Exploring Textual Relationships: A Deep Learning Approach

Janiya Richardson
Advisor: Dr. Christopher Oakley
Spelman College

INTRODUCTION

Extracting meaningful insights from textual data is crucial to making informed decisions and representing knowledge. This project presents an innovative approach to automatically generate mind maps from text by leveraging AI techniques and the PyTorch framework. Semantic relationships between words and phrases are extracted through NLP and deep learning algorithms to create intuitive visual representations of concepts and their interconnections. The system has various stages, including data preprocessing, word embeddings, neural network modeling, and graph-based visualization. Our methodology facilitates the construction of comprehensive mind maps that capture the essence and structure of the underlying information by training neural networks to recognize semantic associations within text. This project contributes to advancing the field of AI-driven text analysis and visualization, offering a novel approach to transforming raw text into actionable insights by creating dynamic and interactive mind maps.

In this project, deep learning algorithms will be used to analyze physics texts and lectures to generate mind maps for main concepts within topics and network maps for word association.

PyTorch

OBJECTIVES

- To generate a comprehensive mind map that capture the theme and structure of the underlying information in a text
- To provide users a comprehensive network map of word associations and themes, enabling more profound insights into the underlying language structure.
- Evaluation of the system's effectiveness demonstrates its capability to accurately capture and represent complex relationships in diverse textual datasets with network metrics.
- Compare texts generated by a variety of actors in an academic environment.(e.g. professors , senior-level students, & highschoolers)

METHODOLOGIES

Pre-Trained Large Language Models and K-Clusters

The pre-trained models GPT and BERT were used to developed the mind and network maps by tokenizing the text into individual tokens and generate contextualized embedding tokens to capture the semantic information into vector representations. The embeddings was used tor thematic and similarity clustering using the KMeans clustering algorithms.

Word/Similarity Matrix

The word matrix was constructed by encoding the input text data into a structured matrix format, where rows represented individual tokens or words, and columns corresponded to contextual features or attributes. Each cell in the matrix contained numerical values representing the frequency. By analyzing the patterns and distributions of words across the matrix dimensions, we uncovered underlying themes, topics, and concepts present in the input text

	Contextual Feature 1	Contextual Feature 2	Contextual Feature 3
Token/Word 1	Frequency 1,1	Frequency 1,2	Frequency 1,3
Token/Word 2	Frequency 2,1	Frequency 2,2	Frequency 2,3
Token/Word 3	Frequency 3,1	Frequency 3,2	Frequency 3,3
...	...	...	...
Token/Word N	Frequency N,1	Frequency N,2	Frequency N,3

Graph Visualization

For both mind and network maps, the thematic relationships and semantic associations within the input text data was modeled as graph structures, wherein nodes represented individual themes, topics, or concepts, and edges denoted relationships or connections between them.

RESULTS

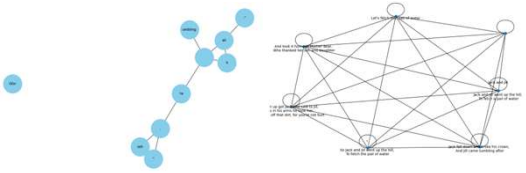


Figure 1

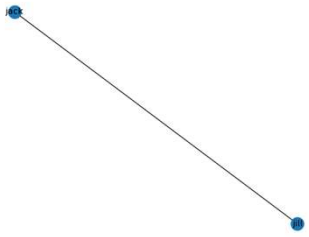
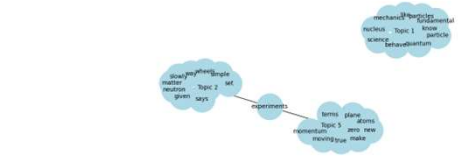
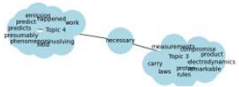


Figure 2





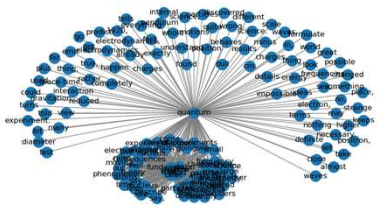


Figure 3

DISCUSSION

As shown in figure 1, the pre-trained models did not generate the maps that we expected for the input text data, the nursery rhyme Jack and Jill. For example, figure 1's mind map (left) was generated using BERT and the network map) was generated using GPT-2. Afterwards, we decided to use a word matrix to generate the maps and it showed better results. For example, in figure 2, the network map (top) only showed the most frequent words in the Jack and Jill besides the stop words, Jack and Jill. The mind map was tested using the 1953 State of the Union speech and was able to show certain topics that were mentioned in the speech, like freedom and nation. After promising results, we decided to test the word matrix algorithm on Feynman's Lecture Volume 1 Section 2-2: Physics Before 1920. In figure 3, the mind map (top) was able to show highlighted words and group them. For instance, quantum, nucleus, and mechanics are in the same topic group in the mind map. The network map (bottom) of figure 3 showed quantum in the center, but one side of the map is hard to see, and certain words can be seen, like atom, phenomenon, and experiments.

LLMs & K-Clusters

We will continue to work with the word matrix algorithm and intend to include the large language models. This will provide a point of comparison for visualizations created by human thematic analysis and network analysis.

Network Analysis

Further work with Network Analysis will limit the nodes to focus the analysis and create a more human-legible format. This will be accomplished by further refining and isolating the content of the text documents used and measuring network metrics like centrality and betweenness in visualizations.

REFERENCES

- Louridas, P. (2020). Algorithms. The MIT Press.Mueller, J.
- Massaron, L. (2018). Artificial intelligence for dummies. John Wiley & Sons, Inc.
- Mariegaard, S., Seidlin, L.D., Bruun, J. (2022). Identification of positions in literature using thematic network analysis: the case of early childhood inquiry-based science education. International Journal of Research & Method in Education. 45(5).

ACKNOWLEDGEMENTS

Thank you to Title III HBCU B, 2022-2023 - Innovation Teaching, Learning and Scholarship for funding this project!