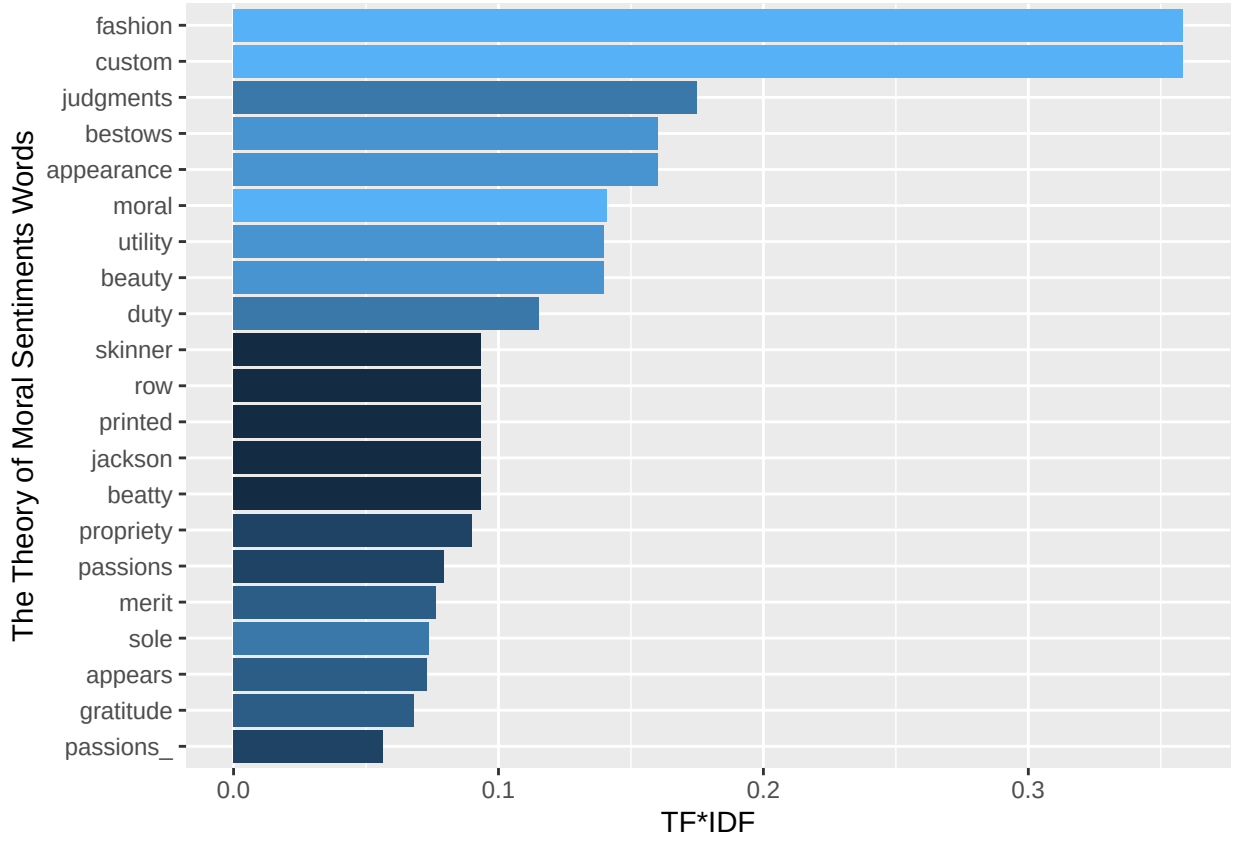
```
Sentiments_TF_IDF %>%
  #We exclude the total column which is not necessary now.
  select(-total) %>%
  #We arrange by TF*IDF in descending order,
  arrange(desc(tf_idf))
```

```
## # A tibble: 246 x 7
##     part word          n Sentiments_Term_Frequency     tf   idf tf_idf
##    <int> <chr>     <int>                      <dbl>  <dbl> <dbl>  <dbl>
##  1     5 custom       39                     0.2    0.2    1.79  0.358
##  2     5 fashion      39                     0.2    0.2    1.79  0.358
##  3     3 judgments    38                     0.0977 0.0977 1.79  0.175
##  4     4 appearance   26                     0.0893 0.0893 1.79  0.160
##  5     4 bestows      26                     0.0893 0.0893 1.79  0.160
##  6     5 moral        25                     0.128  0.128  1.10  0.141
##  7     4 beauty       37                     0.127  0.127  1.10  0.140
##  8     4 utility      37                     0.127  0.127  1.10  0.140
##  9     5 notions      14                     0.0718 0.0718 1.79  0.129
## 10     3 duty         25                     0.0643 0.0643 1.79  0.115
## # i 236 more rows
```

As we can see, some of the most distinctive words are: “custom”, “fashion”, “appearance”, moral”, “notions”, “judgements”, “beauty”, “utility”. In other words, this is what makes this book unique.

Let us visualise this!

```
Sentiments_TF_IDF %>%
  group_by(part) %>%
  slice_max(tf_idf, n = 3) %>%
  ungroup() %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = part)) +
  geom_col(show.legend = FALSE) +
  labs(x = "TF*IDF", y = "The Theory of Moral Sentiments Words")
```



Now, we can observe what has been previously commented on a very nice plot!

So far so good, this summarises the main insights that we can draw from the topic on “term frequency”.

Let us repeat now the same process for “The Wealth of Nations”!

```
Nations_Book_Words <- Tidy_Original_Nations_Book %>%
#I have already tokenised the book, so I will just count the word frequency here.
  count(chapter, word, sort = TRUE)

head(Nations_Book_Words)
```

```
## # A tibble: 6 x 3
##   chapter word         n
```

```
##     <int> <chr>    <int>
## 1      28 price     7280
## 2      28 silver    3435
## 3      28 quantity  3289
## 4      28 corn      2585
## 5      28 produce   2573
## 6      28 0         2526
```

As can be seen, Adam Smith gives a lot of prominance in "The Wealth of Nations" to words such as "silver", "greater", "quantity", "value", "corn", and others.

Next, I will sum up the total number of words contained in each Chapter.

```
Nations_Total_Words <- Nations_Book_Words %>%
  group_by(chapter) %>%
  summarize(total = sum(n))

head(Nations_Total_Words)
```

```
## # A tibble: 6 x 2
##   chapter total
##     <int> <int>
## 1       0   189
## 2       1    79
## 3       2    67
## 4       3   159
## 5       4   257
## 6       5    84
```

As can be easily observed, the structure of "The Wealth of Nations" is highly dissimilar: some chapters are by far lengthier than others.

To be able to continue the work, it is necessary to save the results in a DataFrame.

```
Nations_Words <- left_join(Nations_Book_Words, Nations_Total_Words)
```

```
## Joining with `by = join_by(chapter)`
```

```
head(Nations_Words)
```

```
## # A tibble: 6 x 4
##   chapter word         n  total
##     <int> <chr>    <int>  <int>
## 1      28 price     7280 234693
## 2      28 silver    3435 234693
## 3      28 quantity  3289 234693
## 4      28 corn      2585 234693
## 5      28 produce   2573 234693
## 6      28 0         2526 234693
```

By having done this, we will be able to see now the "term frequency" in "The Wealth of Nations":

```
Nations_Words <- Nations_Words %>%
  mutate(Nations_Term_Frequency = n/total)

head(Nations_Words)
```

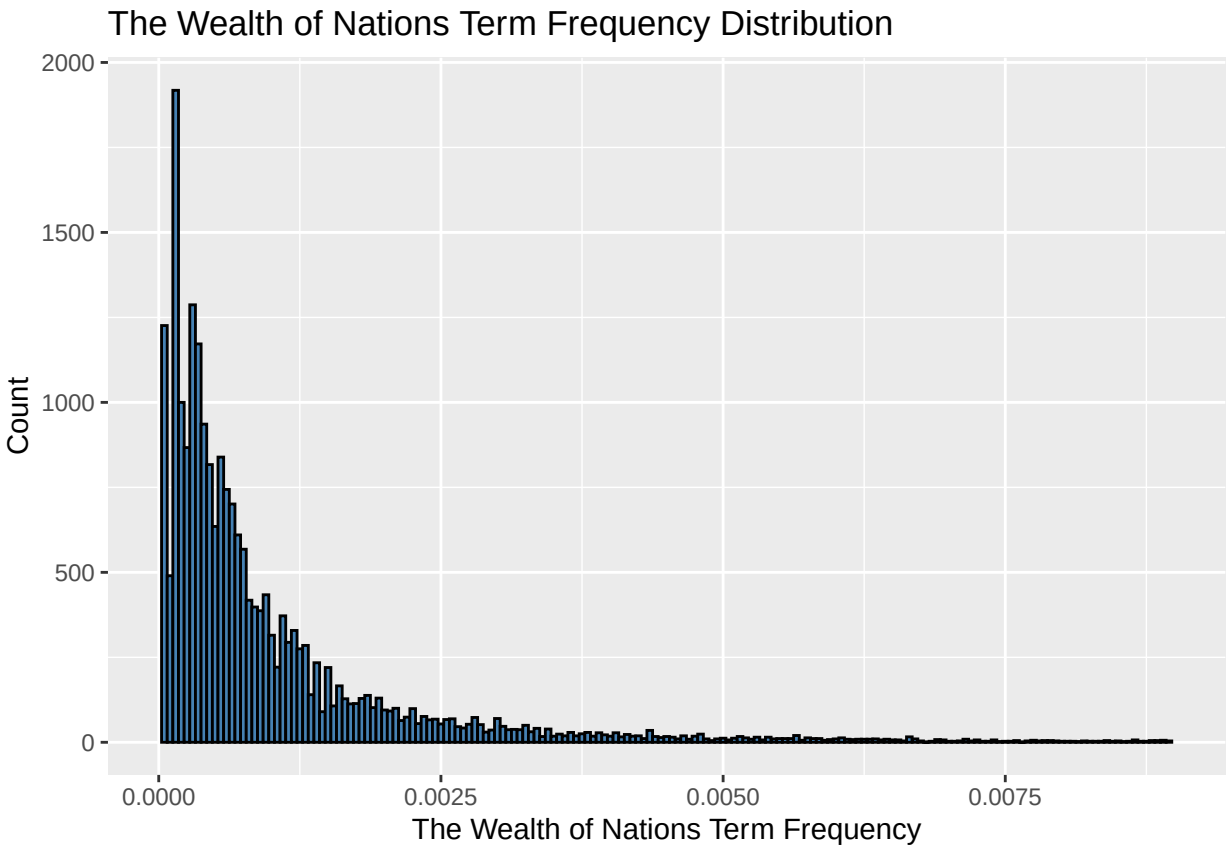
```
## # A tibble: 6 x 5
##   chapter word         n  total Nations_Term_Frequency
##     <int> <chr>    <int>  <int>                  <dbl>
## 1      28 price     7280 234693                 0.0310
## 2      28 silver    3435 234693                 0.0146
## 3      28 quantity  3289 234693                 0.0140
## 4      28 corn      2585 234693                 0.0110
## 5      28 produce   2573 234693                 0.0110
## 6      28 0         2526 234693                 0.0108
```

Let us make a plot of the word distribution in "The Wealth of Nations":

```
ggplot(Nations_Words, aes(x = Nations_Term_Frequency)) +
  geom_histogram(binwidth = 0.00005, color = "black", fill = "steelblue") +
  scale_x_continuous(limits = c(0, 0.009)) +
  labs(title = "The Wealth of Nations Term Frequency Distribution",
       x = "The Wealth of Nations Term Frequency",
       y = "Count")
```

```
## Warning: Removed 499 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



In comparison with what we saw in the case of Adam Smith's previous work, "The Wealth of Nations" does very clearly present a fat-tailed distribution. Mathemathically, this can be interpreted as a "Pareto distribution" or a "power law"; and, linguistically, it means that "The Wealth of Nations" contains many more unique words than "The Theory of Moral Sentiments" does, along with a huge number of "connectors" and the likes, necessary to establish (apologies for the tautology) to make connections among these words.

Let us proceed now with the Zipf's Law (the explanation of which is provided above):

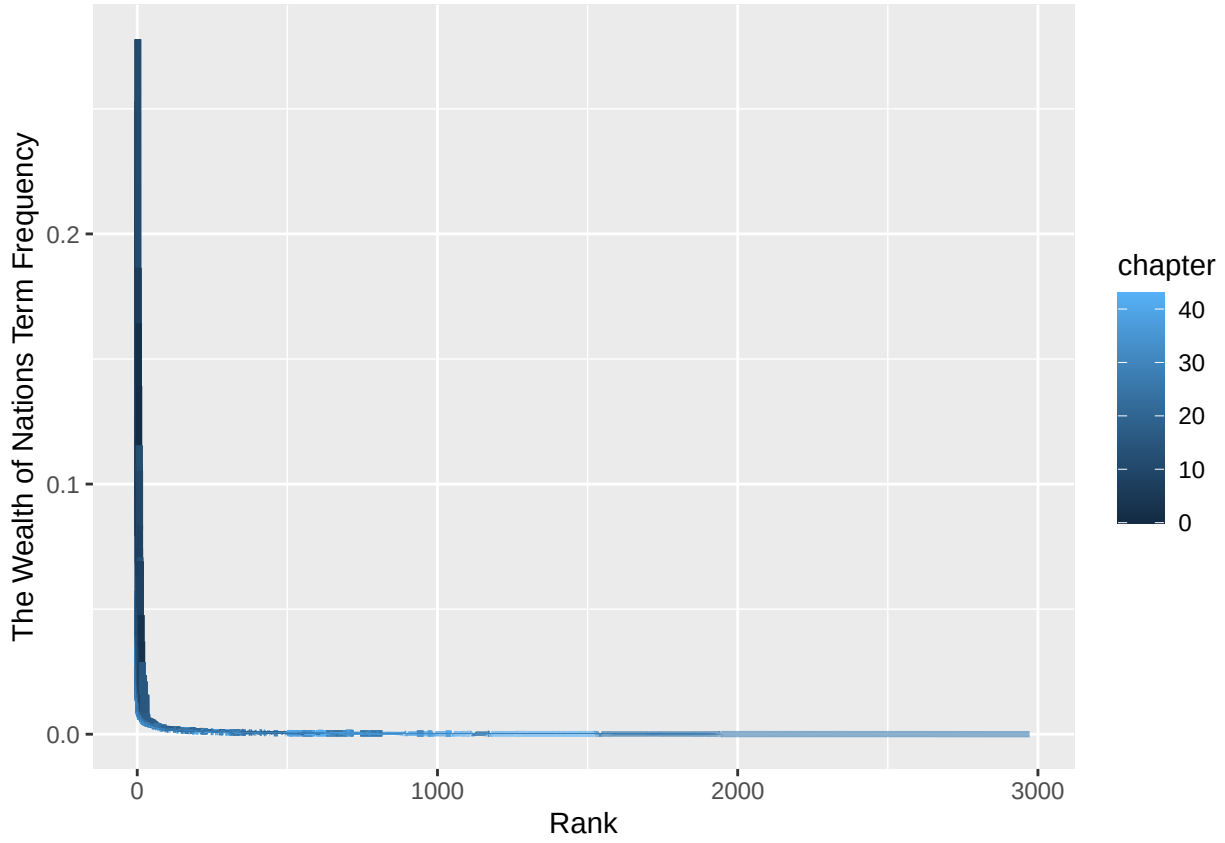
```
Nations_Frequency_By_Rank <- Nations_Words %>%
  group_by(chapter) %>%
  mutate(rank = row_number()) %>%
  ungroup()

head(Nations_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##   chapter word         n  total Nations_Term_Frequency  rank
##     <int> <chr>    <int>  <int>                  <dbl> <int>
## 1      28 price     7280 234693                 0.0310     1
## 2      28 silver    3435 234693                 0.0146     2
## 3      28 quantity  3289 234693                 0.0140     3
## 4      28 corn      2585 234693                 0.0110     4
## 5      28 produce   2573 234693                 0.0110     5
## 6      28 0         2526 234693                 0.0108     6
```

Let's visualise the results of Zipf's Law for "The Theory of Moral Sentiments":

```
Nations_Frequency_By_Rank %>%
  ggplot(aes(rank, Nations_Term_Frequency, colour = chapter)) +
  geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
  xlab("Rank") +
  ylab("The Wealth of Nations Term Frequency")
```



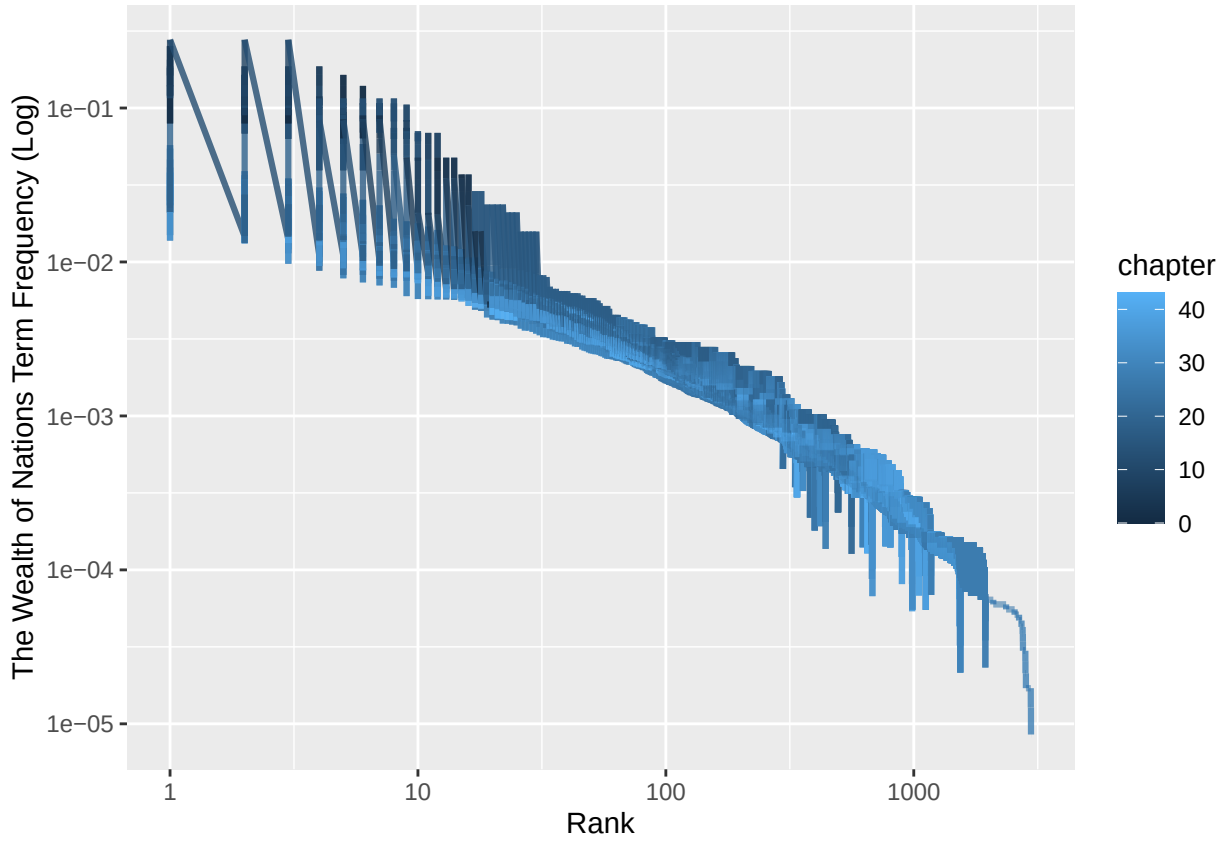

```
head(Nations_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##   chapter word         n  total Nations_Term_Frequency  rank
##     <int> <chr>    <int>  <int>                  <dbl> <int>
## 1      28 price     7280 234693                 0.0310     1
## 2      28 silver    3435 234693                 0.0146     2
## 3      28 quantity  3289 234693                 0.0140     3
## 4      28 corn      2585 234693                 0.0110     4
## 5      28 produce   2573 234693                 0.0110     5
## 6      28 0         2526 234693                 0.0108     6
```

The contrast is even more starking in the case of "The Wealth of Nations". In mathematics and physics, this definitely resembles a "power law". Again, in linguistic terms, what we can take out of this graph is that those words with higher term frequency are lower in the rank.

Let us see how all this plays out on a logarithmic scale!

```
Nations_Frequency_By_Rank %>%
  ggplot(aes(rank, Nations_Term_Frequency, colour = chapter)) +
  geom_line(linewidth = 1.1, alpha = 0.8, show.legend = TRUE) +
  scale_x_log10() +
  scale_y_log10() +
  xlab("Rank") +
  ylab("The Wealth of Nations Term Frequency (Log)")
```




```
head(Nations_Frequency_By_Rank)
```

```
## # A tibble: 6 x 6
##   chapter word         n  total Nations_Term_Frequency  rank
##     <int> <chr>    <int>  <int>                  <dbl> <int>
## 1      28 price     7280 234693                 0.0310     1
## 2      28 silver    3435 234693                 0.0146     2
## 3      28 quantity  3289 234693                 0.0140     3
## 4      28 corn      2585 234693                 0.0110     4
## 5      28 produce   2573 234693                 0.0110     5
## 6      28 0         2526 234693                 0.0108     6
```

Here, we can see again an "inversely proportional relationship will have a constant, negative slope from right to left". However, due to the properties of a "power law" distribution, this relationship follow a line rather than a curve.

Let us proceed with TF*IDF analysis of "The Wealth of Nations"!

```
Nations_TF_IDF <- Nations_Words %>%
  filter(!word %in% c("0", "xi", "iv")) %>%
  bind_tf_idf(word, chapter, n)

head(Nations_TF_IDF)
```

```
## # A tibble: 6 x 8
##   chapter word         n  total Nations_Term_Frequency     tf   idf  tf_idf
##     <int> <chr>    <int>  <int>                  <dbl>  <dbl> <dbl>   <dbl>
## 1      28 price     7280 234693                 0.0310 0.0314 0.526 0.0165
## 2      28 silver    3435 234693                 0.0146 0.0148 1.15  0.0169
## 3      28 quantity  3289 234693                 0.0140 0.0142 0.606 0.00859
## 4      28 corn      2585 234693                 0.0110 0.0111 0.951 0.0106
## 5      28 produce   2573 234693                 0.0110 0.0111 0.417 0.00462
## 6      28 land      2379 234693                 0.0101 0.0102 0.565 0.00579
```

Let us see now what are the most distinctive words in this other major Adam Smith 'oeuvre'!

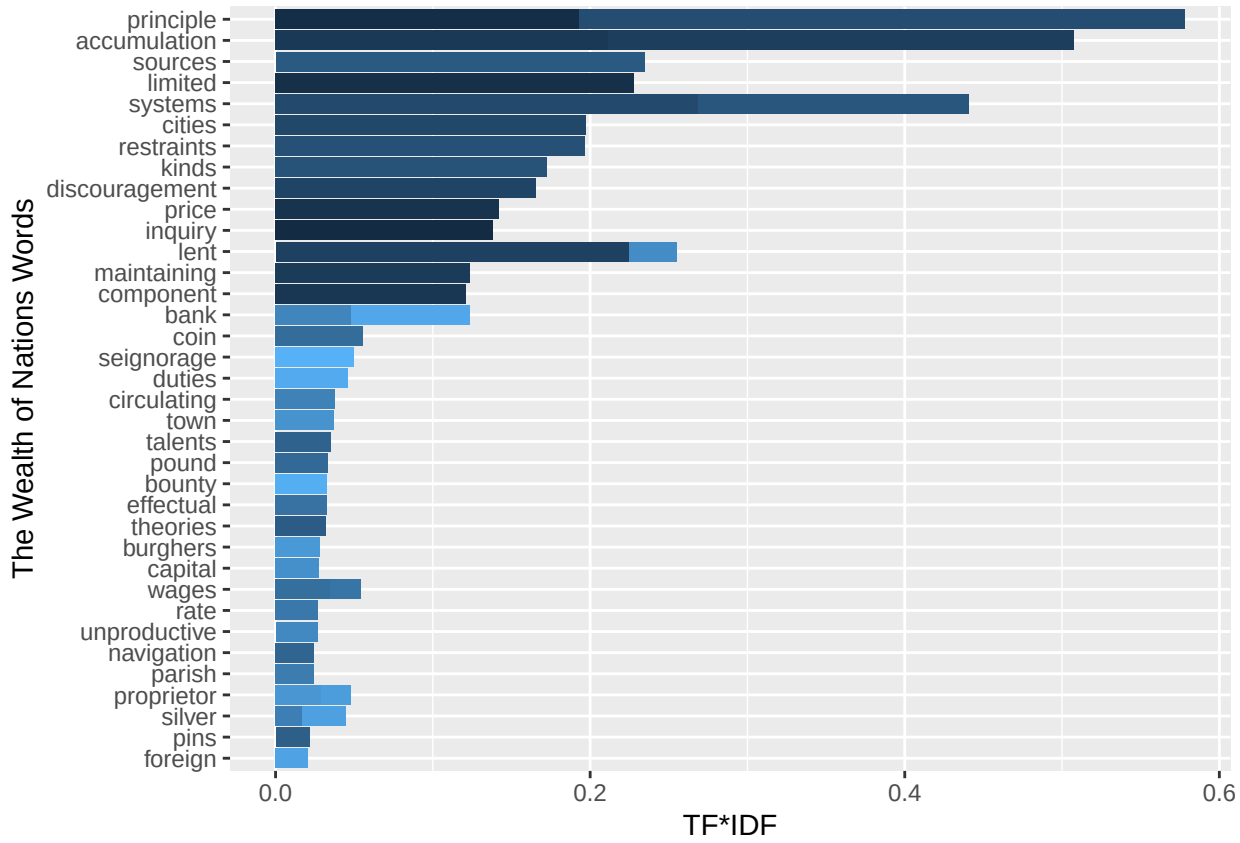
```
Nations_TF_IDF %>%
  #We exclude the total column which is not necessary now.
  select(-total) %>%
  #We arrange by TF*IDF in descending order.
  arrange(desc(tf_idf))
```

```
## # A tibble: 22,818 x 7
##    chapter word            n Nations_Term_Frequency     tf   idf tf_idf
##      <int> <chr>       <int>                  <dbl>  <dbl> <dbl>  <dbl>
##  1      12 principle      10                 0.278  0.278   1.39  0.385
##  2      12 commercial     10                 0.278  0.278   1.22  0.339
##  3       7 accumulation   11                 0.186  0.186   1.59  0.296
##  4      11 systems         9                 0.0818 0.1     2.69  0.269
##  5       7 unproductive    5                 0.0847 0.0847  3.09  0.262
##  6      11 contributed    11                 0.1    0.122   1.99  0.244
##  7      16 sources        11                 0.108  0.108   2.17  0.235
##  8      16 expenses       12                 0.118  0.118   1.99  0.234
##  9       2 limited        11                 0.164  0.164   1.39  0.228
## 10       8 lent            9                 0.0938 0.103   2.17  0.225
## # i 22,808 more rows
```

As we can observe, the words with which Adam Smith is concerned the most in "The Wealth of Nations" are: "price", "silver", "quantity", "corn", and / or "produce".

Let us visualise this!

```
Nations_TF_IDF %>%
  group_by(chapter) %>%
  slice_max(tf_idf, n = 1) %>%
  ungroup() %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = chapter)) +
  geom_col(show.legend = FALSE) +
  labs(x = "TF*IDF", y = "The Wealth of Nations Words")
```



Now, we can neatly view what has been previously commented!

With this, I shall conclude the part on "term frequency".

So far, 3 broad topics have been covered in my piece of research: "working with multiple texts", "sentiment analysis", "and"term frequency".

Concluding remarks: from what we have seen, Text Mining techniques that have been employed in the present piece of analysis of do a relatively good job when it comes to outlining the main characteristics of both Adam Smith works.

However, they are no substitute for human insight, since the similarities that can be found between "The Theory of Moral Sentiments" and "The Wealth of Nations" by means of human, scholar study are hard to be found by employing computational means only.

All in all, the interested reader can find some interesting ideas about both 'oeuvres', but a thorough reading of both works is recommended if one wishes to obtain true insight. Nevertheless, as been previously outlined, what has been done so far is a good initial approximation, one that allows the interested reader to provide relevant information about both texts even without the need to read them first.