Resume Text Analytics: Professional Narratives through NLP

Introduction:

This report interpretates the process of cleaning and analyzing resume data using various natural language processing (NLP) techniques. The data used for this analysis is sourced from a CSV file named "resumes.csv." The analysis will cover data cleaning, text preprocessing, topic modeling, word cloud generation, word statistics, and network graph analysis.

Data Cleaning and Preprocessing:

The first step in any NLP task is data cleaning and preprocessing.

The following steps were performed:

1. non-alphabetic characters were removed from the resume text to ensure that only meaningful words were retained.
2. Tokenization and lowercasing of words were carried out.
3. Common English stop words were removed from the text using the NLTK library.
4. Bigrams and trigrams were generated to capture meaningful word combinations.

Topic Modeling:

Using LDA technique for identifying topics within a corpus of text.

1. Tokenization of the cleaned text was performed.
2. A dictionary was created, mapping words to unique IDs.
3. The corpus was built, representing the text as a bag of words.
4. An LDA model with 5 topics was trained using the Gensim library.
5. The top 5 words for each topic were extracted, providing insights into the main themes present in the resume data.

LDA Topics:

(0, '0.005"development" + 0.005*"vt" + 0.004*"research" + 0.004*"data" + 0.004*"university"')*

(1, '0.007"data" + 0.006*"engineering" + 0.005*"design" + 0.005*"research" + 0.004*"analysis"')*

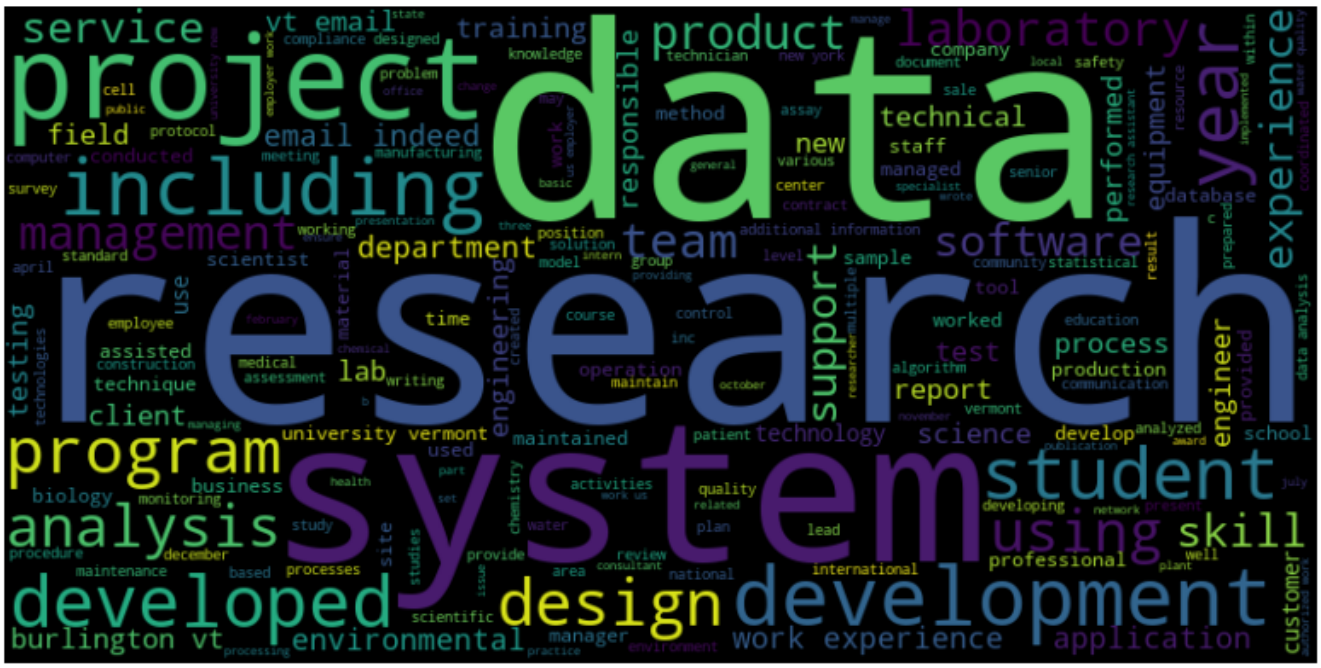
(2, '0.007"vt" + 0.005*"management" + 0.005*"research" + 0.005*"work" + 0.005*"university"')*

(3, '0.012"research" + 0.008*"vt" + 0.008*"data" + 0.007*"university" + 0.007*"laboratory"')*

(4, '0.008"software" + 0.006*"data" + 0.006*"test" + 0.005*"developed" + 0.005*"analysis"')*

Word Cloud:

A word cloud was generated to visually represent the most frequent words in the cleaned resume text. The word cloud provides a quick overview of the most common terms found in the dataset.



Word Statistics:

Several word statistics were calculated to gain a better understanding of the resume text:

Number of words: 64049

Total number of unique words: 11445

Total entropy: 11.74480252985177

The number of words and unique words, along with entropy, provide insights into the diversity and complexity of the language used in the resumes.

Network Graph Analysis:

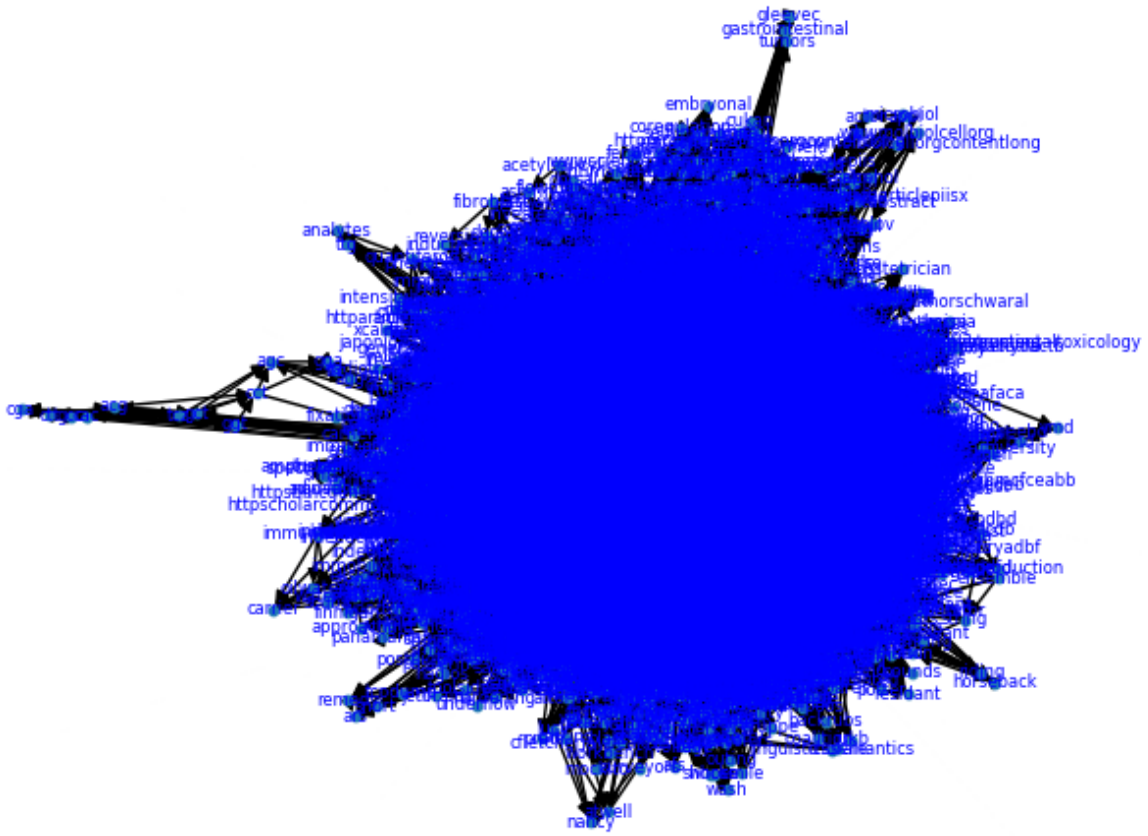
A network graph was constructed to analyze the co-occurrence of words within the resume text.

1. An empty directed graph was created using the NetworkX library.
2. The text from all resumes was combined into a single string.
3. Unique words were identified and added as nodes to the graph.
4. Edges were created between words that co-occurred within a specified window (co-occurrence threshold).

Network Graph Analysis Results:

Top 10 central words: ['research', 'data', 'vt', 'work', 'development', 'new', 'analysis', 'experience', 'including', 'laboratory']

Network graph visualization:



The network graph analysis helps identify words that are highly connected to others, potentially indicating their importance within the corpus.

Conclusion:

This report interprets a comprehensive analysis of resume data, including data cleaning, text preprocessing, topic modeling, word cloud generation, word statistics, and network graph analysis. This analysis provides valuable insights into the content and structure of the resume data, which can be used for various purposes such as keyword extraction, topic classification, and content recommendation.