School Differentiation

Representative Vocabularies

Sentences Sentiment Change

5243_Project1

Code

Jianing Hu

School Differentiation

Code

```
##             school
## 1            plato
## 2        aristotle
## 3       empiricism
## 4      rationalism
## 5         analytic
## 6      continental
## 7    phenomenology
## 8  german_idealism
## 9        communism
## 10      capitalism
## 11        stoicism
## 12       nietzsche
## 13        feminism
```

Because I was so new to the field of philosophy, the moment I opened the data I had a childish question. Why are there so many schools of thought in philosophy, and how do people distinguish between them? As a philosophical “idiot,” may I be able to quickly understand the main ideas of each school? Of course, this is a fantasy. But I remembered that I once saw a video on how to quickly improve reading speed. That video shows that when people read text, they only need to read the first three or four letters of each word to automatically make up the entire word. So we may conclude that the first three to four letters of most vocabulary can be used to distinguish between vocabulary. By analogy, can we then distinguish schools by extracting some of the words from each school of thought? Also, Whether the trend of sentences sentiment change varies in each articles from different schools or not. Whether these differences can help us distinguish between different schools.

Representative Vocabularies

Now, a difficulty emerged. How can we find the **Representative Vocabularies**. Since our data only contain 59 texts in total, it is absurd to use three or four texts (even not books) to find a representative vocabularies of a school. But at the moment I’m just trying to make an attempt at that.

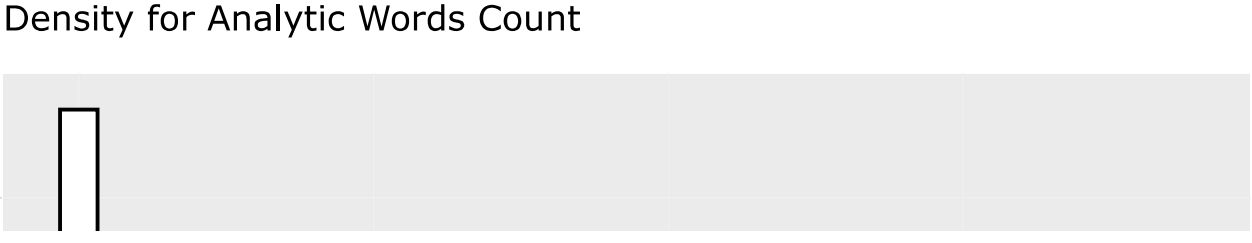
It's obvious we cannot simply pick the words with the highest frequency. Look at the tibble ordered by counts (frequencies) of words here, it's not hard to find the top words are all commonly used in all schools. So, they are not representative of the school – analytic.

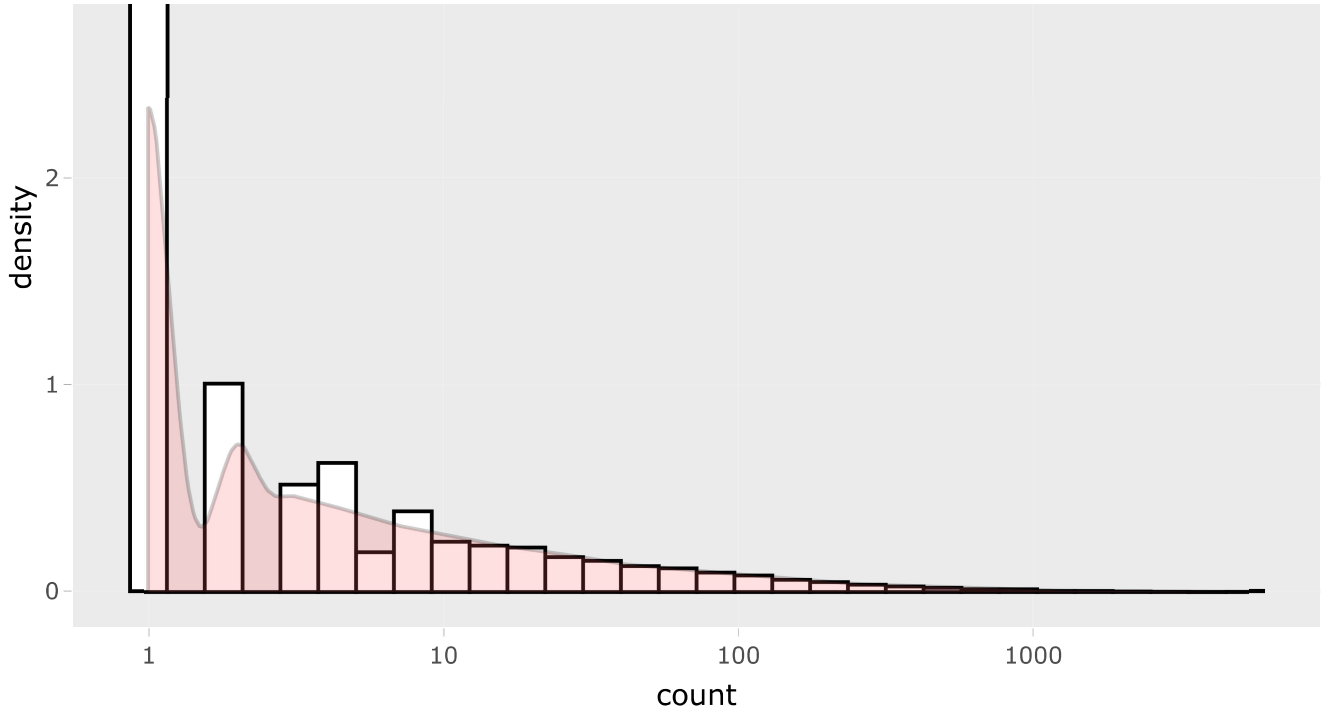
```
## # A tibble: 10 x 3
##    term  count overall_count
##    <chr> <int>         <int>
##  1 one    5194         38579
##  2 the    4840         39104
##  3 can    4670         24106
##  4 may    3414         14651
##  5 but    3409         22275
##  6 say    3295         12107
##  7 true   2541          8366
##  8 will   2498         23028
##  9 sense  2230          8253
## 10 case   2186          8932
```

The words we really want to extract should be the word that one school has a lot of but other schools mention little or even not mention. This means, we can obtain a word's expected frequency in a school according to its frequency in all of the schools. Then, its representativeness can then be judged based on the percentage at which it exceeds the expected frequency. Take the word *one* in the previous table as an example. We know there are `sum(Overall$count)` words in all texts, and `sum(Analytic$count)` words in the texts of Analytic school. Since the word *one* shows up 38579 times in all texts, its expected frequency in Analytic school should be $\text{Overall Frequency} * \frac{sum(\text{Analytic Words})}{sum(\text{Overall Words})}$ which is `38579*sum(Analytic$count)/sum(Overall$count)`. So, the percentage of the word *one* exceed the expected frequency is $\frac{5194 - \text{Expected Frequency}}{\text{Expected Frequency}}$ which is `(5194 - 38579*sum(Analytic$count)/sum(Overall$count)) / (38579*sum(Analytic$count)/sum(Overall$count))`. In this way, we can find the percentage of all words exceeding the expected frequency, and sort the percentages in descending order. And then, is the word at the top of this list what we want?

```
## # A tibble: 2 x 5
##   term         count overall_count expected diff_percent
##   <chr>        <int>         <int>    <dbl>        <dbl>
## 1 unanalysable     4             4    0.495         7.07
## 2 theory        1899          3163  392.            3.85
```

This table can clearly tell us that the selection of representative words cannot be completely dependent on the percentages. We cannot say word *unanalysable* are more representative than *theory* just because *unanalysable* are not mentioned by any other schools. But the frequency of *theory* are mentioned 384.78% more than the general frequency wth 1899 times. It should be a **Representative Vocabularies** for Analytic school. So, the percentages should be weighted according to the frequency (count). But how to set this weight has become a difficult problem to overcome.





Since both the frequencies and the percentages will impact our decision on selecting **Representative Vocabularies**, the weight should be related to the frequencies. Here, we make the histogram and density plot for *count* (frequency). It's a dynamic plot, we can play with it. The **Representative Vocabularies** we really want is those that are close to the end of the tail in this plot. In other words, our selection should be the words with higher percentage of exceeding the general frequency, but its should have some universality for the school at the same time. Then we can infer the percentages of the words with higher frequency should be enclosed with a higher weight, and vice versa. I spent a lot of time on producing am appropriate weight factor for each frequency, but I haven't been able to come to a conclusion until now. It should be close to being proportional to frequency, but it also need to be smaller when the frequency is large enough. At the moment I think the logarithmic ratio might be a good choice, so we choose it for now.

```
## # A tibble: 10 x 6
##    term            count overall_count expected diff_percent weighted_diff
##    <chr>           <int>         <int>    <dbl>        <dbl>         <dbl>
##  1 sentences         919          1023    127.          6.25          42.7
##  2 counterfactual    416           416     51.5         7.07          42.7
##  3 frege             363           364     45.1         7.05          41.6
##  4 russell           397           407     50.4         6.88          41.1
##  5 modal             302           311     38.5         6.84          39.1
##  6 counterfactuals   212           212     26.3         7.07          37.9
##  7 nozick            198           198     24.5         7.07          37.4
##  8 iff               195           195     24.1         7.07          37.3
##  9 sentence         1050          1333    165.          5.36          37.3
## 10 statements        975          1227    152.          5.42          37.3
```

Now, the mimic frequencies are needed for better plo of word cloud.

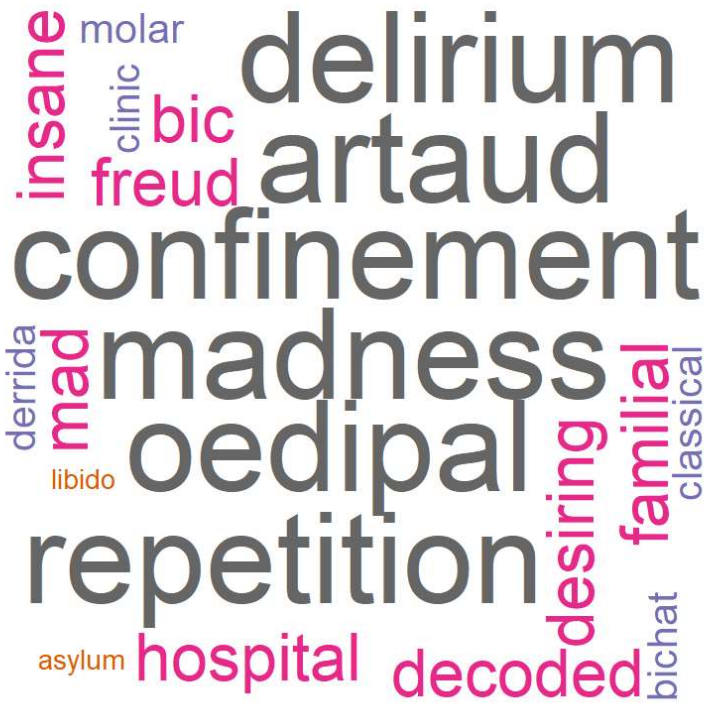
Analytic

Aristotle

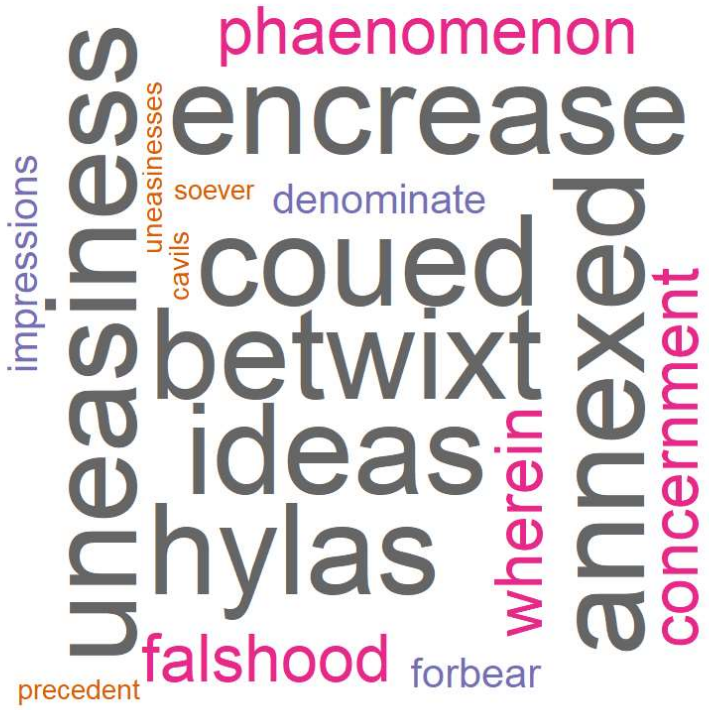
Capitalism

Communism

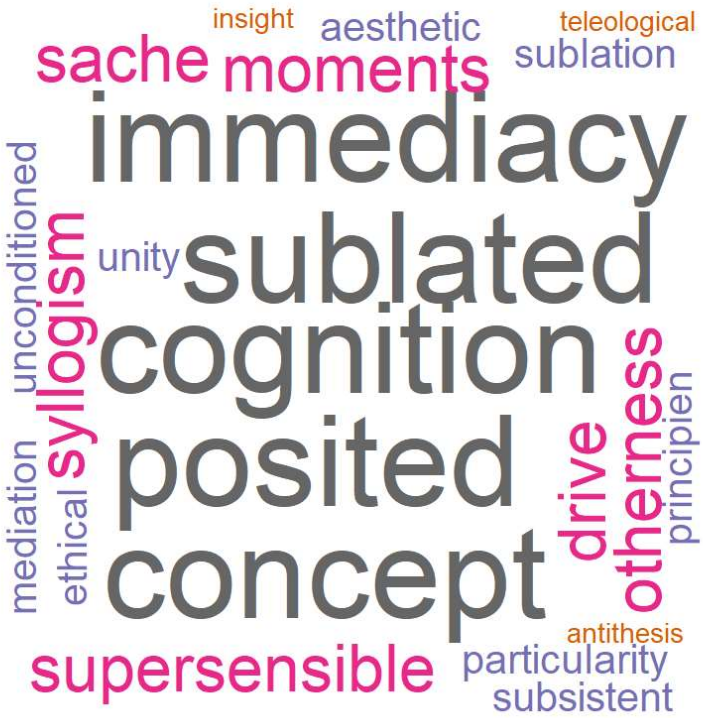
Continental



Empiricism

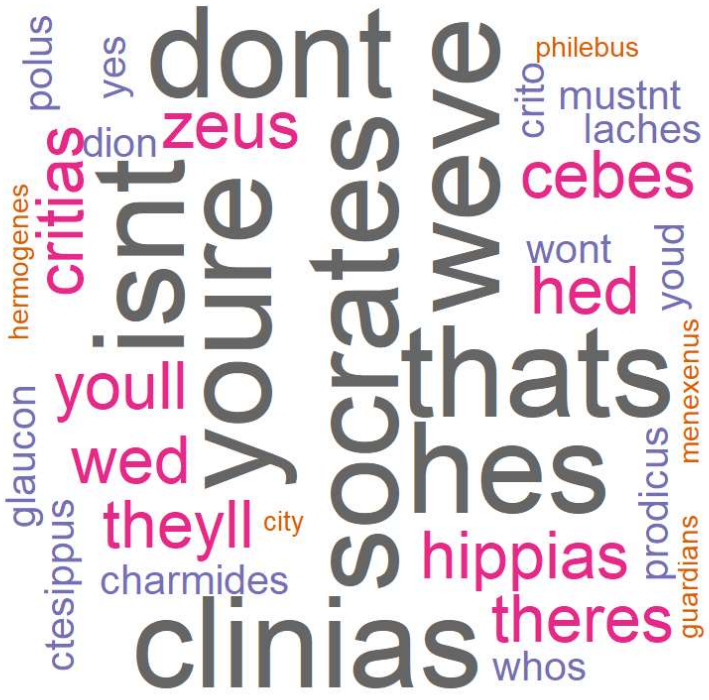


German_idealism



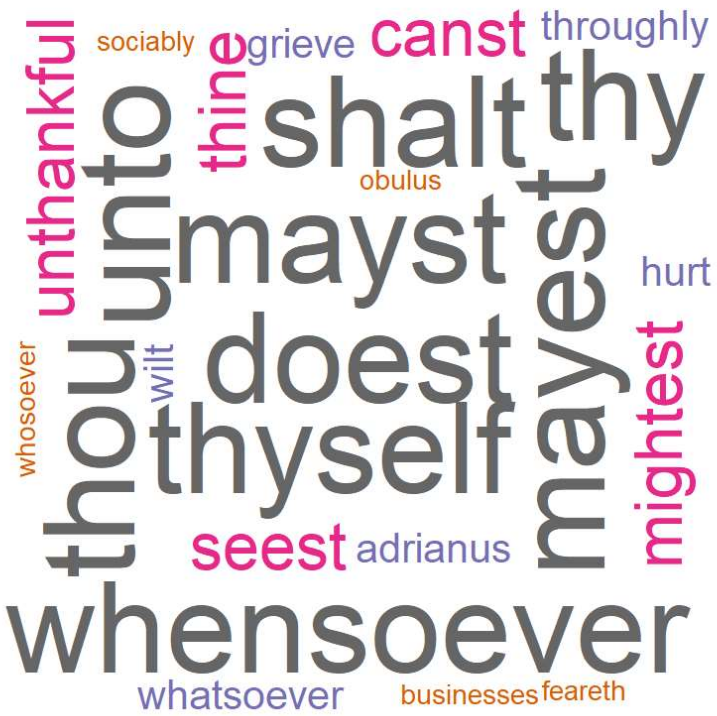
Phenomenology

Plato



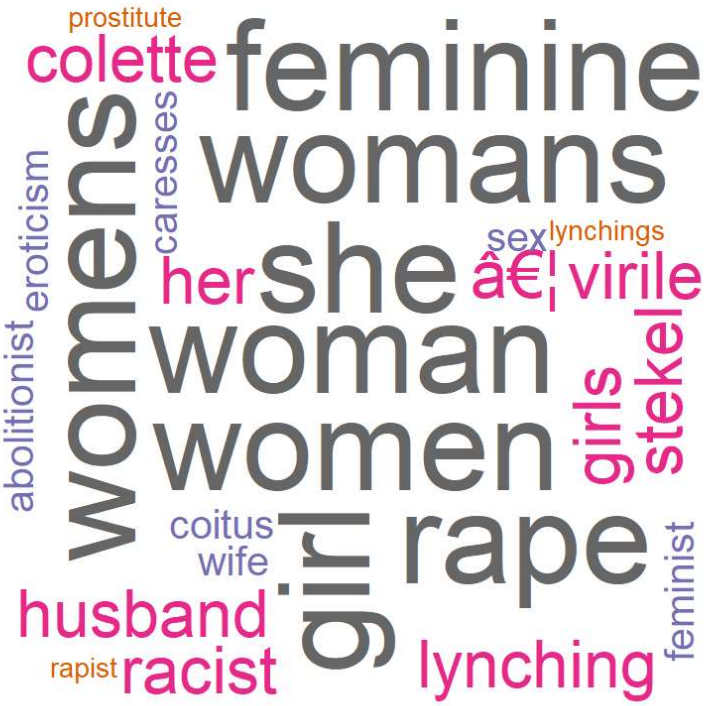
Rationalism

Stoicism

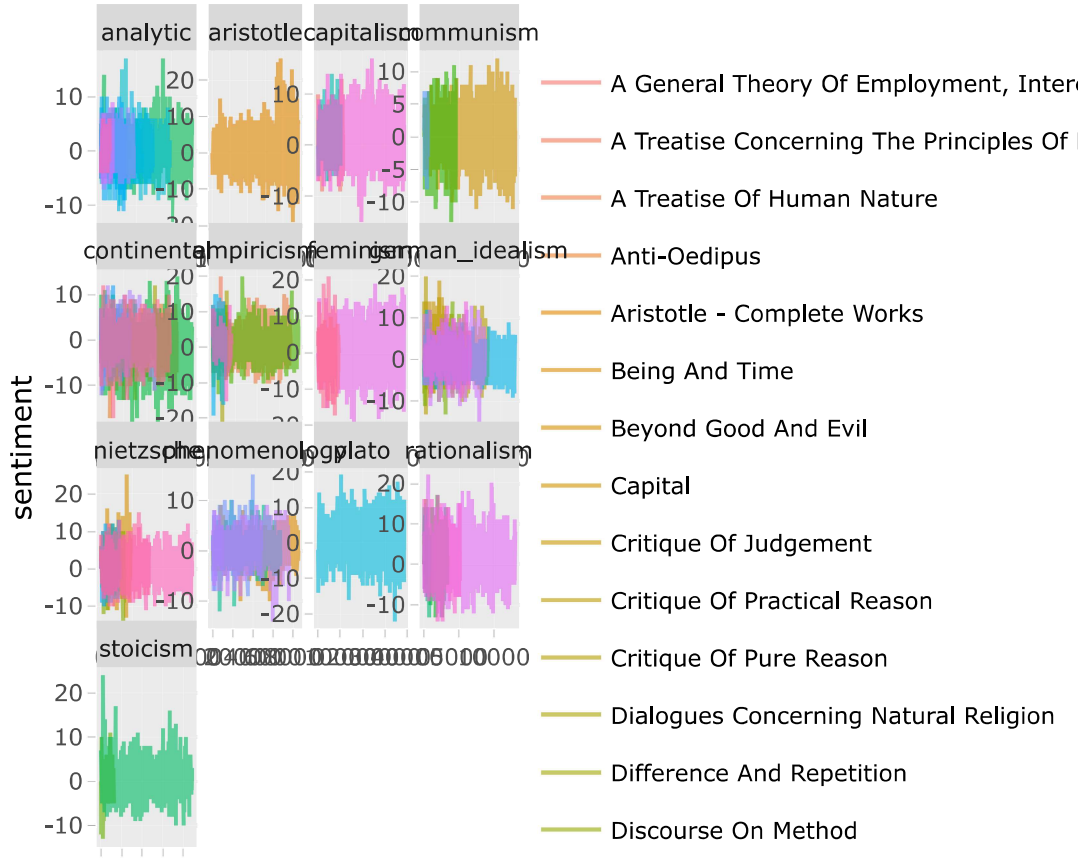


Nietzsche

Feminism



Sentences Sentiment Change



sentence_id