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

Project Progress Report

Tanishq Daniel, Trevor Ouma, Nate Miller

CIS 635

November 8, 2024

OVERVIEW

We have successfully implemented basic preprocessing, argument detection, and theme detection techniques to examine Plato's Republic. We are using the dataset found at this URL:

https://raw.githubusercontent.com/GVSU-CIS635/Datasets/refs/heads/master/republic.txt

Using LDA, we are able to pull out relevant topics that correspond to actual themes in the text, including justice, the ideal state, and truth. We are able to identify and organize arguments into a data frame. The first steps have been made in evaluating the complexity of themes, but our criteria for argument complexity have been challenging, as will be explained in the next section.

CHALLENGES

A major challenge of this work is that the dataset is a work of philosophical prose. It has been challenging to filter out the narrative and get to the technical aspects of Plato's philosophy. In contrast to a purely technical document that has a precise vocabulary and methodology, this work abstracts many of its topics and relies on allegories (e.g. Plato's cave allegory). Arguments can also be challenging due to the usage of abstract metaphors and allegories. The extent of this challenge was not fully considered when choosing LDA. Primarily we have focused on preprocessing as a way to deal with this challenge. Taking more abstract and descriptive words has yielded better results.

Our analysis of argument complexity has proven difficult based on the criteria we selected. It is difficult to identify examples and counter arguments effectively. If we are unable to solve this we will need to adjust that portion of the project by using simpler criteria, such as the word count in the premise and conclusions.

While we will be unable to perfectly extract themes and arguments we can still build a pipeline that extracts some useful information from the text. Our hope is that it could lay the groundwork for more specialized data mining and ML techniques that could overcome that challenge to a greater degree.

COLLABORATION

All three of us meet regularly via Zoom, communicate over text, and have discussions after class about the project. We have all contributed meaningfully to the project.

NEXT STEPS

Theme Extraction:

Improve topic extraction by tuning the hyperparameters of the LDA model. Find better methods of evaluating extraction performance.

Theme Complexity:

Improve theme complexity based on sub-themes, word count, and argument count. Refine the data object used to report this for easier analysis and visualization.

Argument Extraction:

Improve argument extraction to identify more than simple argument structures. Manually validate that the arguments extracted are coherent.

Argument Complexity:

We must either solve our challenging criteria or decide on simpler criteria. Like the theme analysis, we also need to refine the data object returned by this analysis.

Complexity Analysis:

We are trying new methods to evaluate the relationship between theme and argument complexity. Some other ideas include Pearson's Correlation Coefficient or Regression Analysis.

Visualization:

Once we have a good data pipeline, we will use Matplotlib to create visualizations comparing theme and argument complexity.