Classification of Documents Using Graph-Based Features and KNN

Project Supervisor

Mr. Waqas Ali

Submitted By

Abdullah Afzal	2021-CS-80
Muhammad Moaz	2021-CS-101

Department of Computer Science

University of Engineering and Technology, Lahore Pakistan

Contents

Chapter 1 : Introduction

1.1 Project Description

The aim of this project is to develop a document classification system using graph-based features and the K-Nearest Neighbors (KNN) algorithm. By representing documents as directed graphs and leveraging graph similarity measures, the system endeavors to classify documents into predefined topics. This report outlines the methodology employed, including data collection and preparation, graph construction, feature extraction via common subgraphs, implementation of the KNN algorithm, and evaluation of the classification system's performance.

1.2 Purpose of the project

The purpose of this project is to develop an innovative document classification system that leverages graphbased features and the K-Nearest Neighbors (KNN) algorithm. By representing documents as directed graphs and extracting common subgraphs, the system aims to improve classification accuracy and capture the underlying semantic relationships between terms within documents. The project seeks to address the limitations of traditional vector-based models and explore novel approaches to document classification, fostering skills in data representation, algorithm implementation, and analytical thinking.

1.3 Scope of the project

The scope of the project encompasses several key aspects:

- Data Collection and Preparation: Collection of text data for three predefined topics, division into training and test sets, and preprocessing steps such as tokenization, stop-word removal, and stemming.
- Graph Construction: Representation of each document as a directed graph, with nodes representing unique terms and edges denoting term relationships based on their sequence in the text.
- Feature Extraction via Common Subgraphs: Utilization of frequent subgraph mining techniques to identify common subgraphs within the training set graphs, serving as features for document classification.
- Classification with KNN: Implementation of the KNN algorithm using a distance measure based on the maximal common subgraph (MCS) between document graphs, enabling classification of test documents based on the majority class of their k-nearest neighbors.
- Evaluation: Assessment of the classification system's performance using metrics such as accuracy, precision, recall, and F1-score, as well as visualization of classification results through a confusion matrix.

1.4 Project Features

- Hands-on Experience: This project provides practical experience in data representation, algorithm implementation, and analytical thinking through the development of a document classification system.
- Innovative Approach: Unlike traditional vector-based models, this project employs graph-based features to capture the inherent relationships between terms within documents, potentially leading to more accurate classification results.
- Integration of Graph Theory and Machine Learning: By integrating concepts from graph theory with machine learning techniques such as KNN, we have explored the synergy between these disciplines and their applications in real-world problems.

- Feature Extraction via Common Subgraphs: The project utilizes frequent subgraph mining techniques to extract common subgraphs from document graphs, serving as features for classification and enhancing the system's ability to capture shared content across documents related to the same topic.
- Evaluation and Comparison: The performance of the classification system will be rigorously evaluated using metrics such as accuracy, precision, recall, and F1-score. A comparison against traditional vectorbased classification methods will highlight the advantages of the graph-based approach.
- Interdisciplinary Learning: This project engage us with concepts from graph theory, machine learning, natural language processing, and data mining, fostering interdisciplinary learning and expanding their skill set in these areas.
- Potential for Further Exploration: The project lays the groundwork for further exploration and optimization, with opportunities to refine feature extraction techniques, explore alternative graph similarity measures, and scale the classification system to larger datasets.

1.5 Overview of the document

This document serves as a comprehensive report detailing the methodology, results, and insights gained from the project titled "Classification of Documents Using Graph-Based Features and KNN." It begins with an introduction outlining the project's objectives and the motivation behind the chosen approach. The methodology section elaborates on the steps involved in data collection and preparation, graph construction, feature extraction, classification with KNN, and evaluation of the classification system's performance.

Following the methodology, the results section presents the findings of the project, including classification accuracy, precision, recall, and F1-score metrics, as well as a visualization of classification results using a confusion matrix. The conclusion reflects on the significance of the project outcomes and suggests areas for future exploration and optimization.

Throughout the document, emphasis is placed on the innovative nature of the project, its interdisciplinary nature, and its potential implications for advancing document classification techniques.

Chapter 2 : Methodology

2.1 Data collection and preparation

We have scrapped different data from different pages on the topics of food, Health and Fitness, Science and Education. A dataset comprising 15 pages of text for each of the three topics was collected. Each page contained approximately 300 words. The dataset was divided into a training set, consisting of 12 pages per topic, and a test set, comprising 3 pages per topic.

2.2 Graph Construction

Each page of text was represented as a directed graph, with nodes representing unique terms (words) and edges denoting term relationships based on their sequence in the text. All the directed graphs from the text documents have been made and the graph's data is stored in csv.

2.3 Preprocessing

Preprocessing techniques such as tokenization, stop-word removal, and stemming were applied to ensure consistency and reduce noise in the graph representation. To apply these techniques, different algorithms like of stemmation are applied.

2.4 Feature Extraction via Common Subgraphs

Frequent subgraph mining techniques were utilized to identify common subgraphs within the training set graphs. Subgraphs from the set of graphs are made which are used in feature extraction.

2.5 Classification with KNN

The document is classified on the basis of k-nearest neighbors in the feature created by common subgraphs.The KNN algorithm was implemented using a distance measure based on the maximal common subgraph (MCS) between document graphs. The similarity between graphs was computed by evaluating their shared structure, as indicated by the MCS.

2.6 Evaluation

The performance of the classification system is assessed by applying the F1-score algorithm. A confusion matrix was plotted to visualize the classification results.

Chapter 3 : Conclusion

In conclusion, the implementation of a graph-based document classification system using KNN demonstrated promising results. By leveraging graph structures and common subgraphs, the system was able to effectively classify documents into predefined topics. Further experimentation and optimization may lead to enhancements in classification accuracy and efficiency. Overall, this project provided valuable insights into the intersection of graph theory and machine learning in the context of document classification.

3.1 Future Work

Future work could focus on refining the feature extraction process, exploring alternative graph similarity measures, and investigating the scalability of the classification system to larger datasets. Additionally, the integration of deep learning techniques or ensemble methods may further improve the performance of the document classification system.