Thematic and Argument Structure Analysis of Plato's Republic Using Data Mining Techniques

Tanishq Daniel, Trevor Ouma, Nate Miller
CIS 635

October 14, 2024

1 Overview of the Project

This project delves into natural language processing (NLP) and text analysis, focusing on a classic in philosophy— Plato's *The Republic.* The goal is to uncover philosophical themes and analyze the structure of arguments, particularly evaluating the complexity of these themes and their associated arguments. We aim to discover whether Plato uses more complex arguments when discussing highly intricate themes, revealing if there is an intentional alignment between the depth of his reasoning and the topics he addresses. Our motivation is to discover themes and argument structures in *The Republic* to allow for deeper insights into its complex reasoning and layered themes. Our approach will involve thematic analysis, argument structure analysis and complexity analysis. For thematic analysis we will use Latent Dirichlet Allocation (LDA) to categorize key themes in the dialogue and evaluate their complexity. Then we will classify argument types and evaluate their complexity. Finally, we will measure if Plato uses more complex arguments when discussing highly complex themes.

2 Related Work

In Latent Dirichlet Allocation (2003) the authors introduce and describe LDA as a Machine Learning technique for identifying topics. Other research teams have used this technique to evaluate philosophical texts for themes in philosophy and other subjects. The paper Eight Journals over Eight Decades by Jebeile and Kennedy (2020) applies LDA to topic modeling in the field of philosophy of science, analyzing articles from eight journals over eighty years. This project will uniquely use this technique to identify topics within a single philosophical work and then use them for complexity analysis.

3 Data Plan

Data Source

We will use publicly available digital versions of *Plato's Republic.* Project Gutenberg provides the Jowett English translation we rely on for this project.

Preprocessing

Text Cleaning

We will begin with text cleaning, which involves removing metadata, footnotes, and irrelevant sections from the text.

Sentence Tokenization

Tokenization will be used to split the text into manageable units for argument and theme detection.

Data Structuring

We will use Pandas to organize the text data into structured formats such as DataFrames for easier manipulation and access.

Numpy Utilization

Finally, we will utilize Numpy for managing numerical arrays and performing operations involved in text vectorization and statistical measures.

4 Implementation Plan

Topic Modeling for Theme Detection

We will use Scikit-learn to implement Latent Dirichlet Allocation (LDA) to discover main themes. Other themes will be labeled as sub-themes, based on a combination of frequency and the number of sub-topics they straddle.

Theme Complexity Analysis

Theme complexity will be evaluated programmatically based on three criteria: the number of unique sub-themes associated with the theme, the length of discussion measured by the word count of passages focused on the theme, and the argument count, which is the number of unique arguments used to illustrate the theme.

Argument Structure Detection

We will utilize Pandas to store and analyze the detected premises and conclusions in DataFrames.

Argument Complexity Analysis

Argument complexity will be evaluated programmatically based on three criteria: the argument structure, measured by the number of premises and conclusions used to make the argument; the examples used, which counts the number of examples employed to argue points; and the counterarguments, reflecting the number of counterarguments utilized in the argument.

Correlation Analysis

Using Spearman's Rank Correlation, we will utilize Scipy to evaluate the correlation between the ranked complexity of themes and the ranked complexity of their top arguments. Themes and arguments will be ranked based on their complexity scores, and the correlation coefficient will indicate whether more complex themes correspond to more complex arguments. We will identify premises and conclusions within the text using dependency parsing and semantic role-labeling techniques, with Pandas used to store and analyze the detected premises and conclusions.

Visualization

We will use Matplotlib to create visualizations of the relationships between top themes and arguments.

5 Evaluation Plan

Qualitative Analysis

We will manually review a subset of the LDA-extracted topics to ensure alignment with known philosophical themes in *The Republic*. Additionally, argument structures detected through NLP methods will be compared with hand-labeled sections to validate their accuracy.

Quantitative Analysis

We will use cross-validation techniques to test the correlation strength of smaller sections. If individual sections show a correlation between theme and argument complexity, then that is evidence supporting our findings.

Success Criteria

We aim to find a meaningful correlation between the complexity of themes and arguments. A Spearman's Rank Correlation coefficient close to 1 would suggest a strong alignment between thematic and argument complexity, while a low or negative correlation would imply no or an inverse relationship.

6 Plan for Group Collaboration

We will meet weekly via video call for progress reviews. We will work together in a Google Colab project to ensure that we all have the updated files. This will allow us to easily see which member is contributing to which section of code. The repository will be pushed to Github.

7 Timeline

Weeks	**Tasks**
Week 1-2	Text preprocessing and topic modeling with LDA.
Week 3-4	Theme complexity and argument detection.
Week 5-6	Argument complexity analysis and Spearman's rank correlation.
Week 7	Validation and visualization.
Week 8	Prepare final report.

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research.*

[2] Jebeile, J., & Kennedy, E. (2020). Eight journals over eight decades: A computational topic-modeling approach to contemporary philosophy of science. *Synthese*, 198(6), 5239–5274.

[3] Jowett, B. (1892). *Plato's Republic.* Project Gutenberg.