

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

Equipe

Nome: Pensadores de Dados

Membros: Breno Moura de Abreu [1561286]

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

Introdução

A partir da análise dos tópicos e dos sentimentos mais encontrados em diferentes textos filosóficos, pode-se analisar as relações entre autores e escolas de pensamento a fim de facilitar a compreensão e ajudar a entender as diferentes abordagens encontradas na filosofia. Este trabalho tem o objetivo de realizar uma análise de sentimento das principais obras da filosofia ocidental permitindo tirar conclusões sobre as escolas a partir dos tipos de temas e padrões de linguagem utilizados em seus textos. Para tal, serão respondidas as seguintes perguntas:

1. Qual o sentimento predominante em cada escola de pensamento?
2. Quais são as conexões mais fortes entre cada escola de pensamento a partir de seus temas mais comuns?
3. Se separarmos as escolas em N grupos, como os grupos estariam separados?

Processamento de Dados

A base de dados utilizado para a pesquisa, [History of Philosophy](#), foi criada pelo ex-professor de filosofia e atual cientista de dados Kourosh Alizadeh, e apresenta sentenças de 59 obras da filosofia ocidental divididas em 36 autores e 13 escolas de pensamento. A base contém mais de 360.000 linhas e informa a obra, autor, escola de pensamento e ano de publicação de cada sentença.

As seguintes etapas foram realizadas para o processamento dos dados:

1. Remoção das colunas desnecessárias.
2. Transformação dos textos para letras minúsculas.
3. Remoção de duplicatas.
4. Adição da coluna que indica o século de cada sentença
5. Criação de uma nova coluna contendo a sentença normalizada, isto é, sem a presença de *stop words* que interferem na análise.
6. Criação de uma coluna que indica o *score* de sentimento de cada sentença, variando entre -1 e 1.
7. Criação de uma coluna que indica a quantidade de palavras de cada sentença.

8. Criação de uma coluna indicando todas as categorias lexicais e sua quantidade de palavras relacionadas em cada sentença. Esta coluna se apresenta no formato de um dicionário.
9. Contagem das palavras mais presentes para sentenças agrupadas em escola de pensamento.
10. Cálculo da média do *score* de sentimento agrupado por escola de pensamento.
11. Criação de um dicionário que indica todas as categorias lexicais e sua pontuação agrupadas por escola de pensamento.
12. Criação de um novo *data frame* onde cada linha indica uma escola de pensamento e cada coluna uma categoria lexical. Cada célula indica o valor de uma categoria lexical normalizado pelo número total de palavras da respectiva escola.
13. Criação de um novo *data frame* contendo o valor da distância euclidiana entre cada escola com base nos valores das categorias lexicais.
14. Criação de uma nova coluna no *data frame* principal agrupado por escola para indicar a qual agrupamento uma escola pertence com base nos resultados da aplicação do modelo K-means

Resultados

Sentimento Predominante por Escola

Pergunta: Qual o sentimento predominante em cada escola de pensamento?

Hipóteses:

- a. A escola continental possui um sentimento médio mais positivo que a da escola analítica.
- b. Nietzsche apresenta uma média de sentimento negativo.

O modelo utilizado para encontrar o *score* de sentimento para cada escola foi o Sentiment Intensity Analyzer (SIA) que é um modelo treinado para analisar a presença de palavras de cunho positivo ou negativo, intensificadores e o contexto onde as palavras aparecem em uma sentença para indicar uma pontuação relativa ao sentimento apresentado pelo texto. Este *score* é um número na faixa entre -1 e 1 onde -1 indica um sentimento completamente negativo, 1 um sentimento completamente positivo e 0 um sentimento neutro. Para esta pesquisa foi calculada a pontuação de sentimento para cada sentença e então foi feito o cálculo da média dos *scores* por escola.

Os resultados podem ser visualizados na Figura 1 onde é possível perceber que a escola continental, oposto à primeira hipótese, apresenta um *score* de sentimento abaixo das demais escolas e é, inclusive, a única que apresentou um *score* negativo, por mais que próximo do sentimento neutro. Conclui-se também que a segunda hipótese é falsa pois os sentimentos médios encontrados em Nietzsche são positivos.

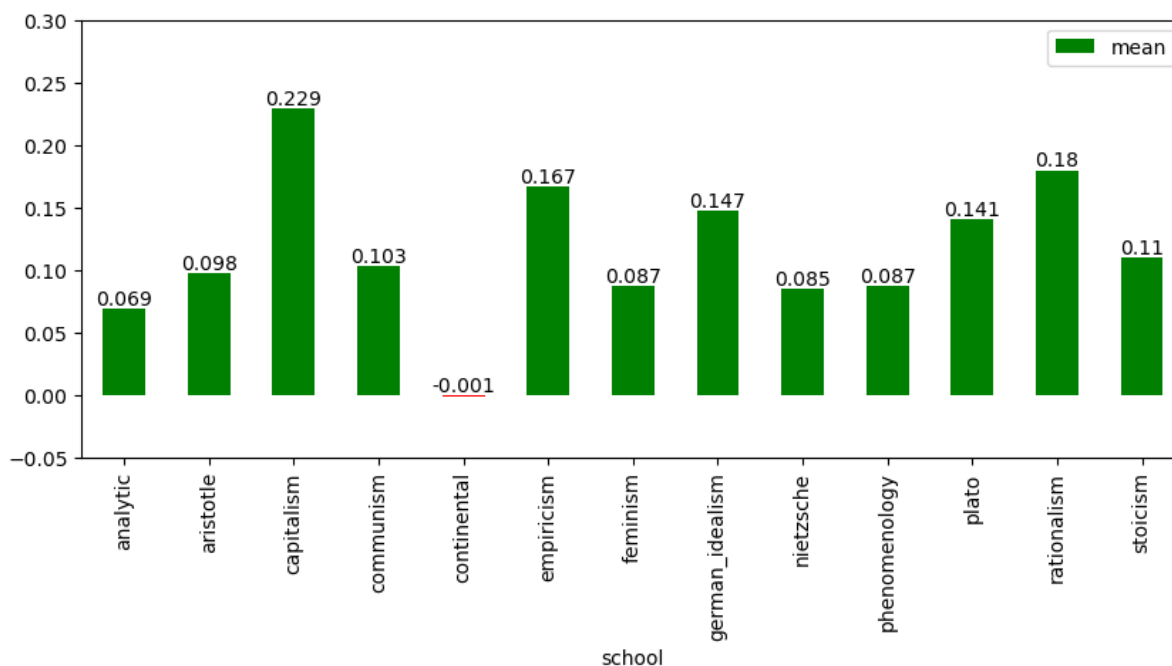


Figura 1. Média do score de sentimento para cada escola

Análise de Categorias Lexicais por Escola

Pergunta: Quais são as conexões mais fortes entre as escolas de acordo com os temas apresentados?

Hipótese: As conexões mais fortes serão com as escolas de pensamento imediatamente anteriores cronologicamente à escola em análise.

O modelo Empath foi treinado utilizando redes neurais e é utilizado para analisar uma sentença apresentando o número de palavras pertencentes a uma das 200 categorias lexicais presentes no modelo como *trust*, *power*, *negative_emotion* e *love*. O modelo é similar ao LWIC mas apresenta uma quantidade maior de categorias que permitem determinar os temas e sentimentos predominantes em um texto. O resultado final é um dicionário que indica uma pontuação, normalizada ou não, para cada categoria.

Para este trabalho, foi gerado um dicionário contendo os valores totais de palavras para cada categoria para cada escola, e então estes valores foram normalizados em relação ao número total de palavras das sentenças de cada escola. A partir desses resultados, um novo *data frame* contendo 130 diferentes categorias foi criado e o valor para cada coluna de cada escola foi anotado. A partir dessa vetorização, a distância entre as escolas foi calculada. O resultado para a escola mais próxima entre apenas escolas cronologicamente anteriores pode ser visualizado na Figura 2. Neste caso, pergunta-se: esta escola compartilha temas e sentimentos com qual escola de um período anterior?

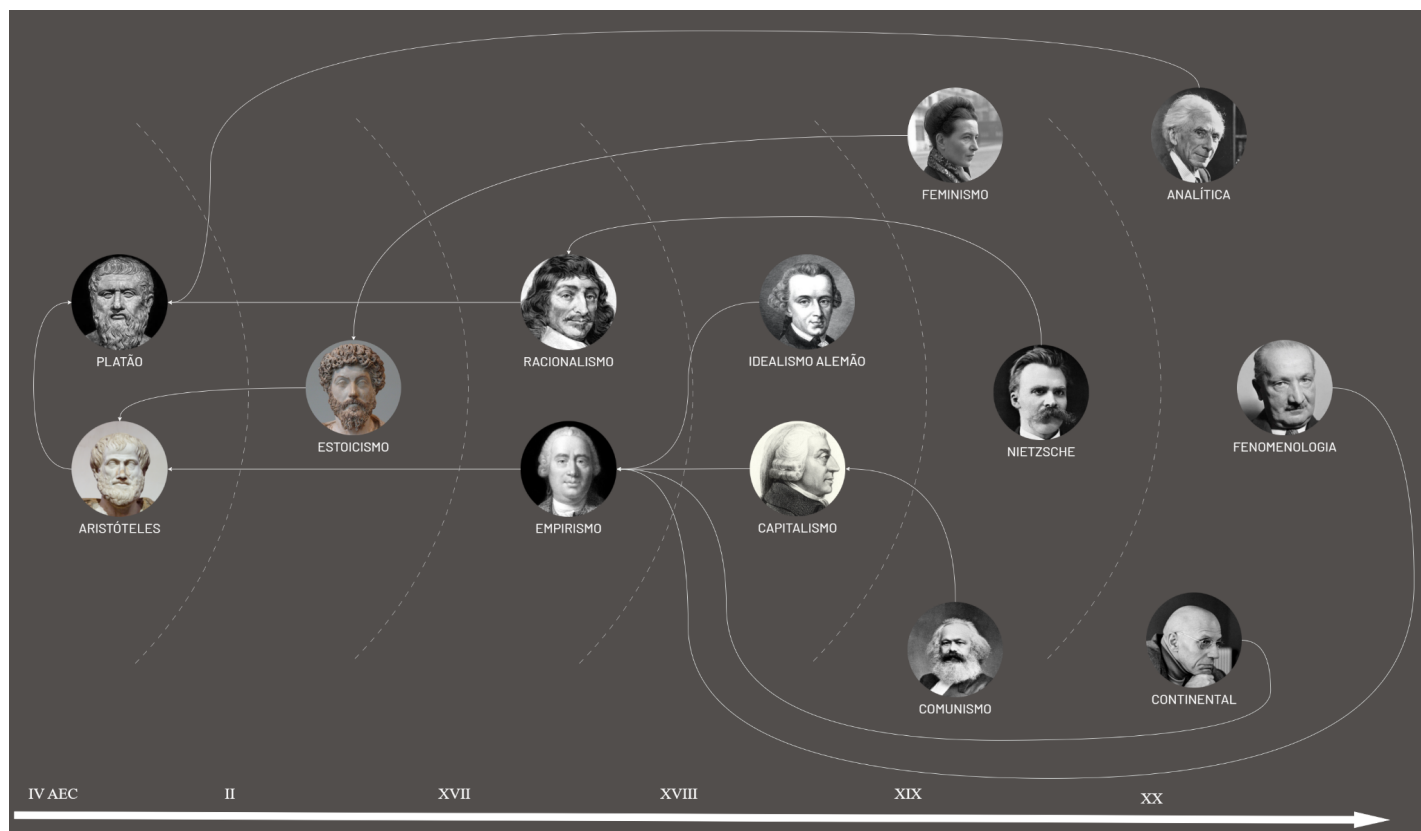


Figura 2. Ligação mais forte com escola de períodos anteriores.

Como é possível observar, a hipótese neste caso é falsa pois não há um padrão entre as conexões em relação a períodos anteriores. Por exemplo, escolas do século XX têm uma relação maior com as escolas do século XVII e IV AEC que do seu século anterior.

Apenas a nível de curiosidade, a Figura 3 apresenta as conexões mais fortes entre as escolas sem levar em consideração o fator temporal. Neste caso, pergunta-se: para esta escola, qual é a outra escola mais próxima em relação a temas e sentimentos sem levar a cronologia em consideração?

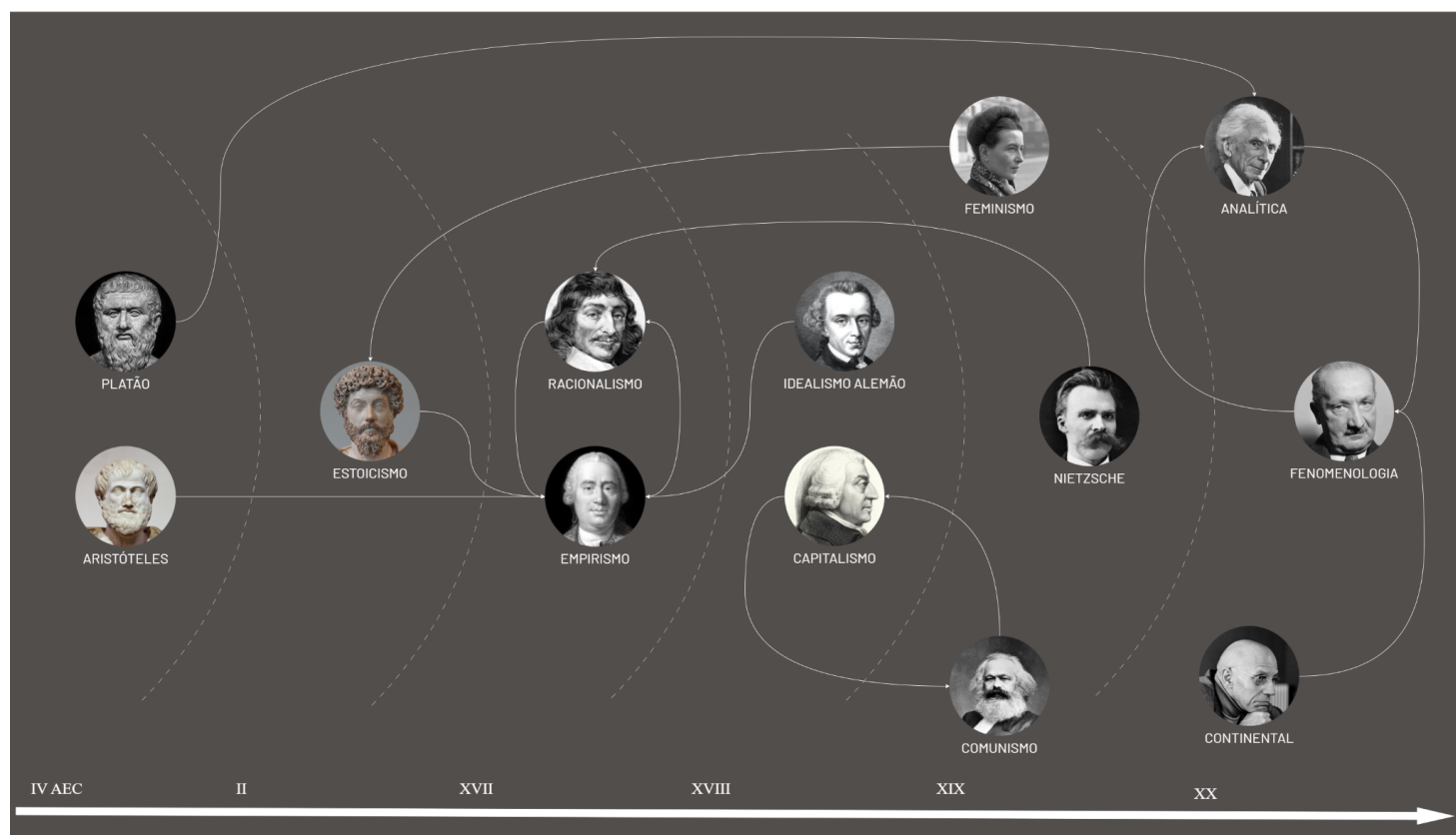


Figura 3. Ligação mais forte com escolas de qualquer período.

Agrupamentos a Partir das Categorias Lexicais

Pergunta: Se separarmos as escolas em N grupos, como os grupos estariam separados?

Hipótese: Se agruparmos as escolas de acordo com o número de séculos presente na base de dados, os grupos formados serão separados por século.

Para realizar o agrupamento das escolas, foi utilizado o modelo K-means que possibilita agrupar as instâncias em K grupos. Com a vetorização dos dados realizada a partir da categorização lexical, pode-se aplicar o modelo para agrupar as escolas. A primeira análise foi realizada com a criação de 6 grupos, possibilitando verificar se as escolas estariam agrupadas corretamente nos 6 séculos distribuídos na base. A Figura 4 demonstra que a hipótese não é verdadeira pois os agrupamentos não refletem a divisão por século. Inclusive, escolas do século XX foram agrupadas com as do século IV AEC, os dois séculos mais distantes encontrados na base.

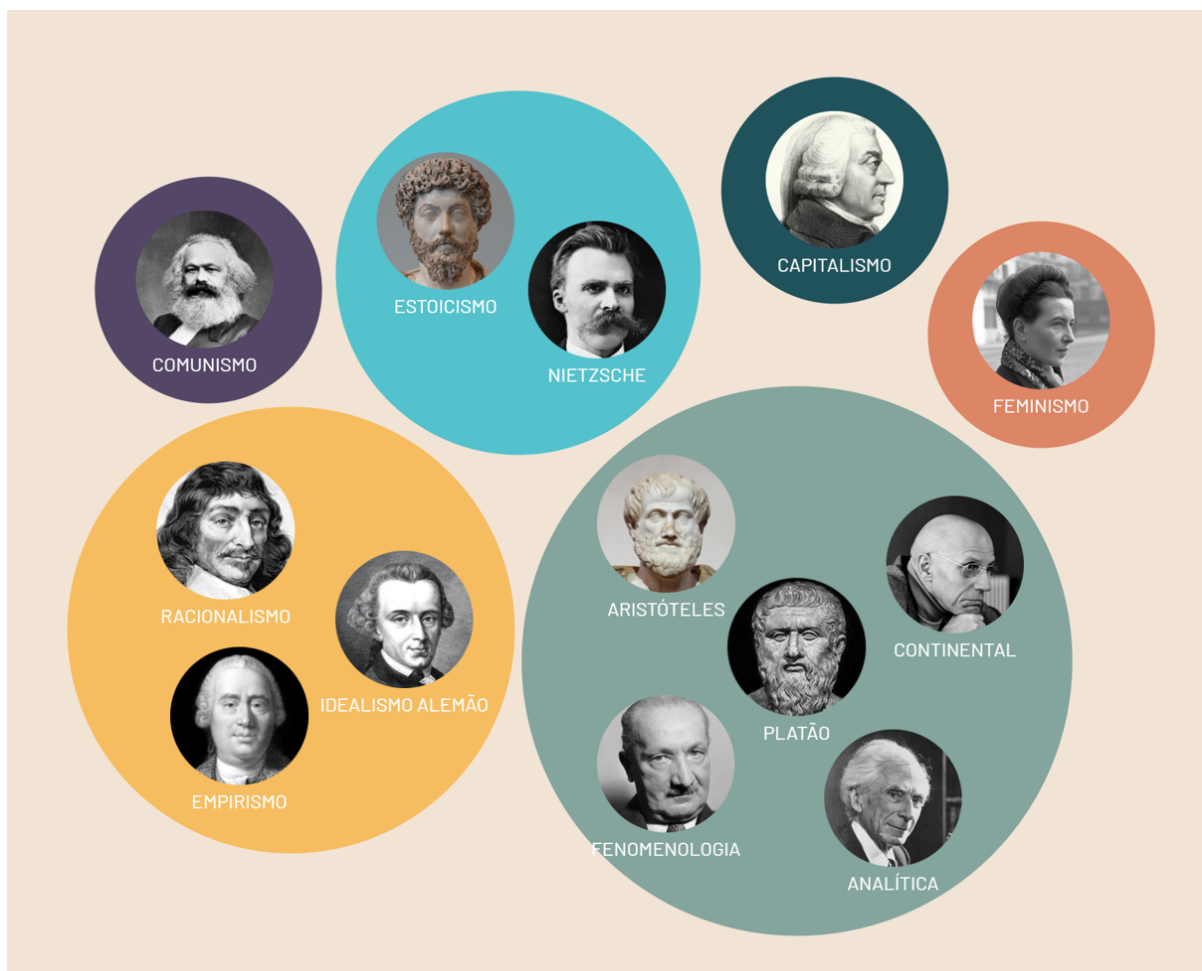


Figura 4. Agrupamento de escolas em 6 grupos.

Para encontrar o melhor número de grupos para as escolas foi criado o dendrograma da Figura 5, que indica que a melhor forma de clusterizar as instâncias é em dois grupos distintos.

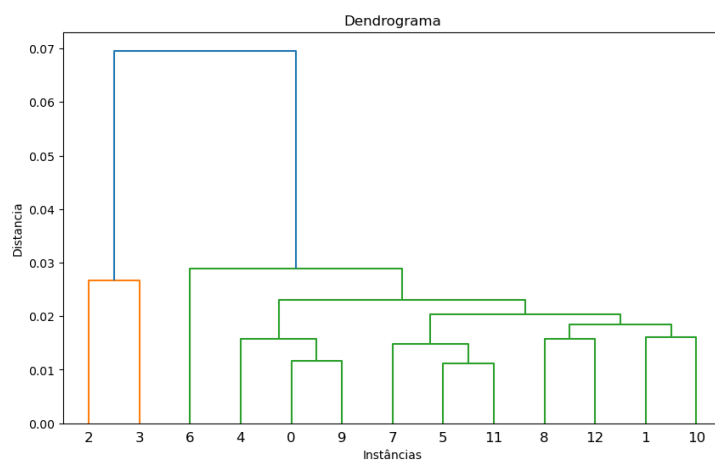


Figura 5. Dendrograma das distâncias entre as escolas.

A partir dessa informação, foi realizado um novo agrupamento apresentado na Figura 6 que separa as escolas em dois grupos distintos, onde um deles contém as escolas Capitalismo e Comunismo e o outro as demais escolas. Uma possibilidade desta divisão está no fato que as duas escolas do primeiro grupo são fortemente ligadas à economia e possuem temas mais ligados a dinheiro, trabalho e produção que as demais escolas.

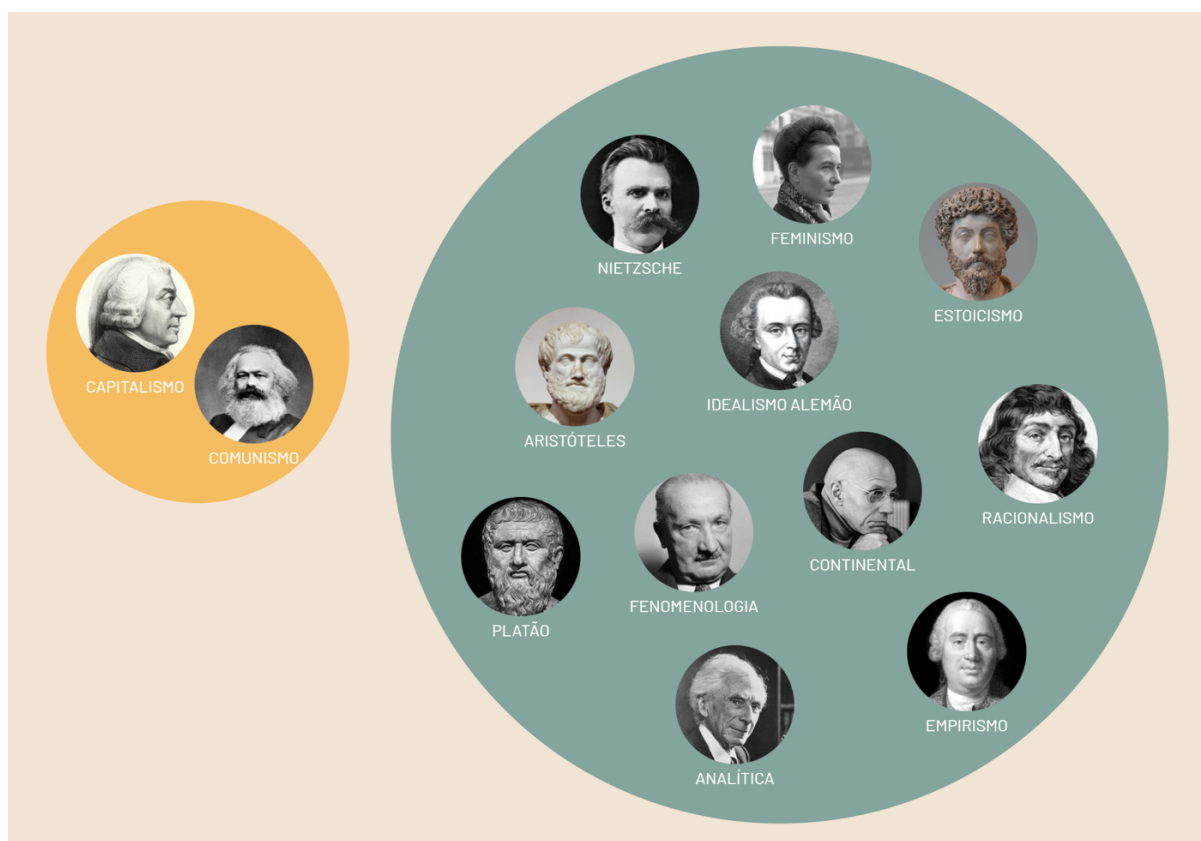


Figura 6. Agrupamento das escolas em dois grupos.

Trabalhos Futuros

Para complementar o projeto será fundamental acrescentar novos registros ao corpus de análise apresentando uma variedade maior de livros e autores pertencentes às escolas já adicionadas e a novas escolas. Percebe-se que períodos da filosofia ocidental, como o medieval e renascentista, estão em falta no corpus assim como escolas de pensamento de outras regiões do planeta, como o confucionismo, daoismo e budismo, as três principais filosofias chinesas. Seria interessante analisar a correlação entre as escolas ocidentais e orientais e criar uma rede mais complexa de conexões entre elas. Para tal, porém, se faz necessário criar um sistema que consegue separar as sentenças de uma obra e adicioná-las no corpus e realizar novamente as análises a fim de gerar novos resultados.