Textbook to Bullet Points

Shreya Nakkala
Texas A&M University
College Station, TX USA

Melissa Zhang
Texas A&M University
College Station, TX USA

Abstract

Textbooks can be dense with words pertaining to a cluster of topics. Reading and searching through such a wall of text is difficult and tedious. In order to create a better reading experience, we introduce a method which splits the paragraphs into individual sentences which are grouped by topic, bullet pointed, and indented based on how general or dense the specific sentence is.

Sentence groupings will be achieved by using latent Dirichlet allocation (LDA) to identify the topics in individual sentences. A pre-trained SVM model would then be applied to the grouped sentences to classify whether the sentence contains a lot of details or if it is a generic and broad sentence in the paragraph. The final classification will then determine the indentation applied to the sentence which will be reflected in the txt of the notes generated.

1. Introduction

Textbooks are an essential part of a student's educational life. In 1994-95, the U.S. Department of Education surveyed 3,994 K-12 teachers and found that 74 percent of teachers used textbooks at least once a week [3, p.18]. As these students continue their education, the need for textbooks doesn't diminish. In a research study sponsored by the Babson Survey Research Group, 68 percent of the 2,700 U.S. faculty surveyed reported that they have a required textbook [8, p. 7].

However, textbooks are not always the easiest to read. Some common reasons for this difficulty in reading was cited to be due to an overload of information and the inability to recognize text structures such as the relationship between main and sub topics [10, p. 3]. After surveying around 25 students about their experience reading textbooks, we found that most of the feedback agreed that textbooks are dense and often structured in a way which loses the readers interest. Students also stated they felt overwhelmed with the amount of different information in paragraphs making it hard to process and connect the context while reading the textbook.

In order to combat this problem, we introduce a method which will break texts within the textbook into individual sentences which are bullet pointed and grouped by topic. We achieve this by using a latent Dirichlet allocation (LDA) model to find the hidden topics of each sentence. We then proceed to use the first sentence of each grouping as the main idea (to provide easier readability) and the rest are indented based on a classification given by the SVM model we trained. By doing so, our goal is to break up the textbook into more readable segments which can show a very basic relationship between the sentences. To prevent a loss of information, we avoided the removal of any sentences. The final result would then be a reconstruction of the textbook formatted as a rough outline.

As our data set we chose to use the human geography textbook titled *Human Geography: People, Place, and Culture 9th Edition* (Erin et al., 2009). We chose this textbook due to the way it's structured with a clear distinction between topics as well as between generic and detailed sentences. During testing, our LDA model was able to mostly group sentences from the inputted paragraphs together into the different topics. These grouped sentences were then passed to the SVM sentence classifier which determined whether the sentence was broad or detailed. The results were then outputted into a Notes.txt file which contained the indented structure of the outline from the paragraphs inputted.

2. Methods

2.1. LDA

Latent Dirichlet allocation (LDA) has been ruled 'the' topic modeling technique since its inception [9]. LDA is a probabilistic topic model which uses unsupervised learning that contains the basic idea that documents are a bag of words with a random mixture of topics. Topics are then defined as a distribution over words. The goal of LDA is to then infer the underlying topic structure existent in a document [1]. Previous research has also shown LDA to be effective even when it comes to short texts [5, 12]. This is beneficial for our research as LDA will be used to infer the underlying topic structure within a paragraph. A paragraph is used as we assume that by writing conventions, authors will have already grouped sentences that contribute to one unifying broad main topic together. Our goal is to then break these broad topics into more easily understandable segments.

2.1.1 Preprocessing Paragraphs

In order to ensure the most accurate results, we used the NLTK Python library to help preprocess the inputted paragraph and remove any noise that may interfere with the topic modeling. First, we used a sentence tokenizer to divide the paragraph into sentences. These sentences were then passed into a word tokenizer to be separated into words. At this point, all stop words were removed (such as 'the' or 'however'). Afterwards, we proceeded to the lemmatizer to have the inputted paragraph undergo lemmatization which prevents words and their various inflected forms from being counted multiple times (such as 'rock' and 'rocks'). However, we opted to skip the stemming process as there was a possibility that harsh stemming could damage the resulting model [7]. After preprocessing the paragraph, the Gensim Python library was used to format the corpus and dictionary. These were then passed into Gensim's LDA model.

2.1.2 Determining Alpha and Eta

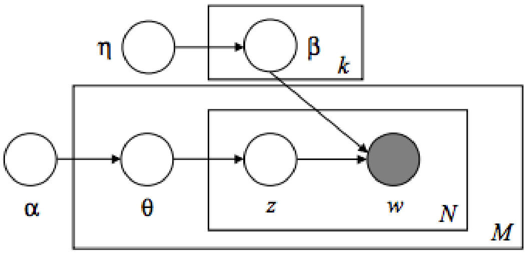


Figure 1: Graphical model of LDA

Figure 1 shows how LDA represents documents as a bag of words with a random mix of topics. In the figure, M is the number of documents, N is the number of words in each document, and k is the number of topics. Each β in k then represents a topic distribution over words. On the left, we see θ which is the topic proportions per document. The z is then the topic assignment per word which is drawn from θ. Finally, w represents a word which is observed in a document with a probability derived from z and β. For LDA, we are given that θ and β are distributions derived from a Dirichlet distribution. Therefore, α and η serve as crucial Dirichlet parameter that represents the prior for the topic distribution for documents and topic distribution per word [1].

In order to determine which combination of α and η would best fit our purposes, a two phase testing process was introduced. During the first phase, all combinations from a determined set of α values and η values were used on eight paragraphs. To produce more coherent groups of sequential sentences, these α and η values needed to be small to discourage any ambiguity that may occur due to the overlapping of topics and topic words. The outputted grouped sentences are then scored on a scale of 0 to 3 where 0 is the worst and 3 is the best. The average value is then computed and compared with every other combination.

Alpha	Eta
Symmetric	None
Asymmetric	Auto
Auto	0.1
0.1	0.01
0.01	0.001
0.001	0.0001
0.0001	0.00001
0.00001	0.000001

Table 1: Set of all alphas and etas used for testing

The best combinations would then move on to phase 2 where they are tested against 105 paragraphs using the same scoring method. The combination with the highest average score would then be used as the value for α and η.

For both phases, the Gensim coherence model was used to measure the coherence of the topic model when the number of topics is 2, 3, 4, or 5. The coherence measure of C_{NPMI} was used for the model due to its accuracy and the usage of a small boolean sliding window which helps capture some degree of proximity between words [6]. After measuring the coherence for all the given number of topics, the one with the highest coherence score was then used to output the best sentence grouping.

Phase 1: 8 paragraphs and All Combinations

After averaging the scores for phase 1, we found that the averages achieved when the α was set to 'asymmetric' was overwhelmingly higher compared to the other α values.

Alpha	Average
Symmetric	1.547
Asymmetric	**2.375**
Auto	1.656
0.1	1.516
0.01	1.219
0.001	1.219
0.0001	1.188
0.00001	1.141

Table 2: Average score per alpha value

Intuitively, this could be attributed to the fact that by common writing convention, sentences in a textbook paragraph should be unified in conveying a singular broad main topic. Therefore, the concentration of topics shouldn't be symmetric. Additionally, this is in line with previous research which demonstrated asymmetric alpha values having a substantial advantage over symmetric alpha values [11].

Phase 2: 105 paragraphs and All Etas

After averaging the scores of each η value in combination with an α set to asymmetric, it was found that the η value of 0.000001 produced the best results.

Asymmetric Alpha	
Eta	Average
None	2.324
Auto	2.324
0.1	2.210
0.01	2.324
0.001	2.343
0.0001	2.381
0.00001	2.333
0.000001	**2.390**

Table 3: Average score for eta values of an asymmetric alpha

2.1.3 Determining Number of Topics

In order to determine the best number of topics to use, we decided to plot the best topic number (based on the highest coherence value) vs the paragraph word count for each η value tested as seen in Figure 6. In the figure, we can observe that a majority of the the sentences (mostly between the word count of about 20 to 130) were shown to have a higher coherence value when the topic number was set to 2. Even though the average best topic number seems to be suggesting 3 as the best topic number, it was important for us to note that this average was likely achieved due to only a slightly better coherence score for sentences in the word count range of 20 to 100. As shown by Figure 7, many of the paragraphs with a word count lower than 109 only exhibited slight changes in coherence values. Therefore, we determined that the topic number of 2 was the best value to use.

2.2. SVM

The SVM classifier is the most fitting to do this job of classifying broad versus detailed sentences due to it being able to pick up several more patterns than other techniques. [4] When testing with clustering models for example, outliers would often change the results of the classification. On the other hand, these outliers do not influence the SVM model as much. Because sentence structure fluctuates, we determined that using an SVM was the better appraoch. Due to this, with the features of entities and sentence length, the SVM model will be able to provide a pretty accurate classification between broad versus more detailed sentences.

2.2.1 Classifying Model

From the textbook paragraphs, we decided to classify the sentences based on whether the sentences being general and broad vs detailed and specific. While testing, we observed that many detailed and specific sentences exhibited a high amount of unique entities. To gather the number of entities that are in each of the sentences, the spaCy Python library was utilized. With spaCy, we used the named entity recognizer (NER) which gathered the number of entities for each sentence.

During testing, we found that using only the number of entities, the SVM was able to clearly distinguish extremely broad sentences and extremely detailed sentences perfectly. However, the use of only one parameter lead to the minor problem when differentiating the sentences with the median amount of entities. Therefore the sentence length was added as a parameter to the classification. If the length of the sentence is too long, we decided to indent the sentence as if it was a detailed sentence. This is due to the fact that readability of outline will improve if extremely long sentences are indented twice inwards (as if they were detailed sentences) instead of only once for broad sentences. In this fashion, readers can avoid having combing across long, tedious sentences first.

3. Experiments and Test Results

3.1. Data Set

For our data set we used *Human Geography: People, Place, and Culture 9th Edition* (Erin et al., 2009). [2] We chose this due to this textbook having a reasonably clear distinction between topics. Paragraphs are also organized so that a mixture of topics exist within a single paragraph. Due to time and memory constraints, only select sections of the textbook will be used. The paragraphs also contain many easily differentiable broad or detailed sentences which can be used to train our SVM.

3.2. Sentence Classification

3.2.1 Testing and Training

For sentence classification we decided to use a total of 100 sentences, 20 sentences for training and 80 sentences for testing. Figure 1 shows the training data plotted alone. Viewing this we can clearly see the difference between the broad sentences and the detailed sentences. Figure 2 provides a boundary that the SVM model chose with the training data. The testing data is plotted to show where the linear decision boundary is positioned.

Correct	Incorrect	Total
70	10	80
.875	.125	1.0

Figure 2: Accuracy From Testing

3.3. Results

The final output of the topic modeling and sentence classification are shown in in Figures 9 and 11 with their corresponding paragraphs in Figure 8 and 10. Given this output, we proceeded to survey students about their reading experience of the textbook excerpt compared to our method's outputted notes. From the survey, 88% of people responded that they do prefer the notes over the textbook.

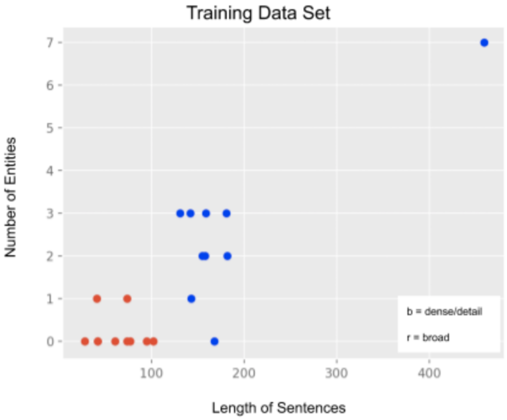


Figure 3: Training Data Set Plot

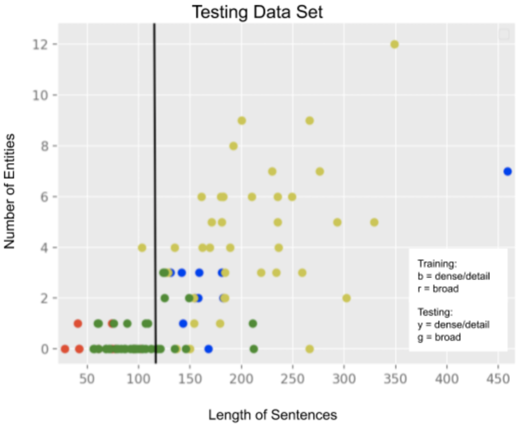


Figure 4: Testing Data Set Plot

4. Conclusion and Future Expectation

In our current project, we currently group 2 topics in each paragraph, but we hope to broaden that in the future. This will allow for a more precise topic grouping, but based on the size of the paragraphs and the coherence measure we chose to go with 2 as the parameter for the number of topics for this model. We also hope to organize the outline better. Some of the feedback we got from our survey after displaying our outline is that it is structured better than textbooks, but it is not structured as concise as notes are. For this we propose, in the future, to remove sentences that are found unnecessary. We also kept the sentences in the outline the same order they were in in the textbook, this was to not lose any information. In the future, something we could do to better this would be to expand from one paragraph to a section in the textbook and move sentences around based on the topics found in the section. This would reorganize the sentence order, but would give the reader more sentences of the same topic grouped together.

5. Task Assignment and Acknowledgement

This project was completed by the cooperation of Melissa Zhang and Shreya Nakkala,under the instruction of Dr.Zhangyang(Atlas) Wang and Zhenyu Wu. Zhang took charge of the Topic Modeling and Nakkala took charge of the sentence classification, and they collaborated on this report. The dataset was provided by the human geography textbook [2].

References

[1] David Blei. Topic models. https://www.youtube.com/watch?v=DDq3OVp9dNA, November 2009.

[2] Erin Hogan Fouberg, Alexander B. Murphy, and De Blij Harm J. *Human geography people, place, and culture.* Wiley, 2012.

[3] Robin R. Henke, Xianglei Chen, and Gideon Goldman. *What Happens in Classrooms? Instructional Practices in Elementary and Secondary Schools, 1994-95. Statistical Analysis Report.* 1999.

[4] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics-Proteomics*, 15(1):41–51, 2018.

[5] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Searching microblogs: Coping with sparsity and document quality. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 183–188, New York, NY, USA, 2011. Association for Computing Machinery.

[6] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[7] Alexandra Schofield, Måns Magnusson, and D Mimno. Understanding text pre-processing for latent dirichlet allocation. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, volume 2, pages 432–436, 2017.

[8] Julia E. Seaman and Jeff Seaman. *Opening the Textbook: Educational Resources in U.S. Higher Education, 2017.* 2017.

[9] Deepak Sharma. A survey on journey of topic modeling techniques from svd to deep learning. *International Journal of Modern Education and Computer Science*, Vol. 9:PP.50–62, 07 2017.

[10] Richard W. Strong, Harvey F. Silver, and Gregory M Perini, Matthew J.and Tuculescu. *Reading for Academic Success: Powerful Strategies for Struggling, Average, and Advanced Readers, Grades 7-12.* 2002.

[11] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981, 2009.

[12] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, 2010.

A. Appendix

Figure 5: width=8in

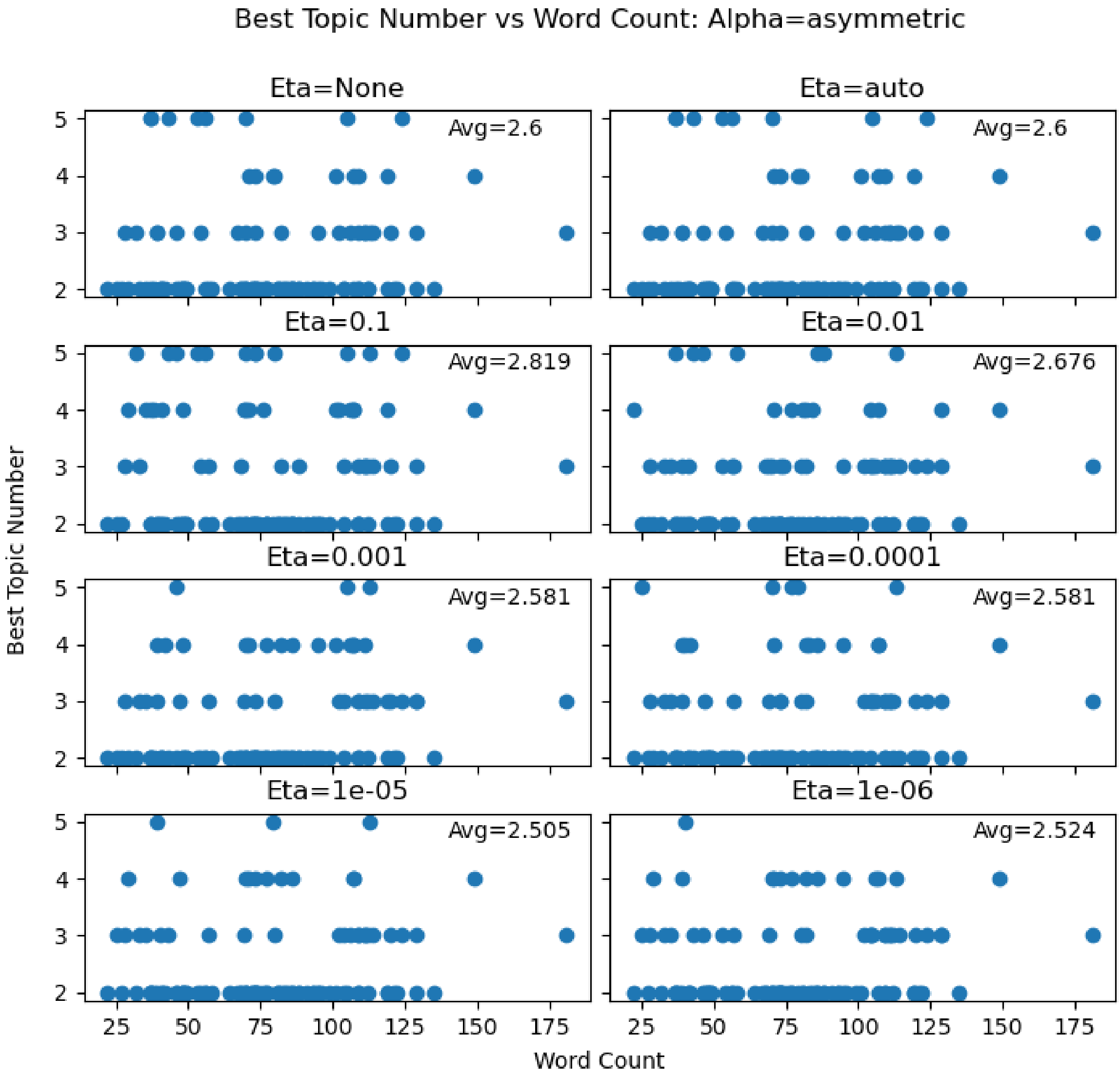


Figure 6: Scatter plots graphing the relationship between Best Topic Number and Word Count

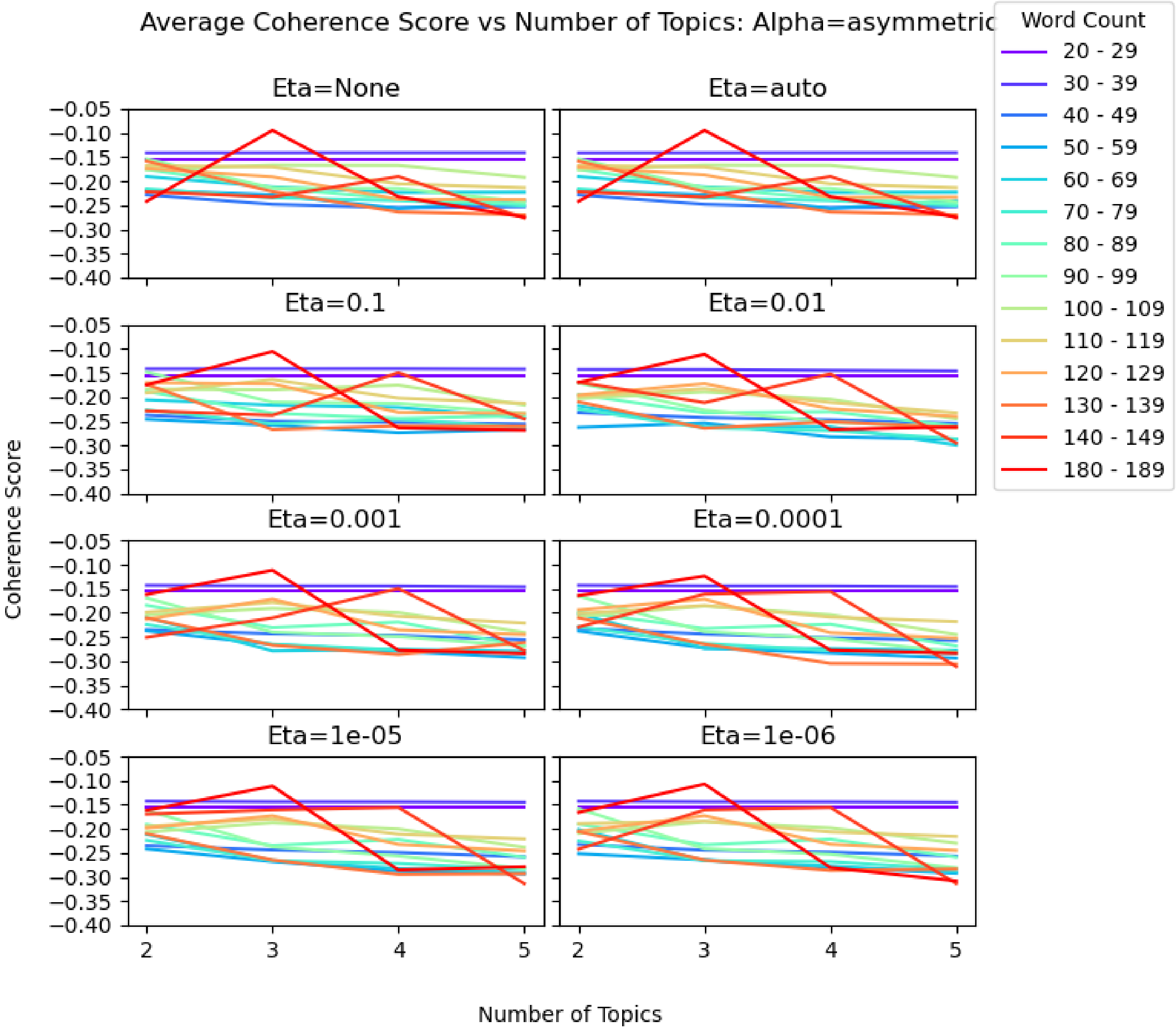
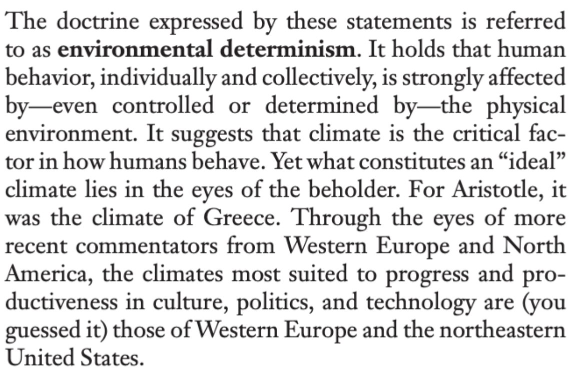
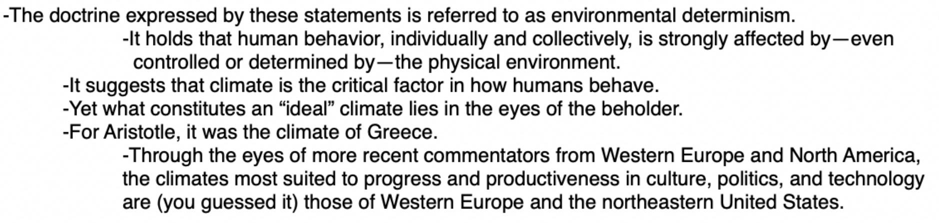


Figure 7: Line plots graphing the coherence scores over number of topics for each word count range



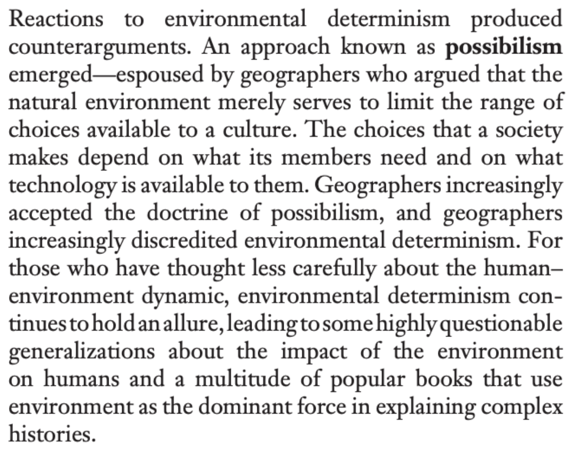
The doctrine expressed by these statements is referred to as **environmental determinism**. It holds that human behavior, individually and collectively, is strongly affected by—even controlled or determined by—the physical environment. It suggests that climate is the critical factor in how humans behave. Yet what constitutes an "ideal" climate lies in the eyes of the beholder. For Aristotle, it was the climate of Greece. Through the eyes of more recent commentators from Western Europe and North America, the climates most suited to progress and productiveness in culture, politics, and technology are (you guessed it) those of Western Europe and the northeastern United States.

Figure 8: Paragraph from the textbook corresponding with figure 10



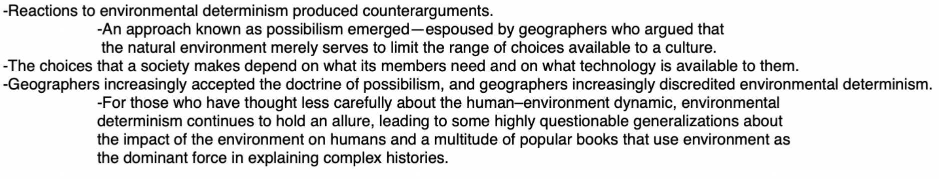
```
-The doctrine expressed by these statements is referred to as environmental determinism.
                -It holds that human behavior, individually and collectively, is strongly affected by—even
                  controlled or determined by—the physical environment.
        -It suggests that climate is the critical factor in how humans behave.
        -Yet what constitutes an “ideal” climate lies in the eyes of the beholder.
        -For Aristotle, it was the climate of Greece.
                -Through the eyes of more recent commentators from Western Europe and North America,
                the climates most suited to progress and productiveness in culture, politics, and technology
                are (you guessed it) those of Western Europe and the northeastern United States.
```

Figure 9: Outline output of paragraph from Figure 9



Reactions to environmental determinism produced counterarguments. An approach known as **possibilism** emerged—espoused by geographers who argued that the natural environment merely serves to limit the range of choices available to a culture. The choices that a society makes depend on what its members need and on what technology is available to them. Geographers increasingly accepted the doctrine of possibilism, and geographers increasingly discredited environmental determinism. For those who have thought less carefully about the human–environment dynamic, environmental determinism continues to hold an allure, leading to some highly questionable generalizations about the impact of the environment on humans and a multitude of popular books that use environment as the dominant force in explaining complex histories.

Figure 10: Paragraph from the textbook corresponding with figure 12



```
-Reactions to environmental determinism produced counterarguments.
                -An approach known as possibilism emerged—espoused by geographers who argued that
                 the natural environment merely serves to limit the range of choices available to a culture.
-The choices that a society makes depend on what its members need and on what technology is available to them.
-Geographers increasingly accepted the doctrine of possibilism, and geographers increasingly discredited environmental determinism.
                -For those who have thought less carefully about the human–environment dynamic, environmental
                determinism continues to hold an allure, leading to some highly questionable generalizations about
                the impact of the environment on humans and a multitude of popular books that use environment as
                the dominant force in explaining complex histories.
```

Figure 11: Paragraph from the textbook corresponding with figure 11