Graph Theory Project Report

Classification-of-Documents-Using-Graph-Based-Features-and-KNN

Submitted by:

Yahya M. Mirza 2021-CS-11 & M. Usman Asghar 2021-CS-46

Submitted to:

Sir Waqas Ali

Department of Computer Science
University of Engineering and Technology, Lahore
Pakistan

Contents

April 21, 2024

1 Introduction

Document classification plays a pivotal role in various domains such as natural language processing, information retrieval, and data mining. In this project, we propose a novel approach to document classification utilizing graph-based features and the k-nearest neighbors (KNN) algorithm. Unlike traditional vector-based methods, our approach aims to capture the structural relationships within documents to enhance classification accuracy.

2 Data

The dataset utilized in this project consists of documents represented as graphs. Each document is transformed into a graph structure, where nodes represent elements such as words, phrases, or concepts, and edges denote relationships between these elements. The dataset is partitioned into training and testing subsets to facilitate model development and evaluation.

3 Methadologies

3.1 Data Collection and Preparation

To begin the project, we collected or created 15 pages of text for each of the three assigned topics, ensuring that each page contained approximately 300 words. The dataset was then divided into a training set, comprising 12 pages per topic, and a test set, comprising 3 pages per topic. This partitioning ensured a balanced representation of each topic in both the training and testing subsets.

3.2 Graph Construction

Each page of text, consisting of approximately 500 words, was represented as a directed graph. In this graph representation, nodes corresponded to unique terms (words) extracted from the text, which were obtained after preprocessing steps such as tokenization, stop-word removal, and stemming. Edges in the graph denoted term relationships based on their sequential occurrence in the text. This graph construction process captured the structural and semantic relationships within the documents, facilitating further analysis and classification.

3.3 Feature Extraction via Common Subgraphs

To extract discriminative features for classification, we employed frequent subgraph mining techniques on the training set graphs. These techniques enabled the identification of common subgraphs shared across documents related to the same topic. By capturing recurring patterns or motifs in the document graphs, these common subgraphs served as informative features for classification, contributing to the accurate categorization of documents.

3.4 Classification with KNN

The classification phase involved implementing the k-nearest neighbors (KNN) algorithm, utilizing a distance measure based on the maximal common subgraph (MCS) between document graphs. This involved computing the similarity between graphs by evaluating their shared structure, as indicated by the MCS. Test documents were classified based on the majority class of their k-nearest neighbors in the feature space created by common subgraphs. This approach leveraged the structural information encoded in the graph-based features to make accurate classification decisions.

3.5 Evaluation

The performance of our classification system was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. Additionally, confusion matrices were generated to visualize the classification outcomes and identify potential areas of improvement. To highlight the advantages of the graph-based approach, the results were compared against those obtained from traditional vector-based classification methods. This comparative analysis provided insights into the effectiveness and robustness of our proposed methodology.

4 Results

The performance of our graph-based classification approach is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. These metrics are computed and compared against results obtained from traditional vector-based methods. Additionally, confusion matrices are generated to visualize the classification outcomes and identify potential areas of improvement.

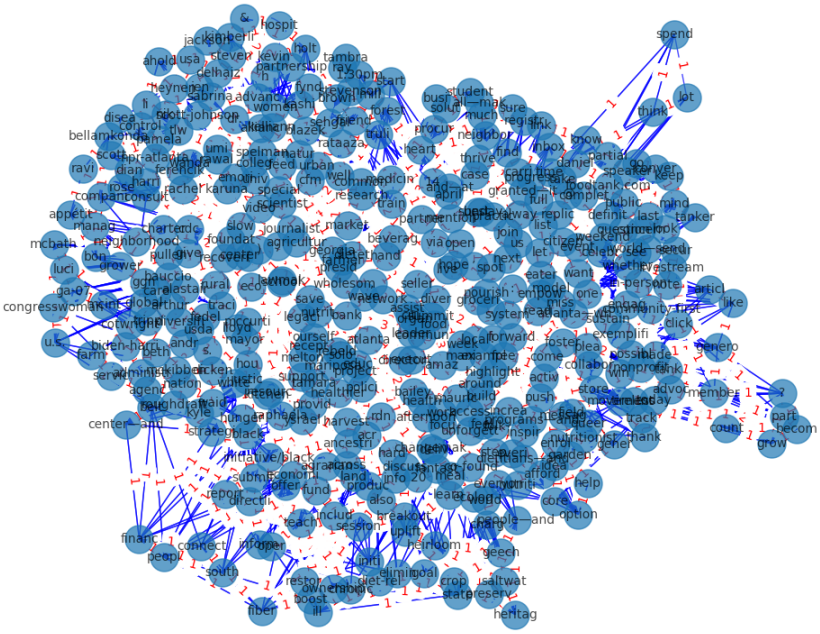

Figure 1: Example of a Document Graph

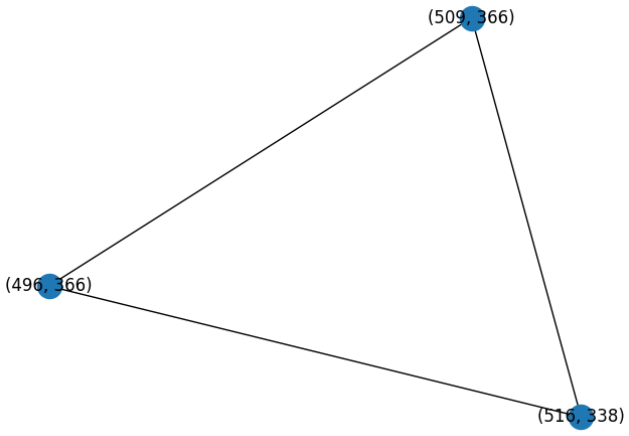

Figure 2: Example of a Common Subgraph

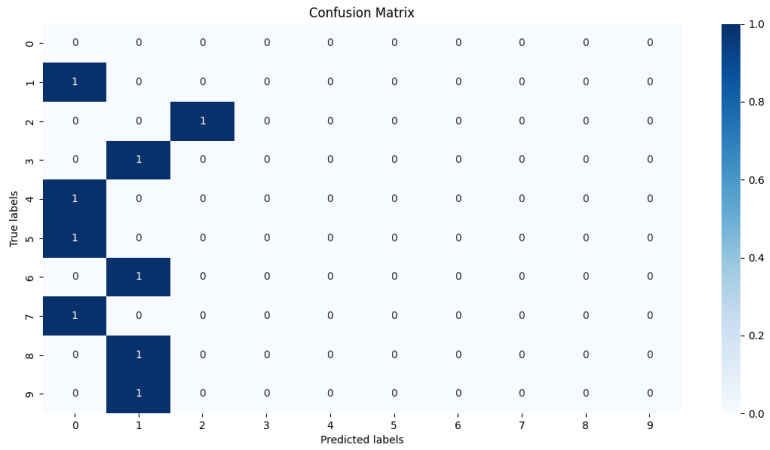

Figure 3: Confusion Matrix for Classification Results

5 Challenges

Throughout the project, we encountered several challenges that influenced the development and evaluation of our classification model. These challenges include:

- Variability in document structures and formats, necessitating robust graph construction techniques.
- Complexity in identifying common subgraphs across a diverse range of documents.
- Optimization of KNN parameters to balance classification accuracy and computational efficiency.

6 Improvements

To address the challenges faced and enhance the effectiveness of our approach, several improvements can be considered:

- Exploring advanced graph construction methods, such as deep learning-based approaches, to capture richer document representations.
- Investigating novel algorithms for common subgraph identification that can handle large-scale graph datasets efficiently.
- Experimenting with alternative classification algorithms or ensemble techniques to further improve classification performance.

7 Implications

The findings of our project have significant implications for the field of document classification:

- The adoption of graph-based features offers a promising avenue for improving document classification accuracy by leveraging structural information.
- Enhanced classification models can facilitate more precise document categorization and retrieval, benefiting various applications including information organization and search engines.
- Continued research and development in graph-based document representation and classification could lead to advancements in document analysis and understanding, with broader implications for knowledge management and artificial intelligence.

8 GitHub Repository

The entire project, including code implementations, datasets, and documentation, is available on GitHub for reference and reproducibility. The repository contains scripts for data preprocessing, graph construction, feature extraction, classification, and evaluation. Additionally, it includes detailed documentation on

each component of the project, along with instructions for setting up the environment and running the code. Collaborators and researchers interested in exploring the methods employed in this study or replicating the experiments can access the codebase and datasets from the GitHub repository at : github.com/Yahya123-hub/Classification-of-Documents-Using-Graph-Based-Features-and-KNN.

9 Conclusion

In conclusion, this project demonstrates the potential of utilizing graph-based features and KNN for document classification. By leveraging the inherent structure within documents, our approach achieves competitive performance compared to traditional methods. Through further refinement and exploration of advanced techniques, we envision continued progress in the field of document analysis and classification.