Análise de Sentimento das Principais Obras da Filosofia Ocidental ao Longo da História

Equipe

Nome: Pensadores de Dados

Membros: Breno Moura de Abreu [1561286] e Ricardo Alexandre de Souza [2506173]

Repositório GitLab: https://gitlab.com/breno-abreu/IntroCD

Obtenção e Processamento de Dados

A fonte de dados escolhida foi um corpus da história da filosofia ociendal encontrado neste link que contém sentenças selecionadas de 59 obras da literatura filosófica ao longo da história. O corpus foi construído pelo ex-professor de filosofia e atual cientista de dados Kurosh Alizadeh e foi atualizado pela última vez em 2021. O arquivo está no formato CSV e apenas esta fonte de dados foi utilizada na fase de análise exploratória.

A linguagem de programação escolhida foi Python através do Jupyter Notebook para realizar as análises. As bibliotecas utilizadas foram: Pandas, Matplotlib, NLTK e Empath.

A fonte de dados contém 360808 linhas e 11 colunas, sendo estas:

1. Título da obra
2. Autor
3. Escola de pensamento
4. Sentença espaçada
5. A sentença em si
6. A data original de publicação
7. A data de edição do corpus
8. O tamanho da sentença (número de caracteres)
9. A sentença transformada apenas para letras minúsculas
10. O texto tokenizado
11. O texto lemmatizado

Após o carregamento dos dados, foi criado um novo Data Frame contendo apenas as colunas 1, 2, 3, 6, 8 e 9. As seguintes colunas foram adicionadas ao novo Data Frame:

1. “century”: informa o século da data de publicação;
2. "sentence_normalized": armazena a sentença normalizada, isto é, com todas as letras transformadas para letras minúsculas, sem pontuação e apenas contendo palavras que não são stopwords;
3. “normalized_words_count”: contém a quantidade de palavras normalizadas da sentença;
4. “sentiment_score”: informa o valor do score da análise de sentimento;

Como é possível perceber na figura 1, todas as colunas possuem dados para todas as linhas, ou seja, não há dados faltantes. A fonte de dados também não apresenta linhas duplicadas. Também podemos perceber que a biblioteca Pandas automaticamente realizou o casting dos dados para os formatos corretos. As datas são classificadas como inteiros pois há apenas o ano, e não a data completa de publicação.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 360808 entries, 0 to 360807
Data columns (total 11 columns):
 #   Column                    Non-Null Count   Dtype 
---  ------                    --------------   ----- 
 0   title                     360808 non-null  object
 1   author                    360808 non-null  object
 2   school                    360808 non-null  object
 3   sentence_spacy            360808 non-null  object
 4   sentence_str              360808 non-null  object
 5   original_publication_date 360808 non-null  int64 
 6   corpus_edition_date       360808 non-null  int64 
 7   sentence_length           360808 non-null  int64 
 8   sentence_lowered          360808 non-null  object
 9   tokenized_txt             360808 non-null  object
 10  lemmatized_str            360808 non-null  object
dtypes: int64(3), object(8)
memory usage: 30.3+ MB
```

Figura 1. Informações sobre as colunas da fonte de dados

Outros Data Frames foram criados ao durante a análise contendo dados de agrupamentos que serão informados ao longo do relatório.

Cobertura e Distribuição dos Dados

A base de dados apresenta sentenças das seguintes obras dos 36 diferentes autores:

1. 'Plato - Complete Works'
2. 'Aristotle - Complete Works'
3. 'Second Treatise On Government'
4. 'Essay Concerning Human Understanding'
5. 'A Treatise Of Human Nature'
6. 'Dialogues Concerning Natural Religion'
7. 'Three Dialogues'
8. 'A Treatise Concerning The Principles Of Human Knowledge',
9. 'Ethics'
10. 'On The Improvement Of Understanding'
11. 'Theodicy'
12. ‘Discourse On Method'
13. 'Meditations On First Philosophy'
14. 'The Search After Truth'
15. 'The Analysis Of Mind'
16. 'The Problems Of Philosophy'
17. 'Philosophical Studies'
18. 'Philosophical Investigations'
19. 'Tractatus Logico-Philosophicus'
20. 'Lewis - Papers'
21. 'Quintessence'
22. 'The Logic Of Scientific Discovery'
23. 'Naming And Necessity'
24. 'Philosophical Troubles'
25. 'On Certainty'
26. 'The Birth Of The Clinic'
27. 'History Of Madness'
28. 'The Order Of Things'
29. 'Writing And Difference'
30. 'Difference And Repetition'
31. 'Anti-Oedipus'
32. 'The Phenomenology Of Perception'
33. 'The Crisis Of The European Sciences And Phenomenology'
34. 'The Idea Of Phenomenology'
35. 'Being And Time'
36. 'Off The Beaten Track'
37. 'Critique Of Practical Reason'
38. Critique Of Judgement'
39. 'Critique Of Pure Reason'
40. 'The System Of Ethics'
41. 'Science Of Logic'

42. 'The Phenomenology Of Spirit'
43. 'Elements Of The Philosophy Of Right'
44. 'Capital'
45. 'The Communist Manifesto'
46. 'Essential Works Of Lenin'
47. 'The Wealth Of Nations'
48. 'On The Principles Of Political Economy And Taxation'
49. 'A General Theory Of Employment, Interest, And Money'
50. 'Enchiridion'
51. 'Meditations'
52. 'The Antichrist'
53. 'Beyond Good And Evil'
54. 'Ecce Homo'
55. 'Twilight Of The Idols'
56. 'Thus Spake Zarathustra'
57. 'Vindication Of The Rights Of Woman'
58. 'The Second Sex'
59. 'Women, Race, And Class'

A figura 2 apresenta como as sentenças estão distribuídas por cada autor. É possível perceber que a quantidade de linhas para Aristóteles e Platão é consideravelmente maior que para os outros autores.

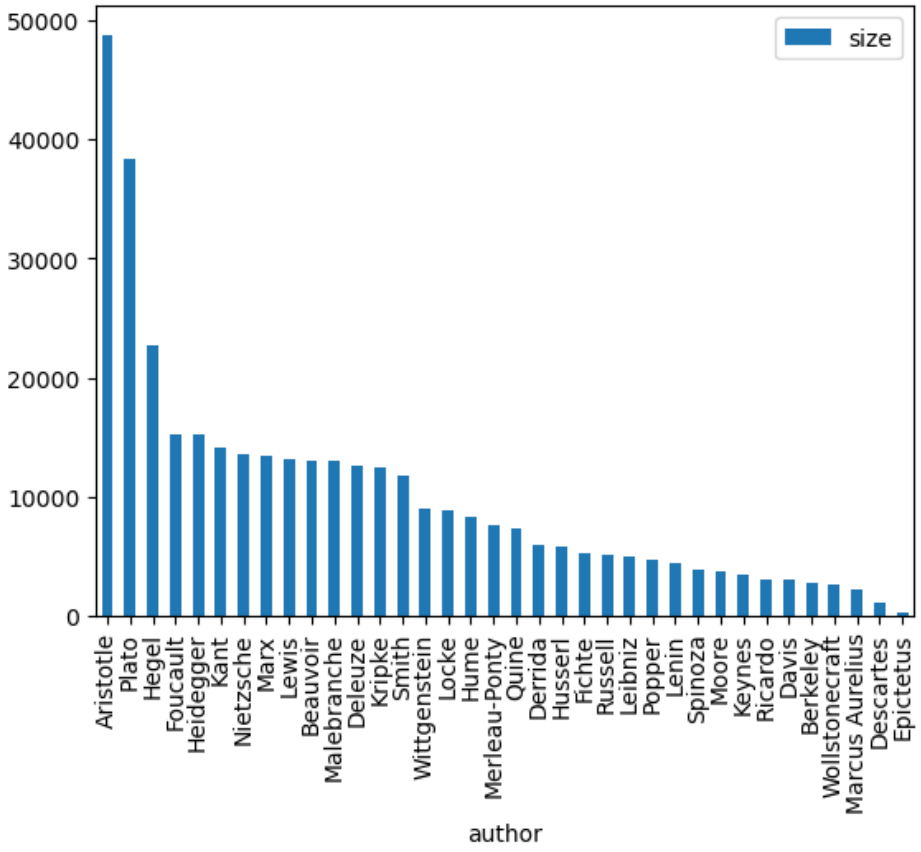

Figura 2. Quantidade de linhas por autor.

A média de sentenças por autor é 10022, sendo que o autor com menos linhas possui apenas 323 entradas, e o autor com mais linhas possui 48779 entradas. A maior parte dos autores possui entre 3761 e 13042 sentenças. A figura 3 apresenta um box plot demonstrando os dados descritos anteriormente. Os dois outliers são Platão e Aristóteles como já citado.

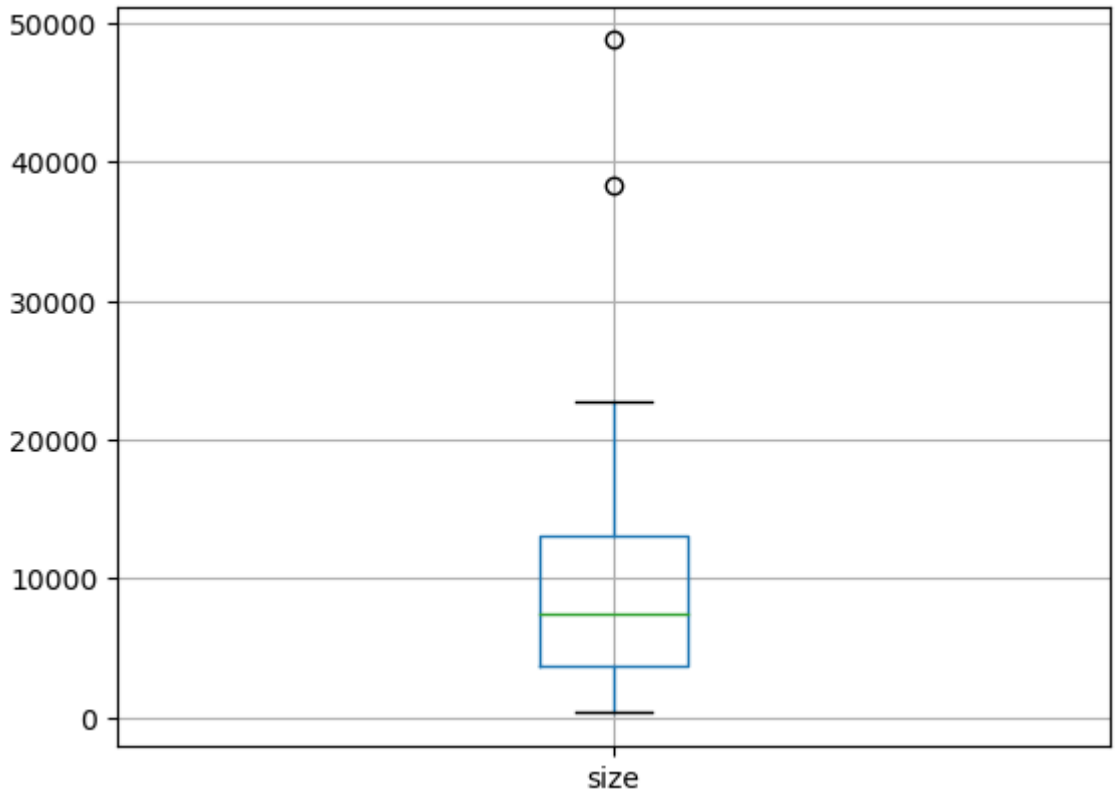

Figura 3. Box plot da distribuição de sentenças por autor.

Podemos visualizar a distribuição dos dados para as escolas de pensamento nos gráficos a seguir. A figura 4 apresenta um gráfico de barras informando a quantidade de linhas para cada escola de pensamento.

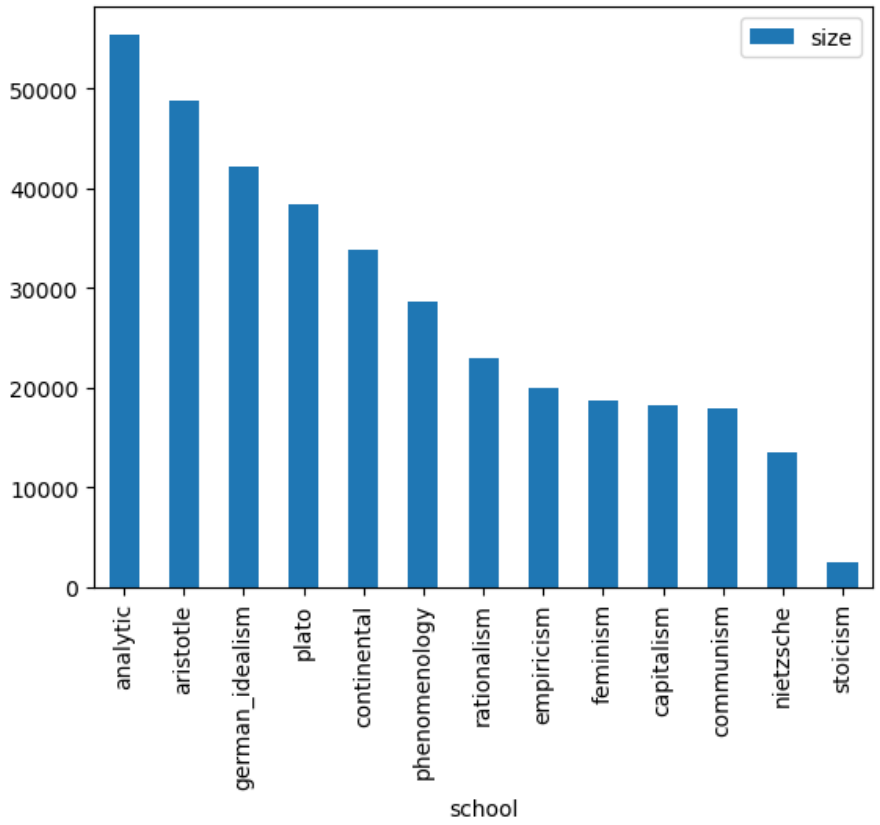

Figura 4. Distribuição de sentenças por escola de pensamento.

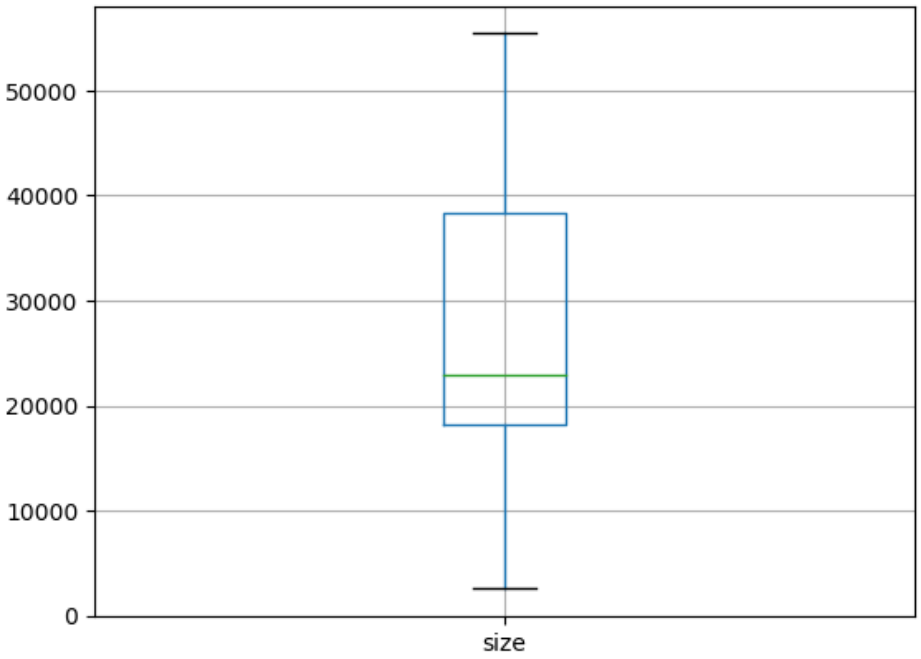

Figura 5. Box plot da distribuição de sentenças por escola de pensamento.

A média de sentenças por escola é 27754, sendo que a escola com mais linhas possui 55425 entradas e a com menos apenas 2535 entradas. A maior parte das escolas possui entre 18194 e 38366 entradas. A figura 5 permite a visualização destes dados.

A figura 6 apresenta a percentagem de linhas para cada escola de pensamento.

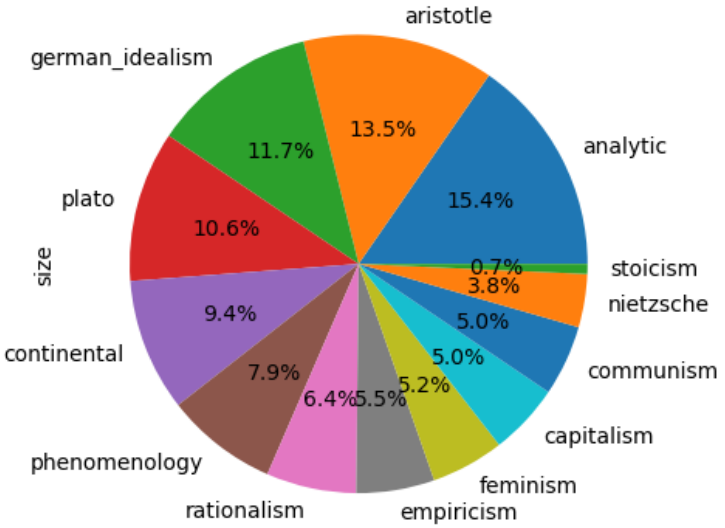

Figura 6. Porcentagem de linhas para cada escola de pensamento

Em relação ao tamanho das sentenças, a média é de 150 caracteres, sendo que a menor sentença possui apenas 20 caracteres, e a maior possui 2649 caracteres. A maior parte das sentenças possui entre 75 e 200 caracteres. A figura 7 apresenta a distribuição de caracteres para as sentenças.

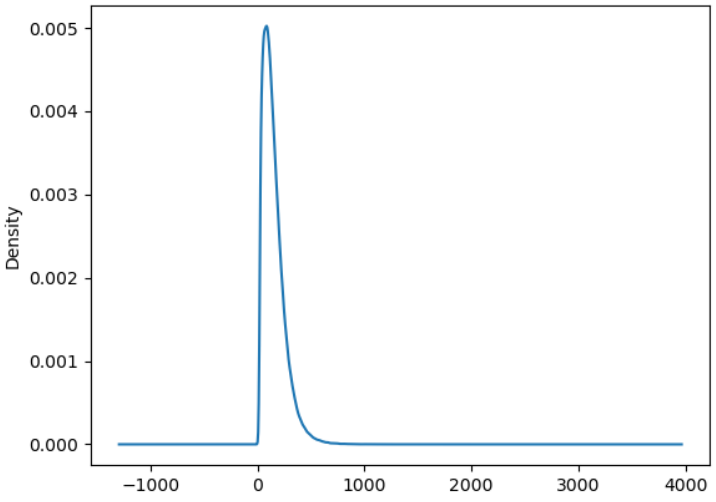

Figura 7. Densidade da quantidade de caracteres das sentenças.

Para a quantidade de palavras para a sentença tokenizada, a média é de 12 palavras, sendo que a sentença com menos palavras tem 0 palavras e a com mais possui 217. A maior parte das sentenças possui entre 6 e 16 palavras tokenizadas. A figura 8 apresenta a distribuição de palavras tokenizadas para as sentenças.

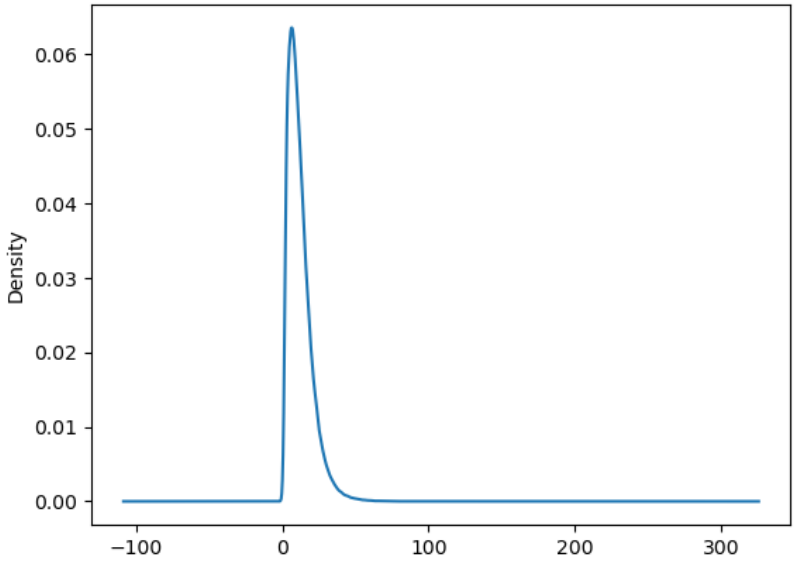

Figura 8. Distribuição de palavras tokenizadas para as sentenças.

A figura 9 apresenta a distribuição ao longo dos anos, e a figura 10 a distribuição para cada século.

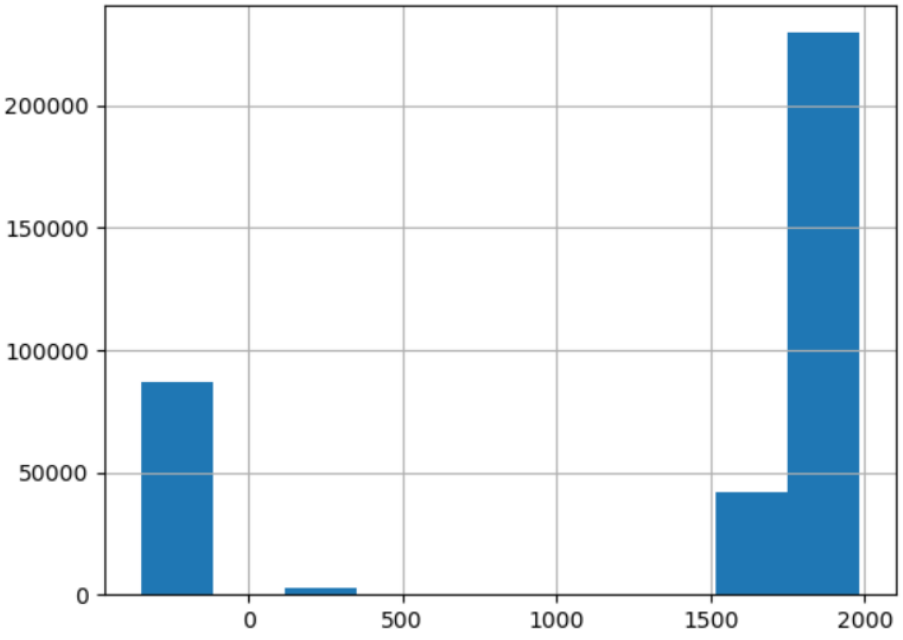

Figura 9. Distribuição das sentenças ao longo dos anos.

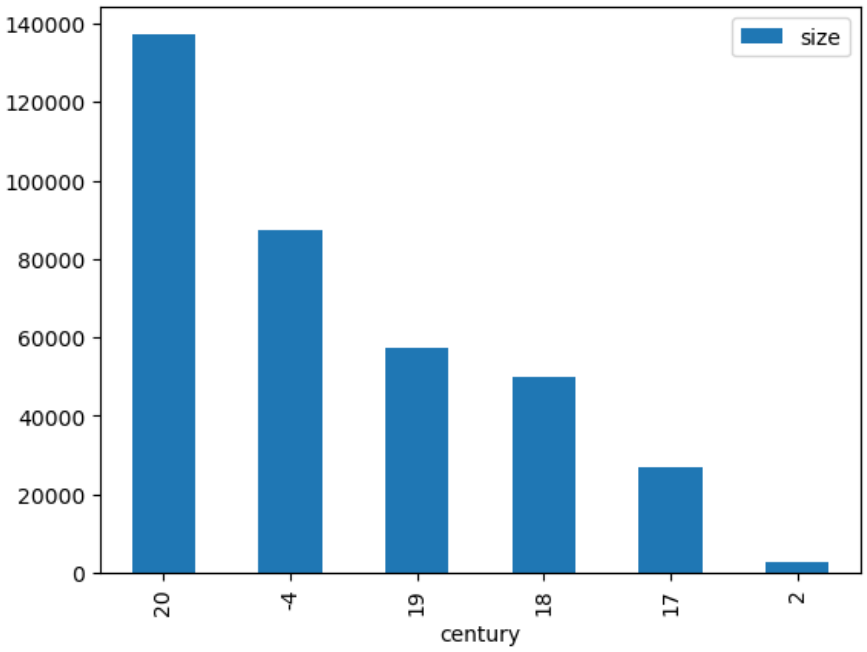

Figura 10. Quantidade de sentenças por século.

Pode-se perceber que existem mais dados relativos à obras publicadas no século XX e no século IV AEC. As demais obras estão entre o século XVII e XIX e poucos dados são relativos à obras publicadas no século II. Não há dados sobre obras do período medieval ou renascentista.

Conclui-se que os dados apresentam uma grande disparidade em relação a distribuição de sentenças para autores, escolas de pensamento e épocas diferentes. A análise não poderá ser justa dessa forma e um corpus que apresenta uma distribuição mais equilibrada poderia servir melhor para gerar análises mais confiáveis.

Análise Exploratória

Após realizar a tokenização das sentenças, normalizando-as e retirando as stopwords, foi aplicado para cada escola de pensamento a contagem de palavras. Os gráficos a seguir apresentam as palavras mais comuns para as sentenças de cada escola de pensamento.

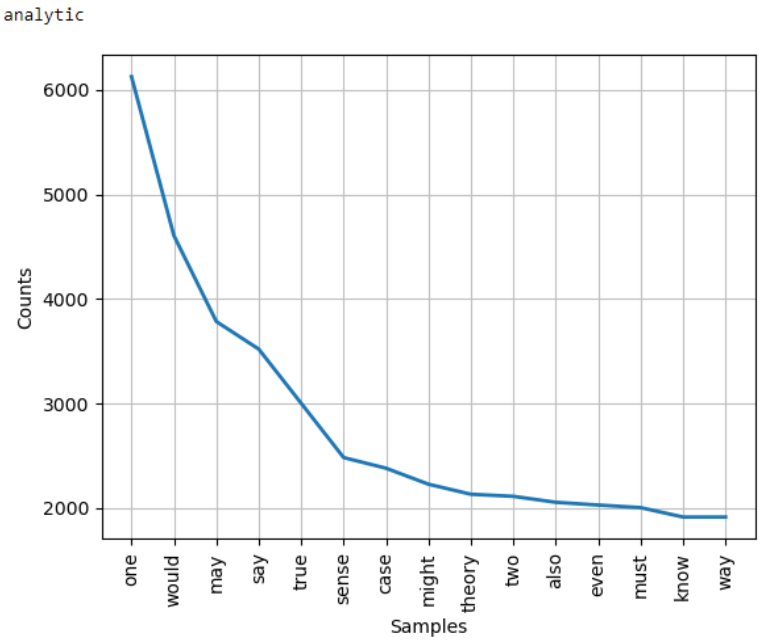

Figura 11. Palavras mais comuns para a escola Analítica

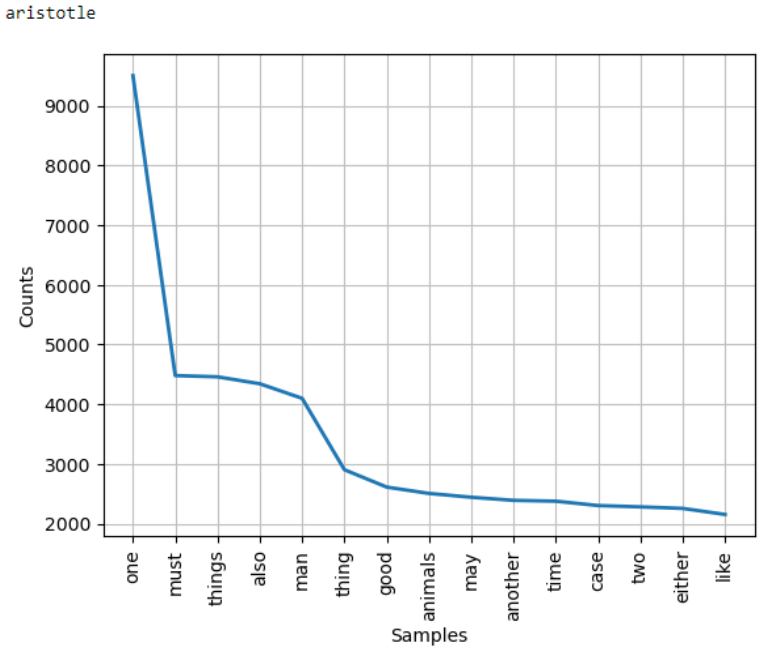

Figura 12. Palavras mais comuns para a escola Aristóteles

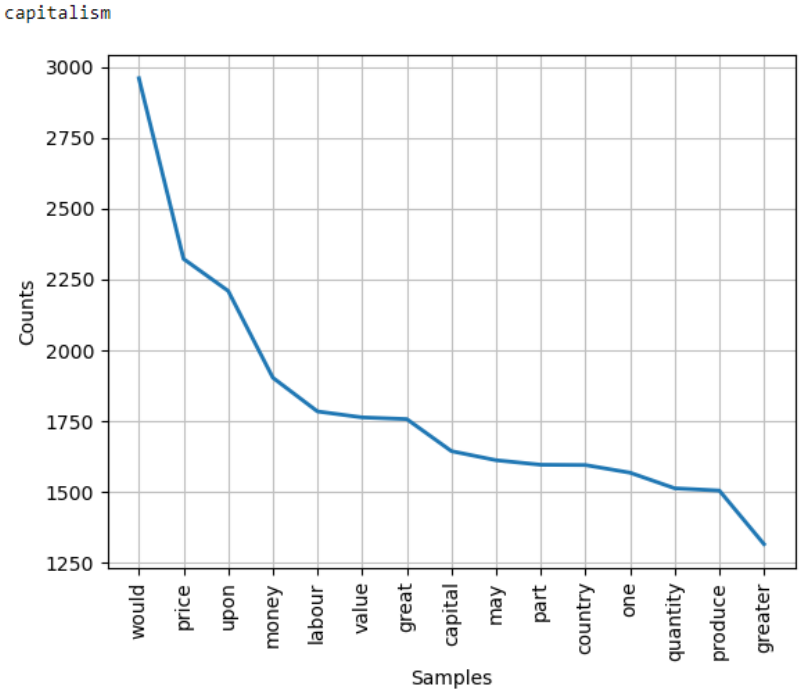

Figura 13. Palavras mais comuns para a escola Capitalismo

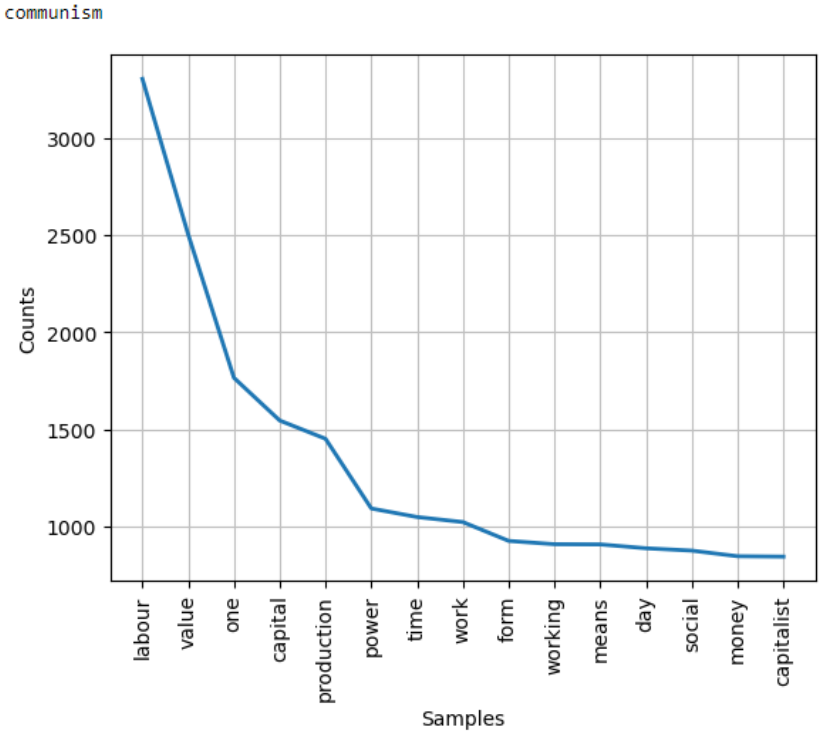

Figura 14. Palavras mais comuns para a escola Comunismo

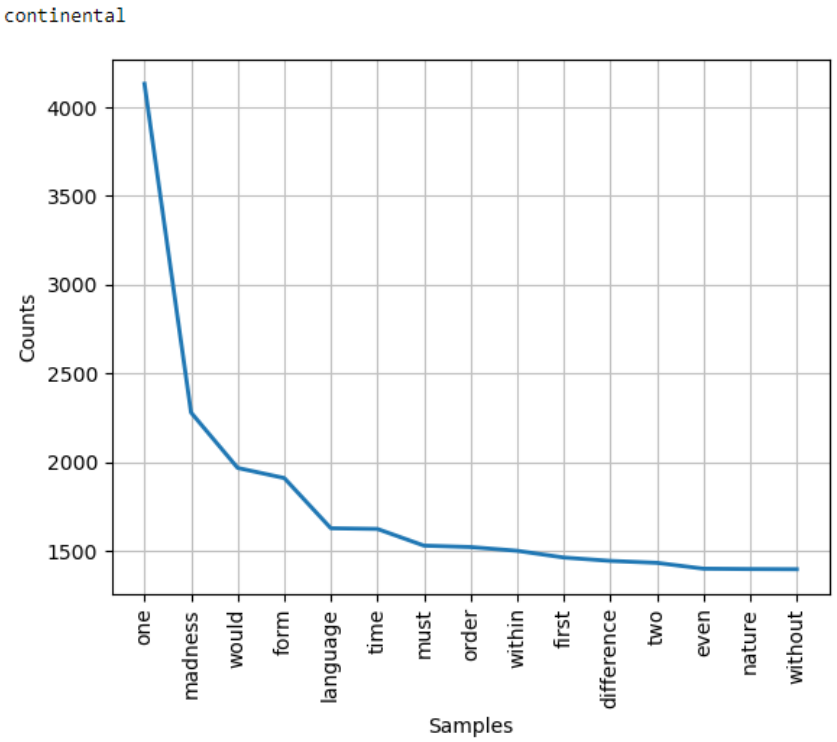

Figura 15. Palavras mais comuns para a escola Continental

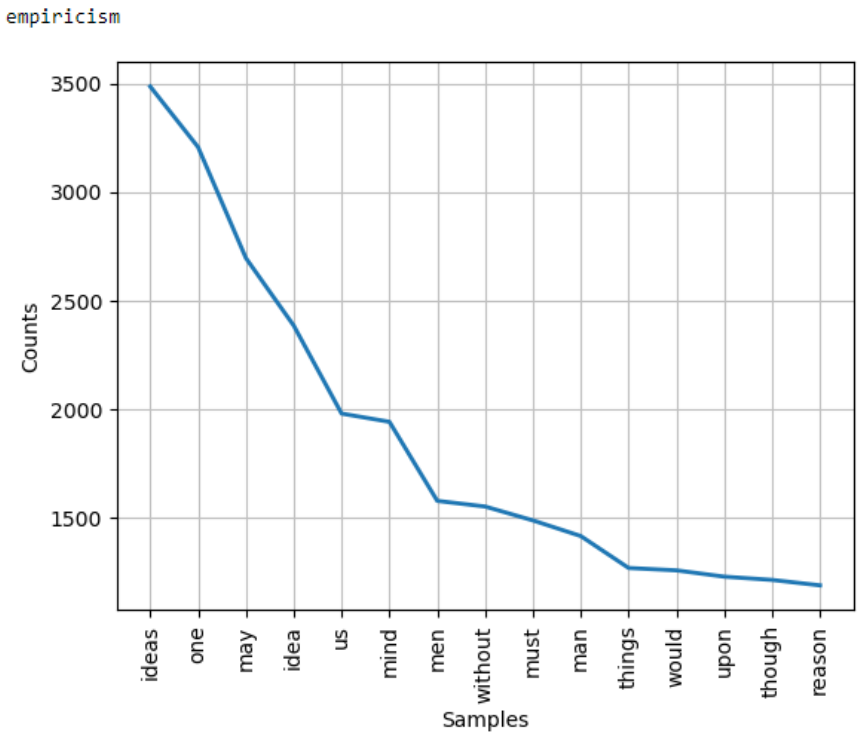

Figura 16. Palavras mais comuns para a escola Empirismo

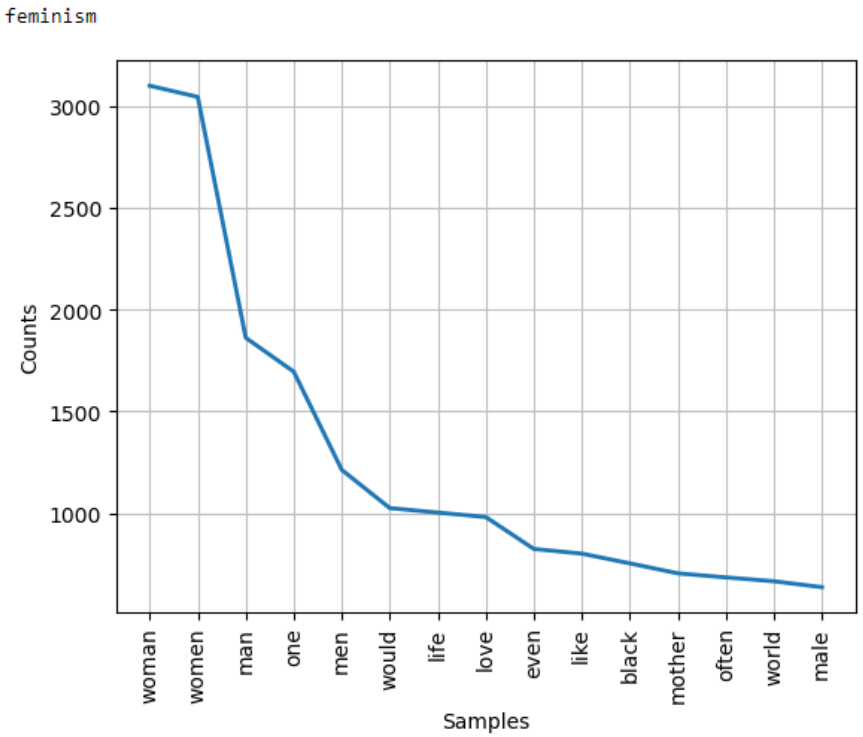

Figura 17. Palavras mais comuns para a escola Feminismo

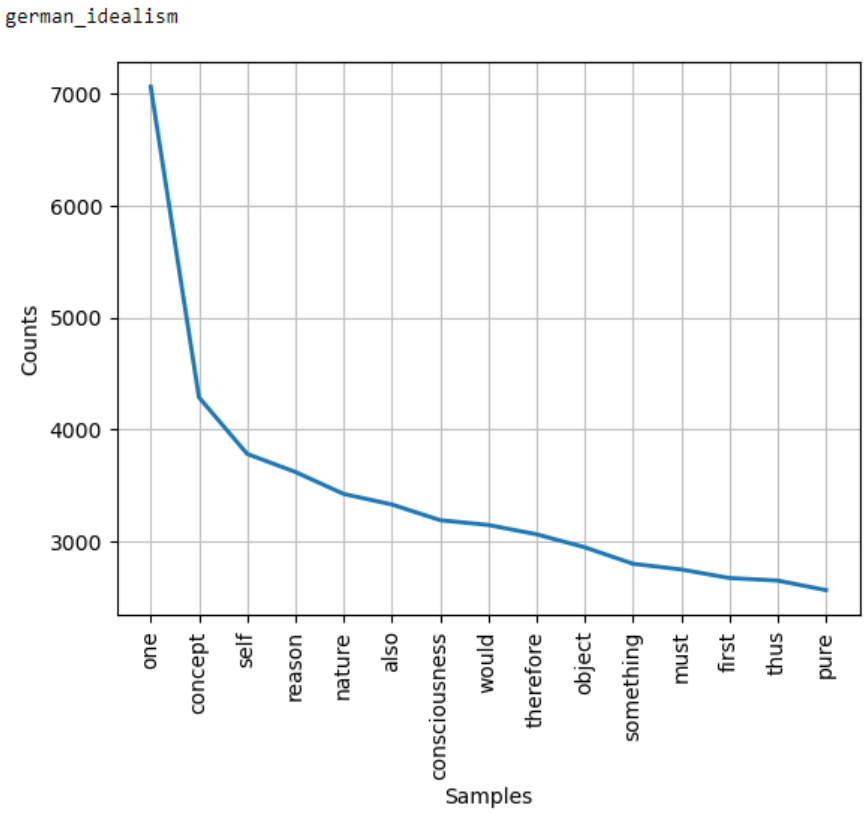

Figura 18. Palavras mais comuns para a escola Idealismo Alemão

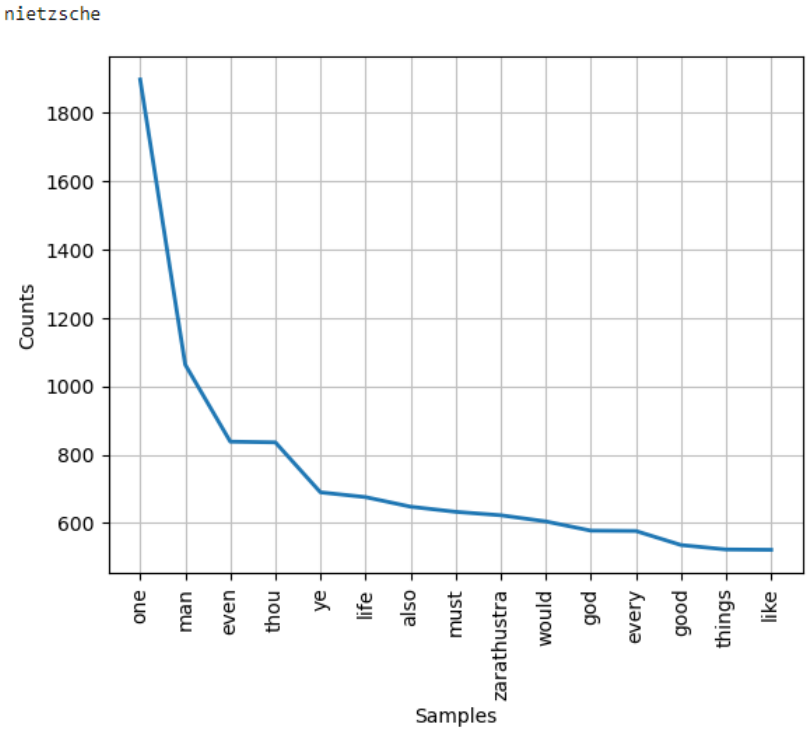

Figura 19. Palavras mais comuns para a escola Nietzsche

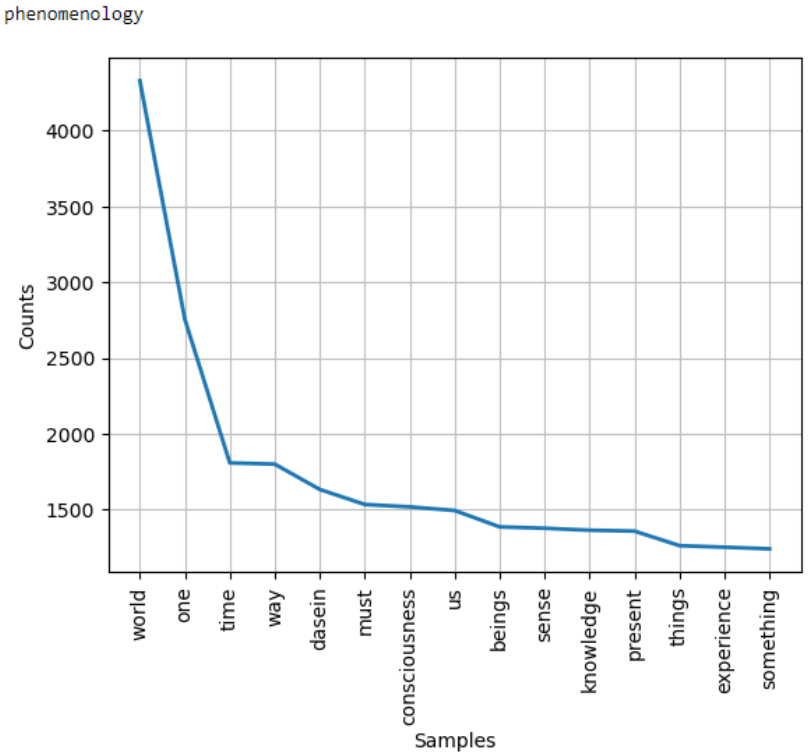

Figura 20. Palavras mais comuns para a escola Fenomenologia

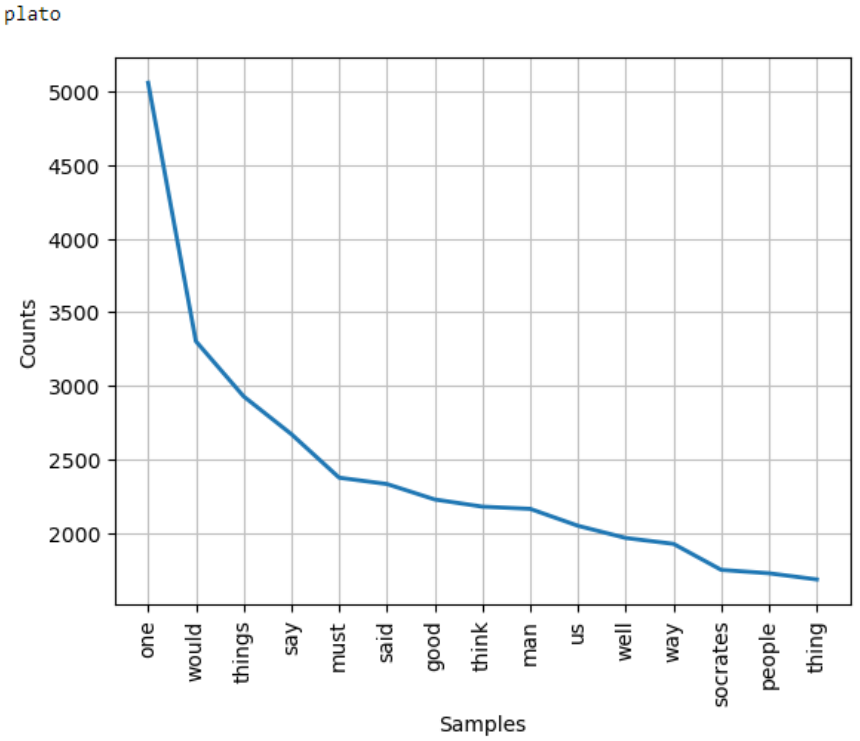

Figura 21. Palavras mais comuns para a escola Platão

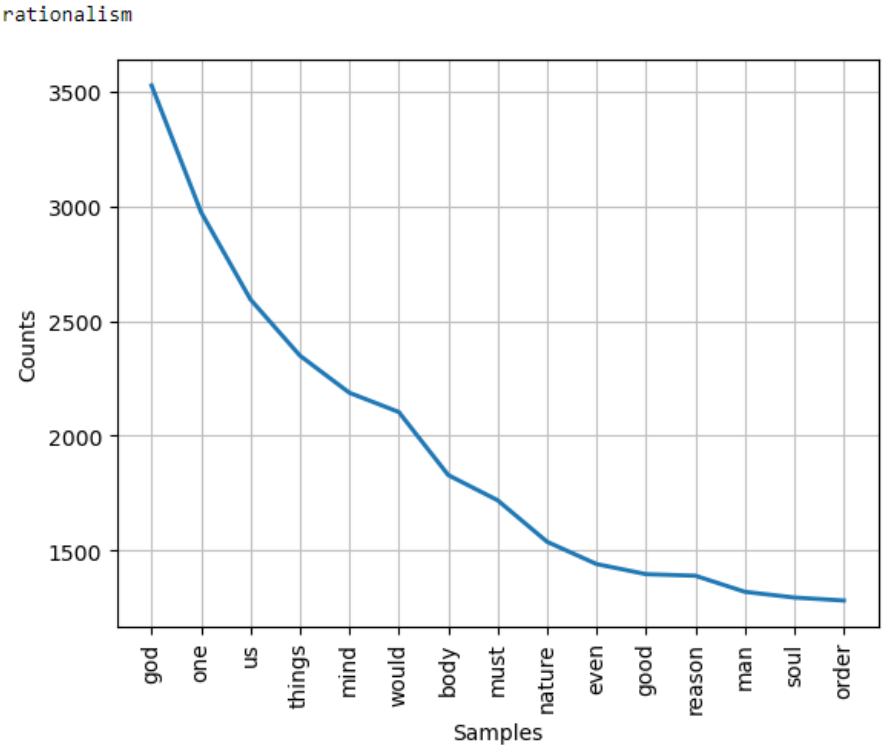

Figura 22. Palavras mais comuns para a escola Racionalismo

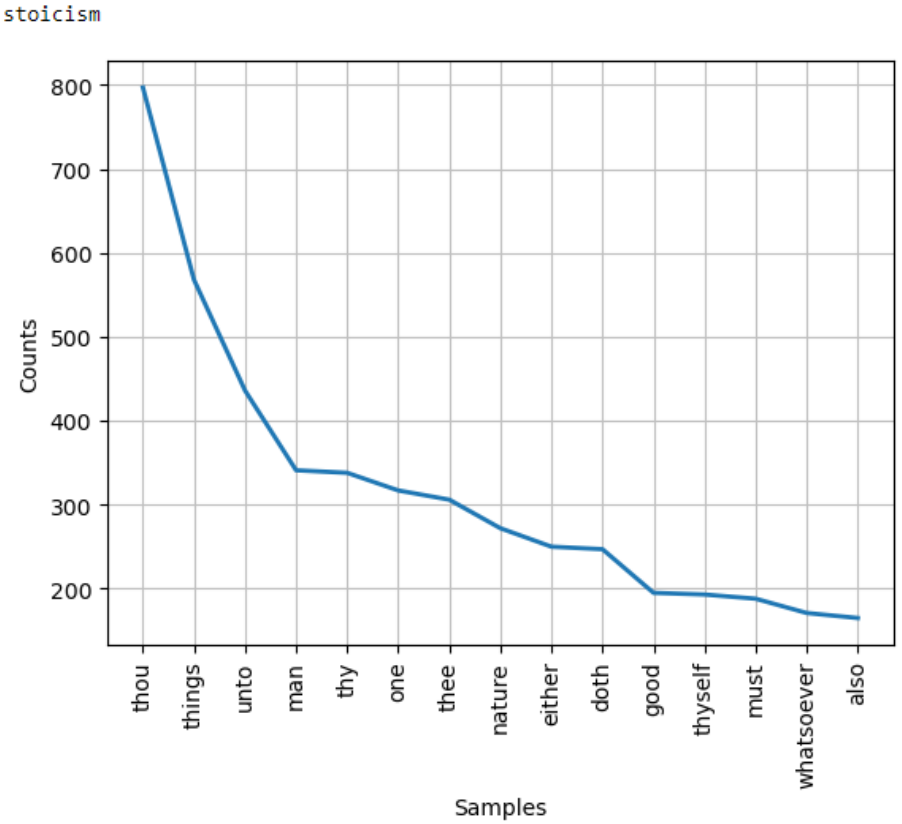

Figura 23. Palavras mais comuns para a escola Estoicismo

Um problema apresentado nas análises é que certos autores utilizam uma linguagem mais antiga, como para o estoicismo e Nietzsche, e apresentam certas stopwords (como thou, thy, thee) que deveriam ser retiradas para permitir uma análise melhor com apenas os termos relevantes.

Outra análise realizada foi a Análise de Sentimento para todas as sentenças do texto, permitindo observar como se os sentimentos das sentenças estão distribuídas. A figura 24 apresenta o gráfico da análise realizada. Números positivos representam sentenças com sentimento positivo, de modo contrário, números negativos representam sentimentos negativos. Valores próximos de 0 identificam sentenças neutras.

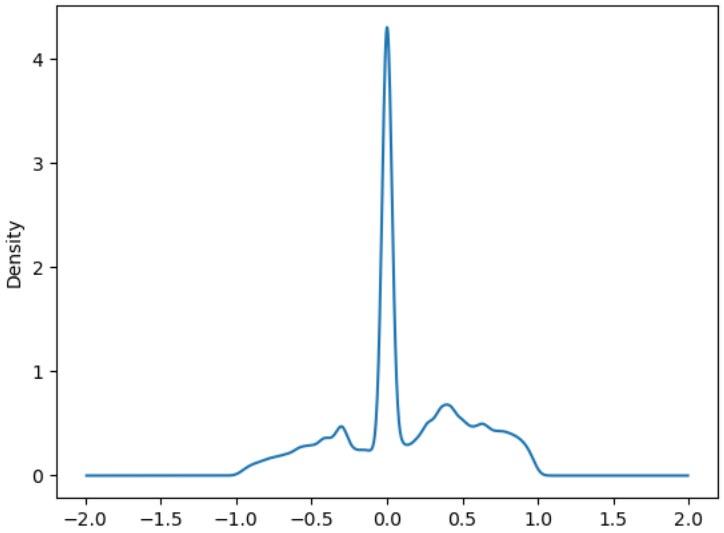

Figura 24. Análise de sentimento para todas as sentenças.

Percebe-se pelo gráfico que a maior parte das sentenças apresentam sentimentos neutros. Para as demais sentenças, há mais textos com sentimentos positivos que negativos.

A última análise realizada foi a de categorização léxica das sentenças. A ferramenta utilizada foi a biblioteca Empath que indica quais as categorias que um determinado texto pertence. Um texto pode pertencer a diferentes categorias e com diferentes graus de pertencimento dependendo das palavras utilizadas e como elas estão distribuídas. A figura 25 apresenta um gráfico com as 20 principais categorias para todas as sentenças, sem distinção de grupo.

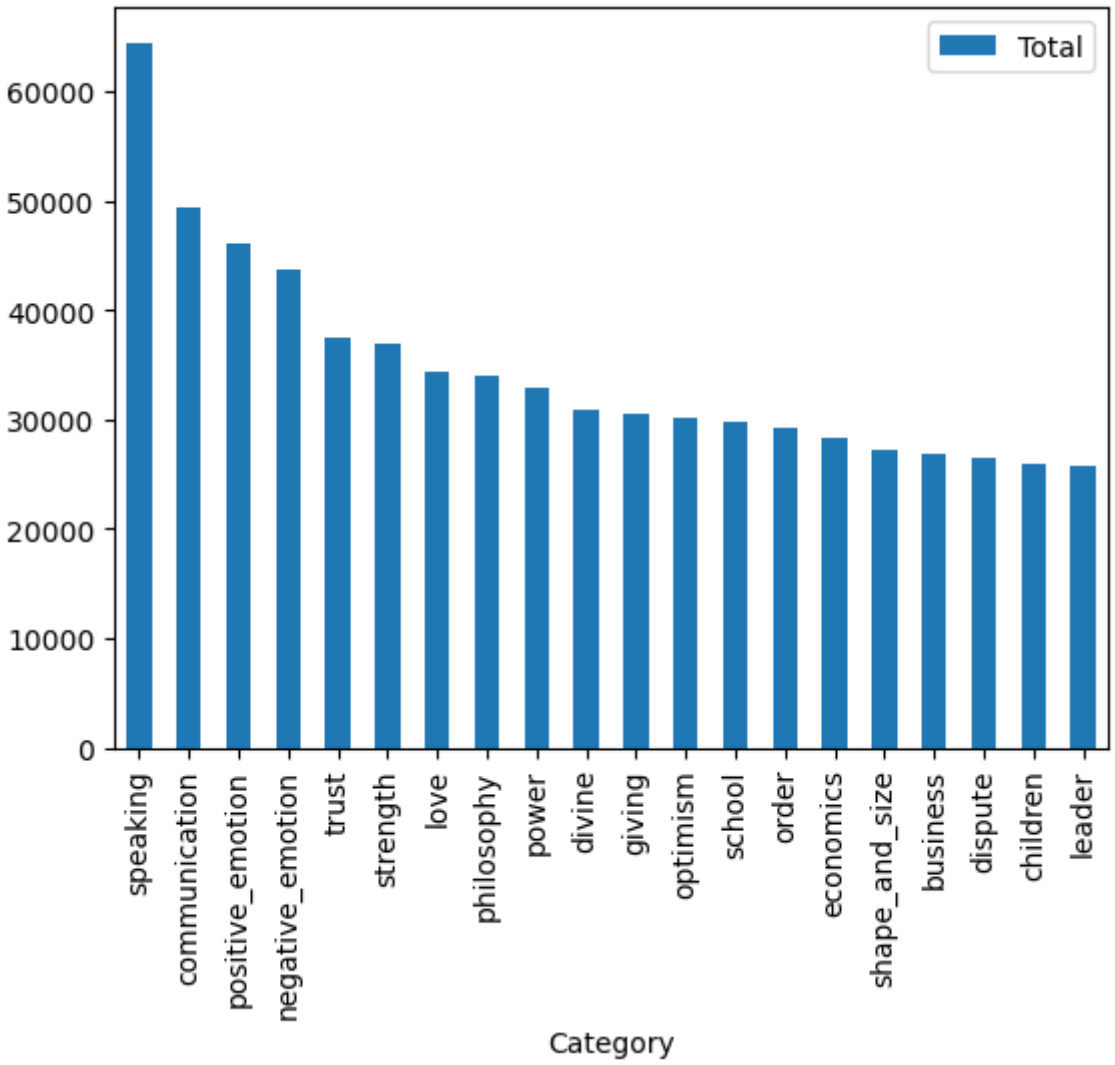

Figura 25. 20 principais categorias léxicas para toda a base de dados.

Assim como indicado pela análise de sentimentos, há mais sentenças que apresentam emoções positivas que negativas.

Discussão e Próximos Passos

O objetivo da pesquisa é conseguir compreender as diferenças linguísticas entre as escolas de pensamento e apontar as semelhanças e diferenças entre os autores e suas obras. Além disso, a pesquisa também tem o propósito de analisar as obras agrupando-as em diferentes categorias, como por escola de pensamento, século em que foram publicadas e por semelhanças na linguagem e tema apresentado.

A análise exploratória permitiu visualizar os tipos de agrupamento mais prováveis de serem analisados ao longo do trabalho. Inicialmente percebe-se a classificação por escola de pensamento, dado presente na base de dados original. Outro agrupamento é pelo século quando as obras foram publicadas. Infelizmente os dados não abordam todos os séculos da história da filosofia, como por exemplo, percebe-se que obras de período medieval e renascentista não estão presentes na base de dados. Além disso, a distribuição de obras para cada século e escola de pensamento não é equilibrada, tendo algumas escolas e séculos muito mais dados disponíveis que para outros, como é o exemplo das obras do século IV AEC e II EC. A disparidade entre a quantidade de sentenças não permitirá uma análise justa para os textos disponíveis. Talvez seja necessário retirar aleatoriamente sentenças de certos autores para permitir que a distribuição de dados seja mais equilibrada.

A análise temporal de transformação dos sentimentos e características linguísticas também não poderá ser realizada para toda a história da filosofia dado que há vacâncias entre os períodos apresentados na base de dados. Como já citado, há uma falta de obras e autores medievais e renascentistas, o que não permitirá a análise desde a antiguidade. Uma possível alteração nos objetivos de pesquisa seja apenas analisar temporalmente as obras entre os séculos XVII e XX para verificar se há transformações linguísticas que se dão ao longo apenas dessa faixa de tempo.

A contagem de palavras também será um desafio para algumas escolas como o Estoicismo e Nietzsche que utilizam stopwords que não estão na lista comum da biblioteca pois são palavras utilizadas em textos mais antigos apenas. Seria importante acrescentar as palavras que devem ser retiradas da análise em uma fase posterior. Ainda assim, a contagem de palavras irá permitir visualizar os termos mais comuns utilizados em cada grupo.

Outra forma importante de agrupamento e análise é utilizando a biblioteca Empath que permite realizar a classificação léxica dos textos e encontrar categorias e emoções mais encontradas por tipo de grupo.

A última técnica importante para o trabalho é a análise de sentimento, que será analisada posteriormente de acordo com o século de publicação e para cada outro agrupamento relevante. Todas essas técnicas irão permitir criar um perfil para cada grupo apontando mais claramente as semelhanças e diferenças entre eles.