# Is There Any Similarity between Philosophy Schools? Can We Roughly Understand Philosophy by Algorithms?

Nichole Zhang uni:qz2446

## Part 1 Introduction

Philosophy is hard to understand for most of us who are not majoring in Philosophy. However, is it feasible for us to learn more about Philosophy with the help of algorithms? In this project, I use the dataset from Kaggle to find the similarities from different schools through different dimensions and try to extract main ideas from some schools' texts.

## Part 2 Data Processing for this Project

### 2.1 Import Data

For this project, I use the dataset called **"philosophy data"**, created by Kourosh Alizadeh. This dataset contains over 300,000 sentences from over 50 texts spanning 13 major schools of philosophy.

```
philo_data <- read.csv("../data/philosophy_data.csv")
```

### 2.2 The Packages Used In This Project

For this project, I use the following nine packages:

**ggplot2** - a system for declaratively creating graphics, based on The Grammar of Graphics

**gplots** - a collection of R programming tools for plotting data, including heatmap.2

**tidyr** - tools to help to create tidy data, where each column is a variable, each row is an observation, and each cell contains a single value

**syuzhet** - a package comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction tool developed in the NLP group at Stanford

**tm** - A framework for text mining applications within R

**wordcloud** - Functionality to create pretty word clouds, visualize differences and similarity be- tween documents, and avoid over-plotting in scatter plots with text

**RColorBrewer** - a tool to choose sensible colour schemes for figures in R

**dplyr** - a grammar of data manipulation

**factoextra** - a package provides some easy-to-use functions to extract and visualize the output of multivariate data analyses

## 2.3 Some Basic Data Summaries for the data

The dataset contains 59 titles, 36 authors, 13 schools and 360780 sentences. The average length of sentences is 151.

```
length(unique(philo_data$title))
```

```
## [1] 59
```

```
length(unique(philo_data$author))
```

```
## [1] 36
```

```
length(unique(philo_data$school))
```

```
## [1] 13
```

```
length(unique(philo_data$sentence_lowered))
```

```
## [1] 360780
```

```
summary(philo_data$sentence_length)
```

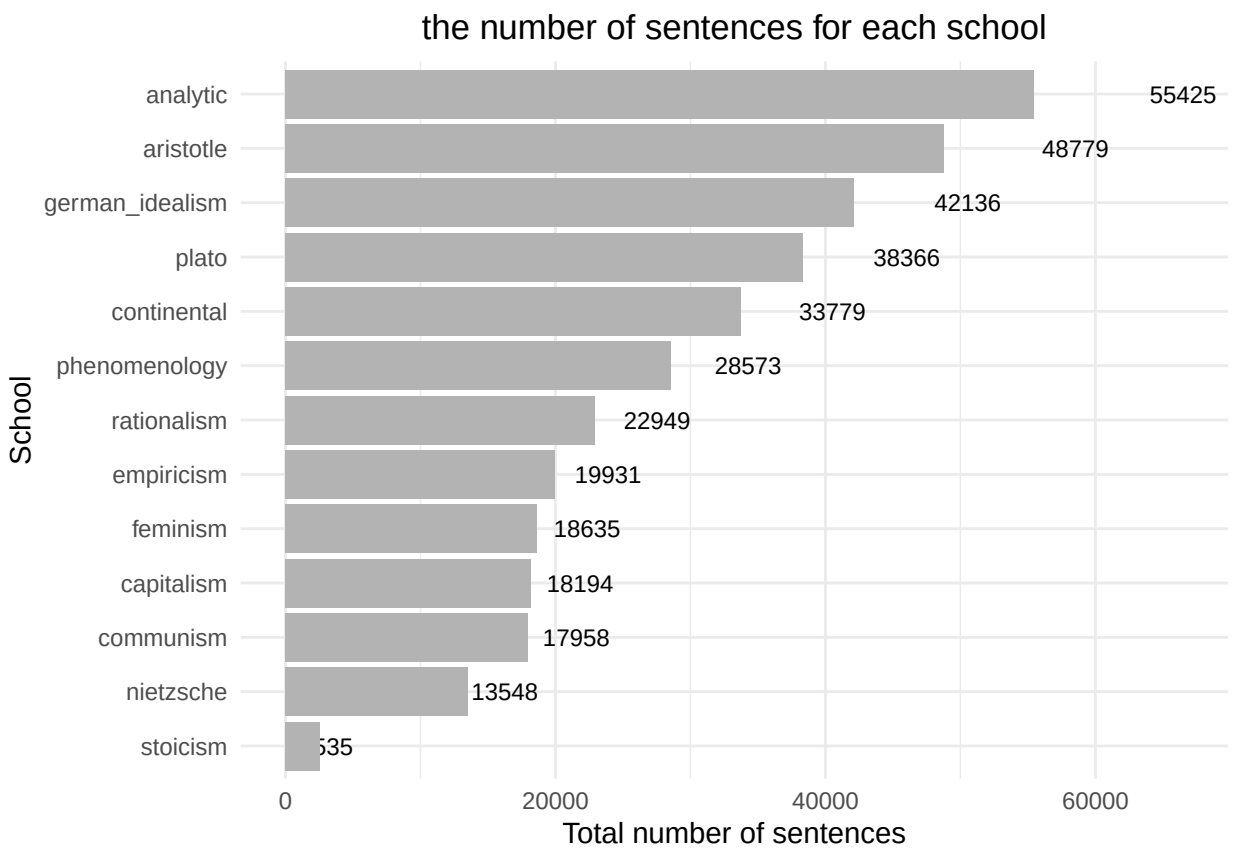
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    75.0   127.0   150.8   199.0  2649.0
```

## 2.3 Limitation

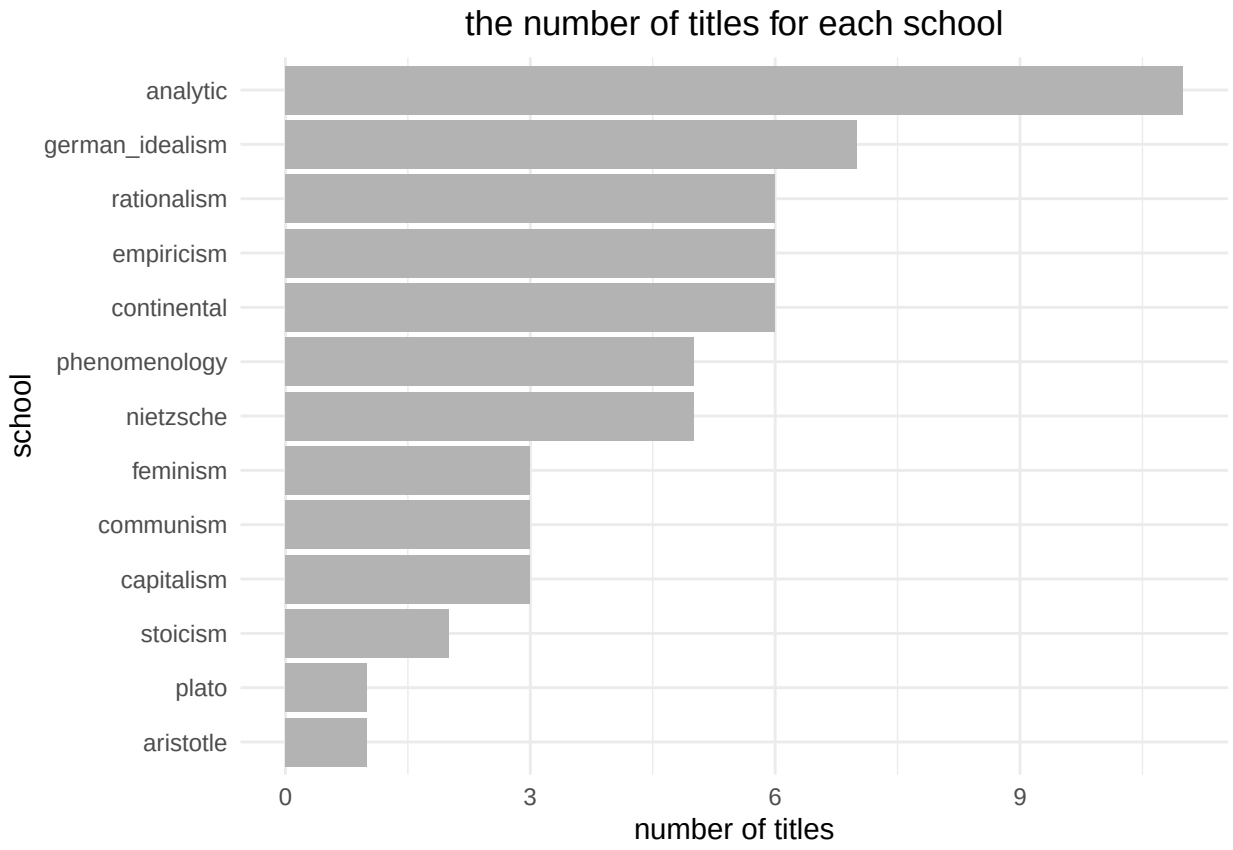
Most of the schools have more than 10,000 sentences in the dataset, however, Stoicism only has 535 sentences. It might be biased for concluding the main idea of stoicism by such a few sentences.

On average, there are only 5 books and 3 authors for each school, which means our conclusion for these schools may not reflect the true idea of these schools. We need more lines, if we want to understand their philosophy idea more accurately.

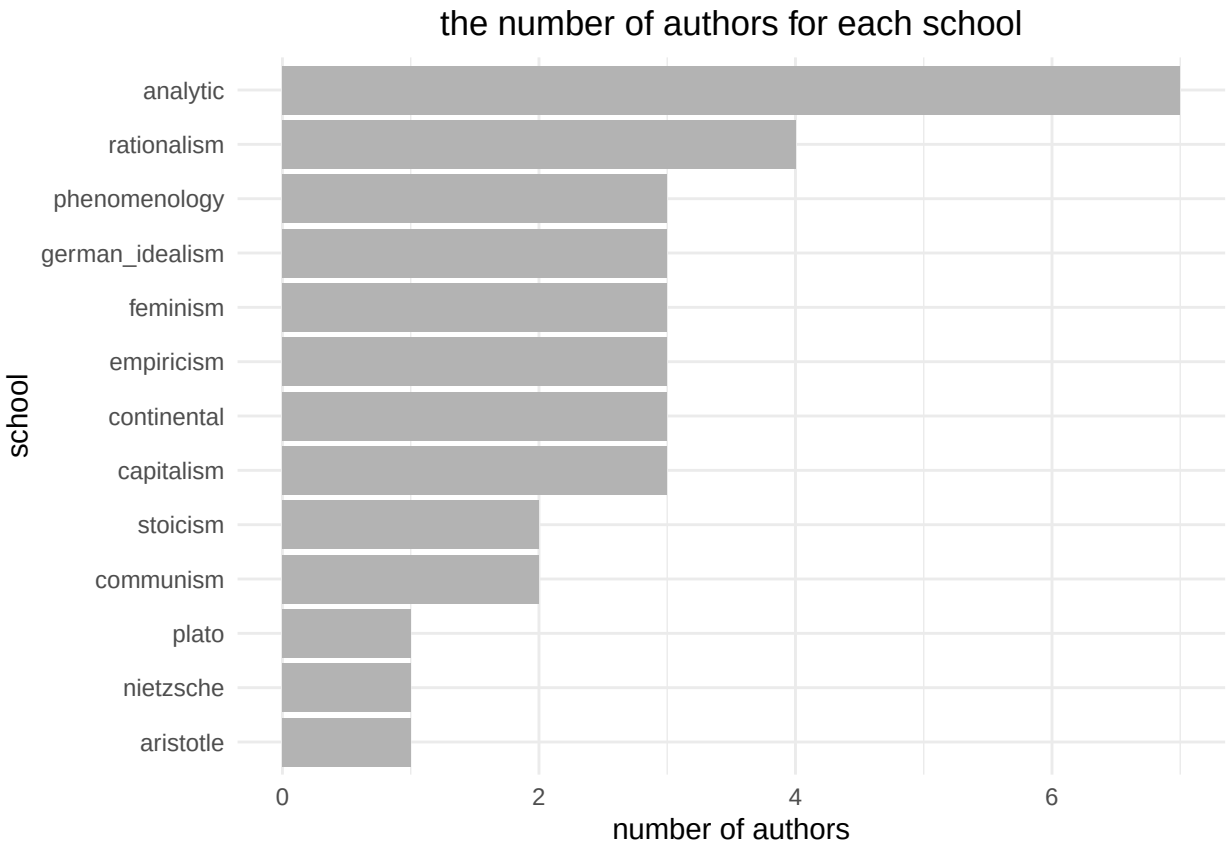
```
par(mfrow=c(2,2))
#show the number of sentences for each school
sen_school <- philo_data %>%
  group_by(school) %>%
  summarise(count = n())
ggplot(sen_school, aes(x = count, y = reorder(school,count))) +
  geom_text(aes(label = count),position = position_stack(vjust = 1.2), size = 3) +
  geom_col(fill = "gray70") +
  theme_minimal() +
  xlab("Total number of sentences")+
  ylab("School")+
  labs(title = "the number of sentences for each school")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
#show the number of titles for each school
tmp <- unique(philo_data[, 1:3])
show_tmp <- tmp %>%
             group_by(school) %>%
             count(school, sort = TRUE)
ggplot(show_tmp, aes(x = n, y = reorder(school,n))) +
  geom_col(fill = "gray70") +
  theme_minimal()+
  xlab("number of titles")+
  ylab("school")+
  labs(title = "the number of titles for each school")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
#show the number of authors for each school
tmp1 <- unique(philo_data[, c(2,3)])

show_tmp1 <- tmp1 %>%
             group_by(school) %>%
             count(school, sort = TRUE)
ggplot(show_tmp1, aes(x = n, y = reorder(school,n))) +
  geom_col(fill = "gray70") +
  theme_minimal()+
  xlab("number of authors")+
  ylab("school")+
  labs(title = "the number of authors for each school")+
  theme(plot.title = element_text(hjust = 0.5))
```




# Part 3 Deep Diving into Sentences

## 3.1 Preparation

Besides the stopwords in **tm** package, I also add some other words, which usually appear in philosophy texts.

```
stopword <- c(stopwords("english"),"will","can","one","may","must","things","without","certain","yet","
```

## 3.2 Clustering Schools by Two Ways

### 3.2.1 Which schools are discussing about similar topics?

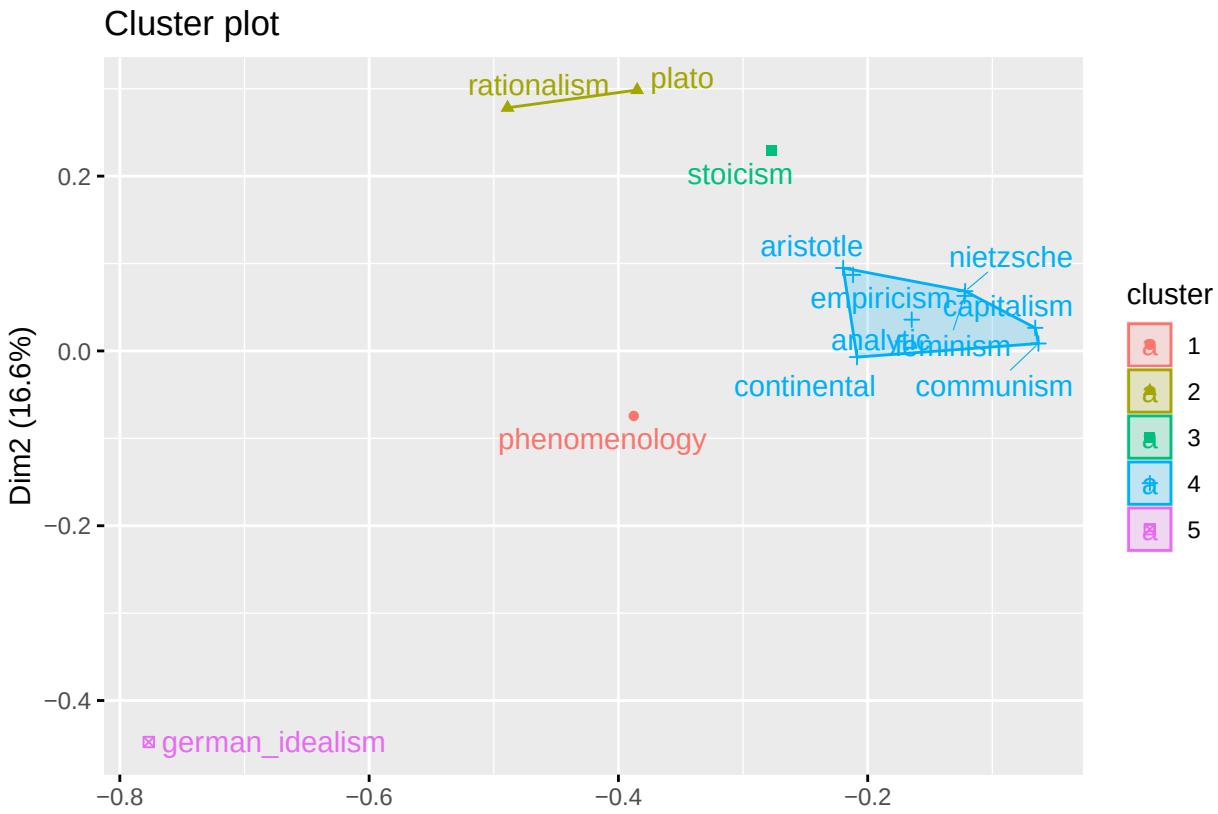
After filtering stopwords, white spaces and punctuations, I find the top 200 high-frequency words for each school, and combine these 2,600 words to build a Document-Term Matrix. Then, I use a K-means Clustering algorithm to find the clusters of these schools.

```
tweets <- rep(NA,13)
tdm.result <- data.frame()
l <- rep(NA, 13)
schools <- c("analytic", "aristotle", "german_idealism", "plato", "continental", "phenomenology", "ratio
i = 1
for (name in schools) {
  data_tmp <- philo_data[philo_data$school == name, ]
```

```
  docs <- Corpus(VectorSource(data_tmp$sentence_str))
  docs<-tm_map(docs, stripWhitespace)
  docs<-tm_map(docs, content_transformer(tolower))
  docs<-tm_map(docs, removeWords, stopword)
  docs<-tm_map(docs, removeWords, character(0))
  docs<-tm_map(docs, removePunctuation)
  l[i] <- length(docs)
  tweets[i] <- iconv(docs)
  i = i+1
  tdm.all<-TermDocumentMatrix(docs)
  tdm.tidy=tidy(tdm.all)
  tdm.overall=summarise(group_by(tdm.tidy, term), sum(count))
  tmp <- tdm.overall %>%
    arrange(desc(`sum(count)`)) %>%
    head(200)
  tmp <- as.data.frame(tmp)
  tmp <- cbind(rep(name, 200), tmp)
  tdm.result <- rbind(tdm.result, tmp)

}
```

```
#combine words and build a dtm
colnames(tdm.result) <- c("school","term","time")
tdm.overall <- tdm.result %>%
  spread(key = term,value =time)
tdm.overall <- as.data.frame(tdm.overall)
column <- length(tdm.overall)
tdm.overall[is.na(tdm.overall)] <- 0
rownames(tdm.overall) <- tdm.overall[,1]
tdm.overall <- tdm.overall[ ,2:column]
```

```
# Use k-means clustering
km.res.word=kmeans(tdm.overall/l, iter.max=200, 5)
fviz_cluster(km.res.word,
             stand=F, repel= TRUE,
             data = tdm.overall/l, xlab="", xaxt="n",
             show.clust.cent=FALSE)
```

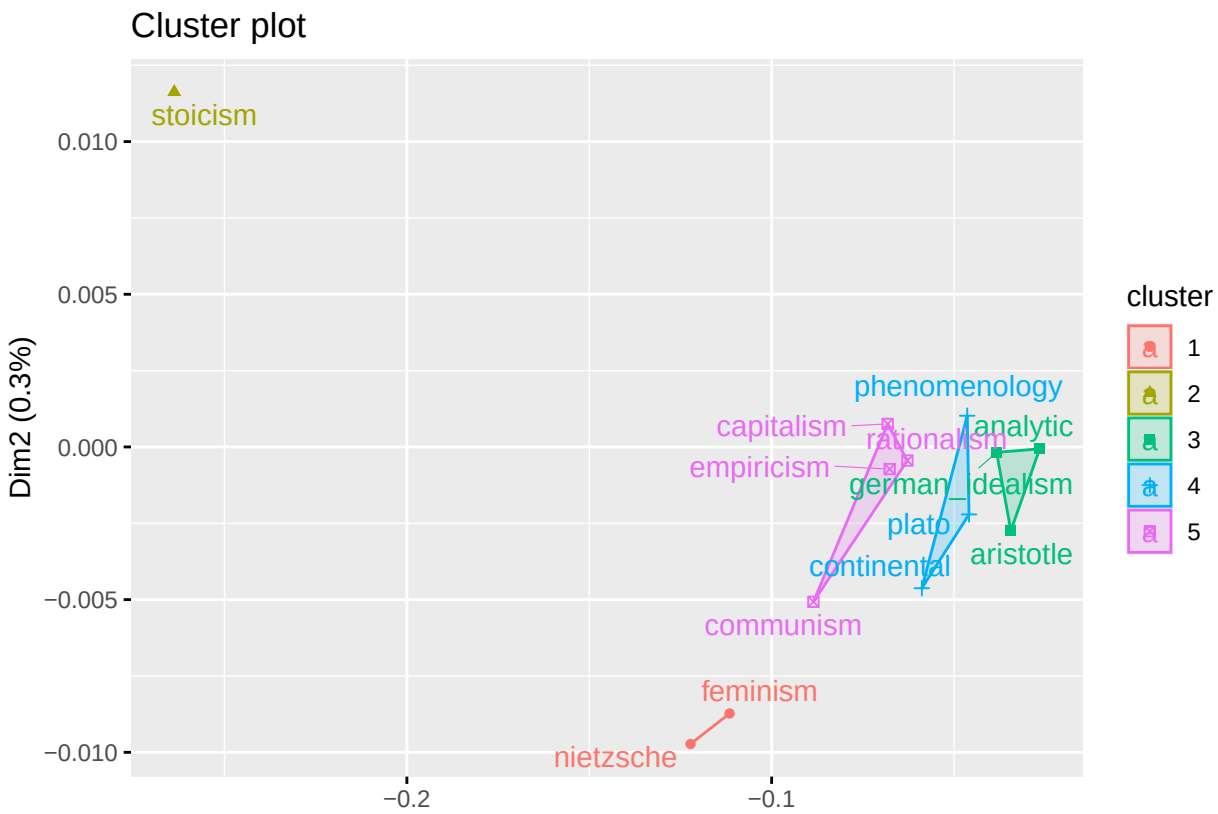



By k-mean clustering, we can figure out that those schools, **Aristotle**, **Nietzsche**, **Empiricism**, **Capitalism**, **Analytic**, **Feminism**, **Continental** and **Communism**, are more likely to discuss similar topics in texts. **Plato** and **rationalism** also focus on similar topics.

### 3.2.2 Which schools are writing with similar sentiments?

By analyzing sentiments for each word of all schools and clustering algorithm, we can find those schools writing with similar sentiments.

```
s <- get_nrc_sentiment(tweets[1:(i-1)])
rownames(s) <- schools
senti_summary <- s/l
km.res=kmeans(senti_summary[,1:8], iter.max=200,
              5)
fviz_cluster(km.res,
             stand=F, repel= TRUE,
             data = senti_summary[,1:8], xlab="", xaxt="n",
             show.clust.cent=FALSE)
```

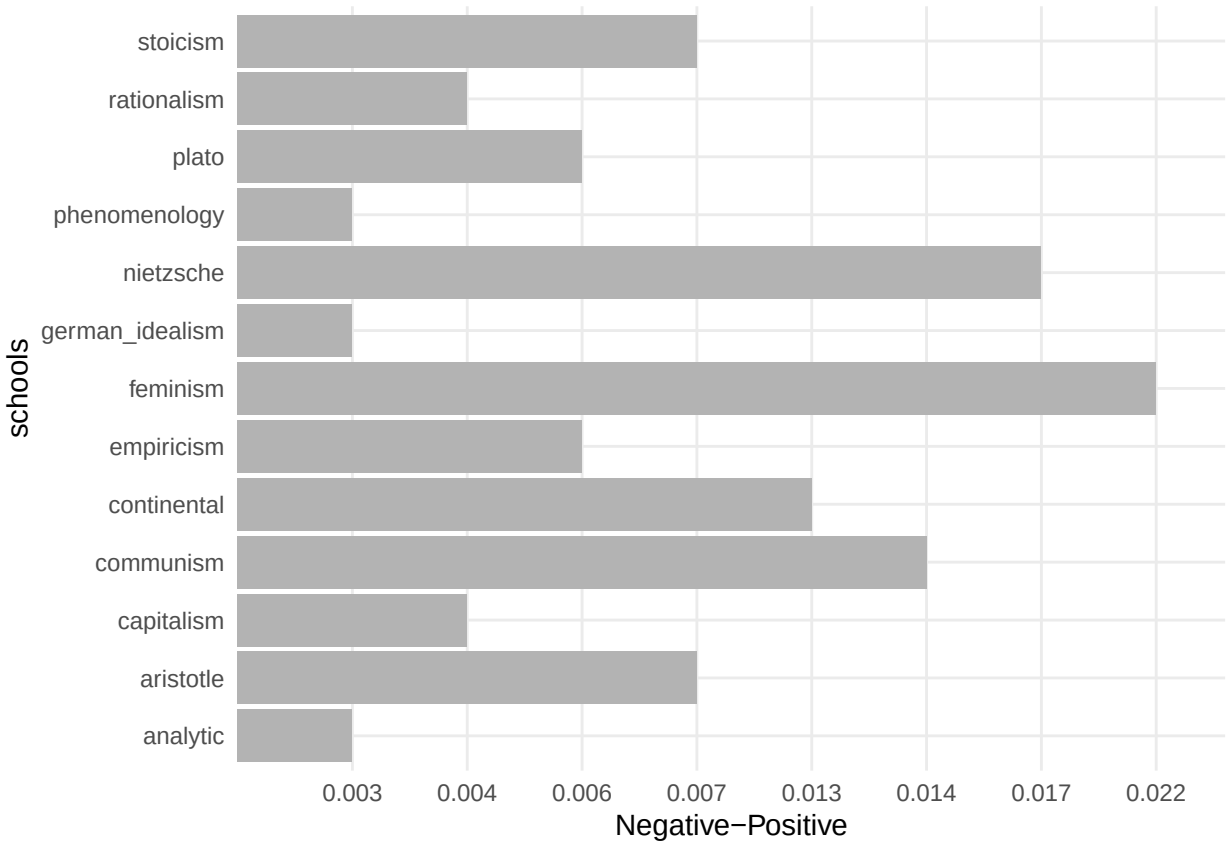



By the graph above, we can point out that four groups: {**Feminism** & **Nietzsche**}, {**Plato** & **Phenomenology** & **Continental**}, {**Capitalism** & **Rationalism** & **Communism** & **Empiricism**}, {**Aristotle** & **German Idealism**}. Each group's texts are written by similar sentiments.

### 3.3 Sentiment of each school is positive or negative?

We can use the sentiment dtm built before to do the analysis.

```
neg_pos <- round(senti_summary[,9] - senti_summary[,10],3)
neg_pos_res <- cbind(schools, neg_pos)
neg_pos_res <- as.data.frame(neg_pos_res)
ggplot(neg_pos_res, aes(schools, neg_pos))+
  geom_col()+
  coord_flip()+
  geom_col(fill = "gray70") +
  theme_minimal()+
  ylab("Negative-Positive")
```




All schools' sentiments are negative. Feminism is the most negative one.

# Part 4 Extract Main Ideas for Some Schools from Texts

For most people, philosophy texts are obscure. Analyzing the dataset can help us understand the basic ideas about some schools. Two examples are given as follows.

## 4.1 Feminism

The main topic that Feminism is always discussing is **Woman**. Therefore, I extract all sentences containing **women are** from Feminism texts, and build a word cloud.

```
woman_data <- philo_data[philo_data$school == "feminism", ]
woman_docs <- woman_data$sentence_lowered[grepl("women are", woman_data$sentence_lowered)]
woman_docs <- Corpus(VectorSource(woman_docs))
woman_docs <-tm_map(woman_docs, stripWhitespace)
woman_docs <-tm_map(woman_docs, content_transformer(tolower))
woman_docs <-tm_map(woman_docs, removeWords, stopword)
woman_docs <-tm_map(woman_docs, removeWords, character(0))
woman_docs <-tm_map(woman_docs, removePunctuation)
len1 <- length(woman_docs)
tdm.all<-TermDocumentMatrix(woman_docs)
tdm.tidy=tidy(tdm.all)
tdm.overall=summarise(group_by(tdm.tidy, term), sum(count))
wordcloud(tdm.overall$term, tdm.overall$`sum(count)`,
```

```
          scale=c(5,2),
          max.words=100,
          min.freq=1,
          random.order=FALSE,
          rot.per=0.3,
          use.r.layout=T,
          random.color=FALSE,
          colors=brewer.pal(12,"Paired"))
```


```
tdm.overall %>%
  arrange(desc(`sum(count)`))%>%
  head(10)
```

```
## # A tibble: 10 x 2
##    term  `sum(count)`
##    <chr>        <dbl>
##  1 women          218
##  2 men             57
##  3 woman           19
##  4 even            17
##  5 world           14
##  6 made            13
##  7 man             12
##  8 like            11
##  9 love            10
## 10 two             10
```

The word **even** is a high frequency word in these sentences, showing that Feminism advocate women and men should be even. It is quite close to the main idea of Feminism: "Feminism incorporates the position that societies prioritize the male point of view, and that women are treated unjustly within those societies. Efforts to change that include fighting against gender stereotypes and establishing educational, professional, and interpersonal opportunities and outcomes for women that are equal to those for men."

## 4.2 Rationalism and Empiricism

There always exists a dispute between rationalism and empiricism about how can people acquire knowledge. Thus, I extract all sentences containing **knowledge** from both Rationalism and Empiricism texts, and build a word cloud.

### Rationalism

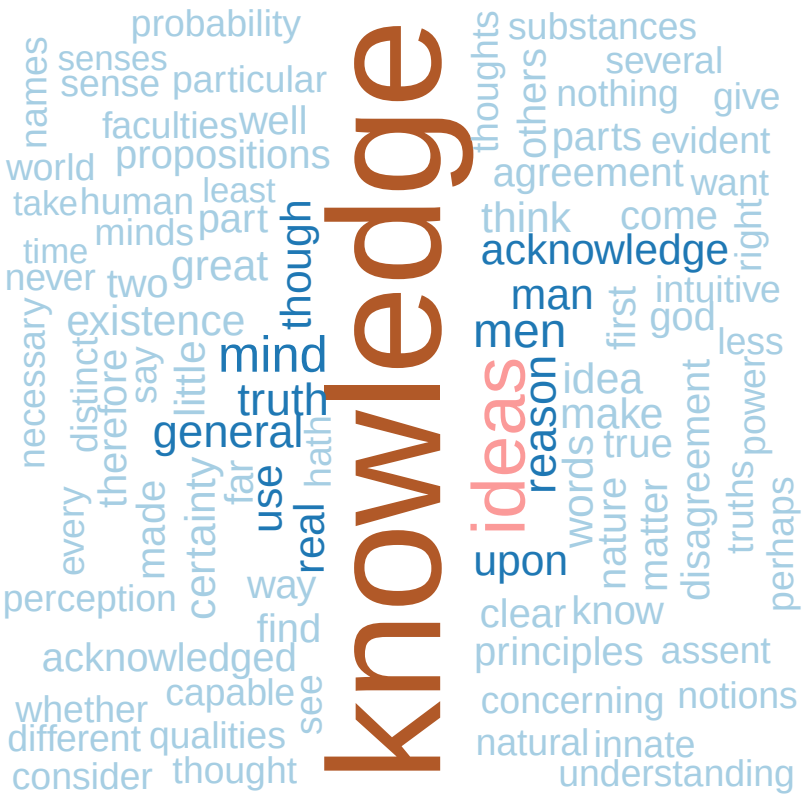
```
tweet_rat_emp <- rep(NA,2)

rational_data <- philo_data[philo_data$school == "rationalism", ]
rational_docs <- rational_data$sentence_lowered[grepl("knowledge",rational_data$sentence_lowered)]
rational_docs <- Corpus(VectorSource(rational_docs))
rational_docs <-tm_map(rational_docs, stripWhitespace)
rational_docs <-tm_map(rational_docs, content_transformer(tolower))
rational_docs <-tm_map(rational_docs, removeWords, stopword)
rational_docs <-tm_map(rational_docs, removeWords, character(0))
rational_docs <-tm_map(rational_docs, removePunctuation)
len1 <- length(rational_docs)
tweet_rat_emp[1] <- iconv(rational_docs)
tdm.all<-TermDocumentMatrix(rational_docs)
tdm.tidy=tidy(tdm.all)
tdm.overall=summarise(group_by(tdm.tidy, term), sum(count))
wordcloud(tdm.overall$term, tdm.overall$`sum(count)`,
          scale=c(5,2),
          max.words=100,
          min.freq=1,
          random.order=FALSE,
          rot.per=0.3,
          use.r.layout=T,
          random.color=FALSE,
          colors=brewer.pal(12,"Paired"))
```


```
tdm.overall %>%
  arrange(desc(`sum(count)`))%>%
  head(15)
```

```
## # A tibble: 15 x 2
##    term      `sum(count)`
##    <chr>            <dbl>
##  1 knowledge          735
##  2 god                240
##  3 mind               135
##  4 nature              89
##  5 body                83
##  6 idea                82
##  7 know                74
##  8 far                 71
##  9 good                69
## 10 true                69
## 11 ideas               64
## 12 love                59
## 13 kind                58
## 14 truth               58
## 15 order               57
```

**Empiricism**

```
empiricism_data <- philo_data[philo_data$school == "empiricism", ]
empiricism_docs <- empiricism_data$sentence_lowered[grepl("knowledge",empiricism_data$sentence_lowered)]
empiricism_docs <- Corpus(VectorSource(empiricism_docs))
empiricism_docs <-tm_map(empiricism_docs, stripWhitespace)
empiricism_docs <-tm_map(empiricism_docs, content_transformer(tolower))
empiricism_docs <-tm_map(empiricism_docs, removeWords, stopword)
empiricism_docs <-tm_map(empiricism_docs, removeWords, character(0))
empiricism_docs <-tm_map(empiricism_docs, removePunctuation)
len2 <- length(empiricism_docs)
tweet_rat_emp[2] <- iconv(empiricism_docs)
tdm.all2<-TermDocumentMatrix(empiricism_docs)
tdm.tidy2=tidy(tdm.all2)
tdm.overall2=summarise(group_by(tdm.tidy2, term), sum(count))
wordcloud(tdm.overall2$term, tdm.overall2$`sum(count)`,
          scale=c(5,1),
          max.words=100,
          min.freq=1,
          random.order=FALSE,
          rot.per=0.3,
          use.r.layout=T,
          random.color=FALSE,
          colors=brewer.pal(12,"Paired"))
```




```
tdm.overall2 %>%
  arrange(desc(`sum(count)`))%>%
  head(15)
```

```
## # A tibble: 15 x 2
##    term        `sum(count)`
##    <chr>              <dbl>
##  1 knowledge            919
##  2 ideas                308
##  3 mind                 129
##  4 men                  113
##  5 truth                 90
##  6 general               89
##  7 reason                88
##  8 though                84
##  9 use                   84
## 10 man                   83
## 11 real                  81
## 12 upon                  80
## 13 acknowledge           79
## 14 great                 76
## 15 words                 76
```

The high frequency words in **Rationalism** are **"god"**, **"nature"**, **"love"** and etc, which are more related to virtual side. In contrast, the high frequency words in **Empiricism** are **"truth"**, **"real"**, and etc, which are more focus on reality.

They are quite similar with the main difference between Rationalism and Empiricism: "Rationalism is the knowledge that is derived from reason and logic while on the other hand empiricism is the knowledge that is derived from experience and experimentation."

# Part 5 Reference


Kaggle - History of Philosophy: https://www.kaggle.com/kouroshalizadeh/history-of-philosophy

Wikipedia - Feminism: https://en.wikipedia.org/wiki/Feminism

Difference Between Rationalism and Empiricism: https://askanydifference.com/difference-between-rationalism-and-empiricism-with-table/