G11

Classification of Web Documents Using a Graph Model

(GRAPH THEORY)

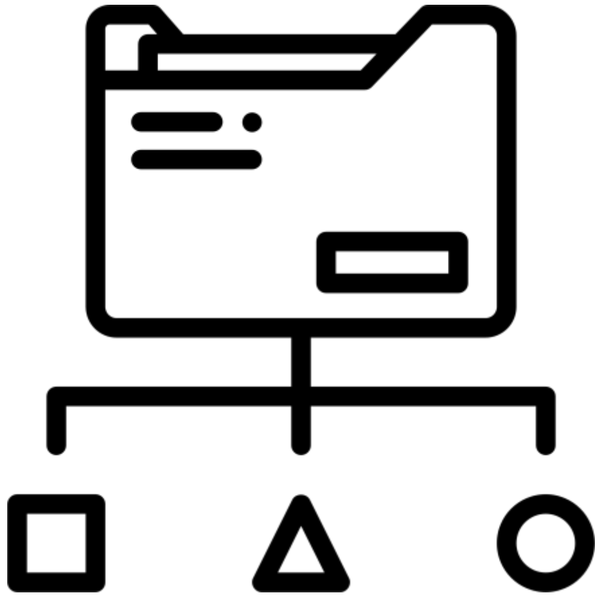
Presentated by:

Messam Naqvi
Muhammad Muneeb

INTRODUCTION

- Document classification is the process of assigning **categories** to natural language documents based on their content.
- Classical approaches like rule induction and Bayesian approaches are based on **vector** models they:

1. Assign values to terms based on frequency.
2. Often discard structural information.
3. Potentially affect classification accuracy.



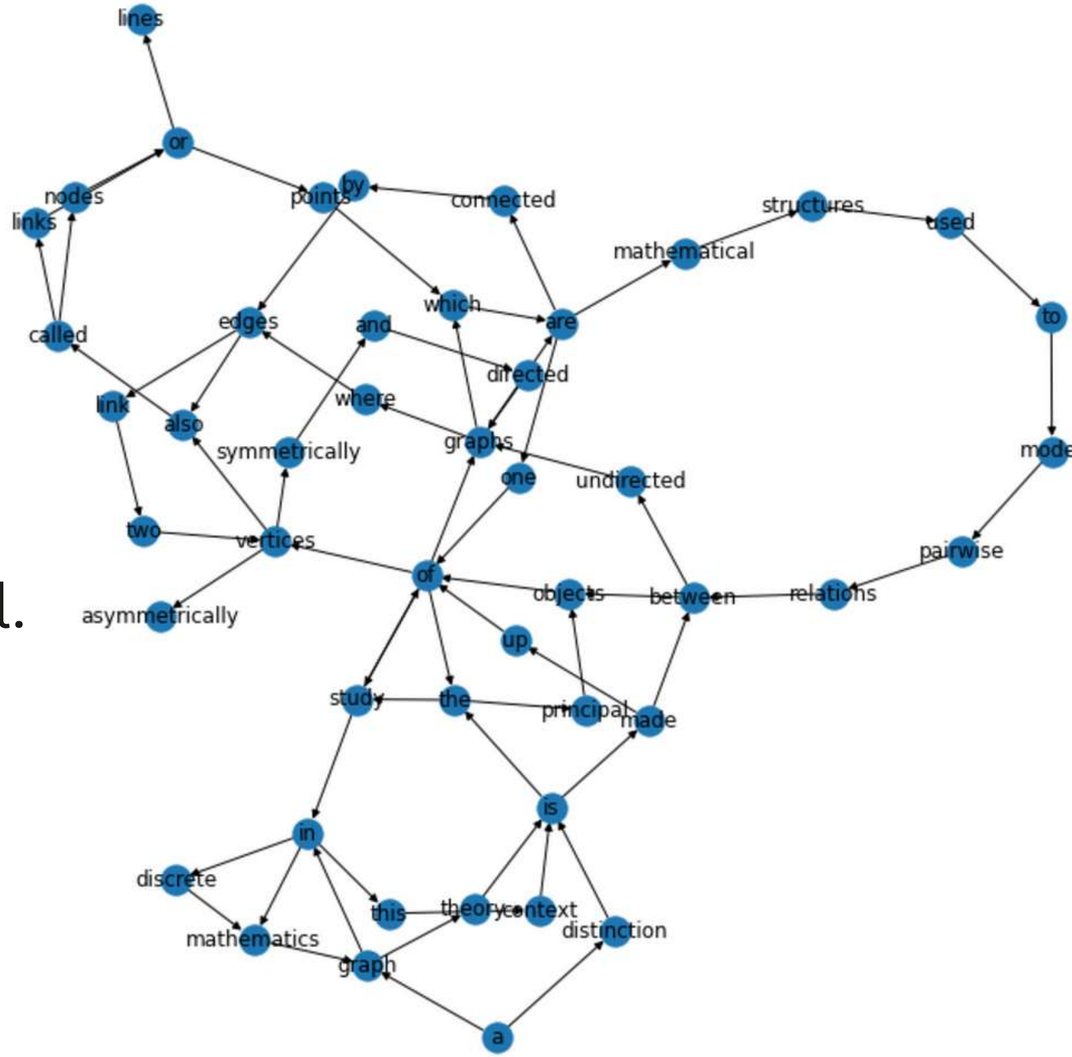
INTRODUCTION

- Graph-Based Approach:

1. Representing documents as **nodes** and **edges**.

2. Structural relationships between terms in documents are captured in the graph model.

3. Graph-based classification methods can outperform traditional vector-based approaches in **accuracy** and **execution** time.



PROBLEM STATEMENT

- Traditional classification methods are limited to numeric feature vectors, losing structural information.
- The challenge is to develop algorithms that **retain structural information** for improved classification accuracy.
- Our project introduces a **graph-based classification approach.**

METHODOLOGY

1

Data Collection

Gather 45 web documents with scrapping

2

Preprocessing

Remove conflicting classifications

3

Graph Representation

Represent all 45 documents as graphs

4

MCS

Measure Distances from graph structure

5

KNN

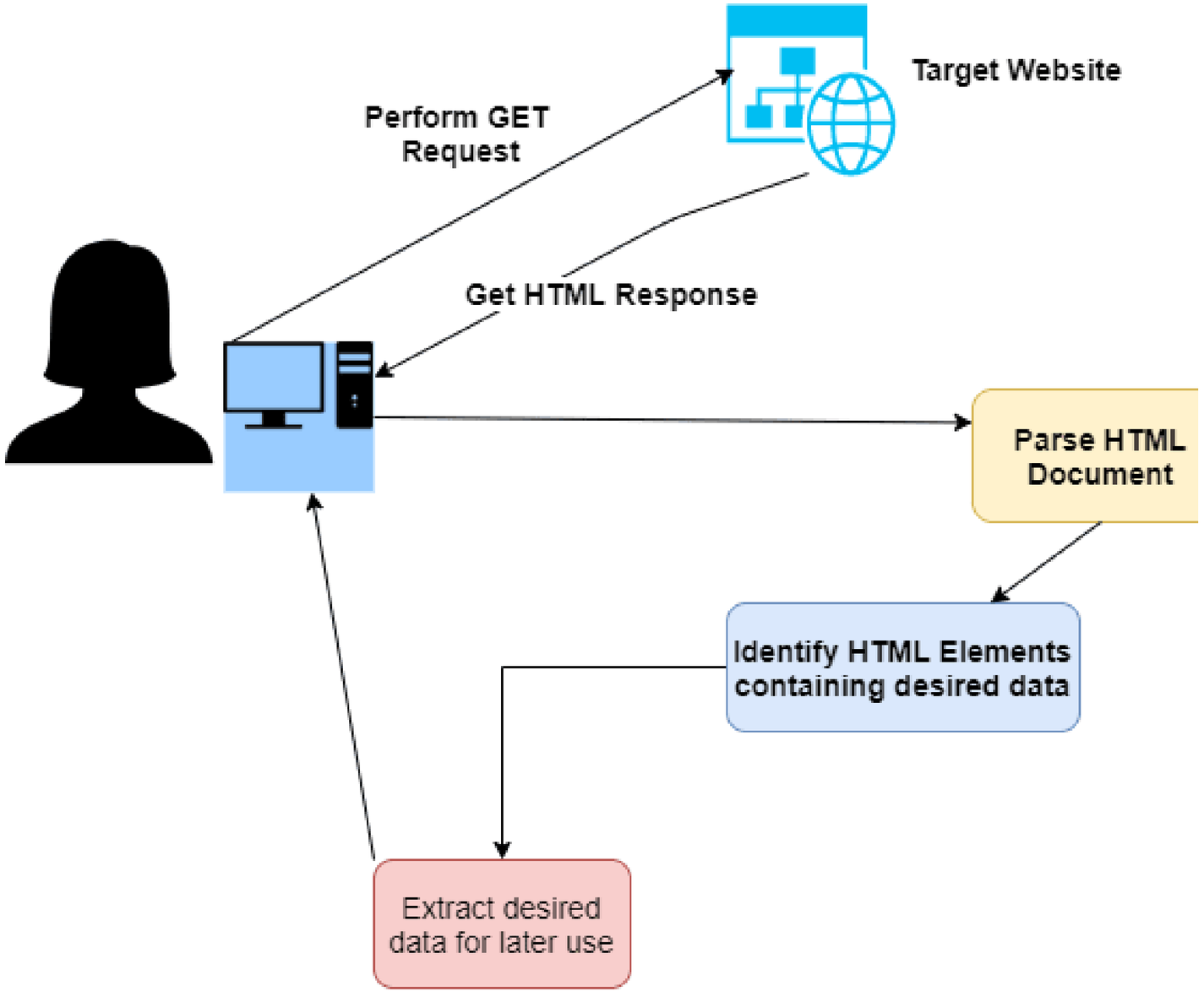
Experiments using the k-NN algorithm with graph-based data

6

Confusion Matrix

Create a confusion matrix

METHODOLOGY



METHODOLOGY

- Tokenization
- Stopword Removal
- Lowercasing
- Removing Non-Alphabetical Words
- Stemming/Lemmatization

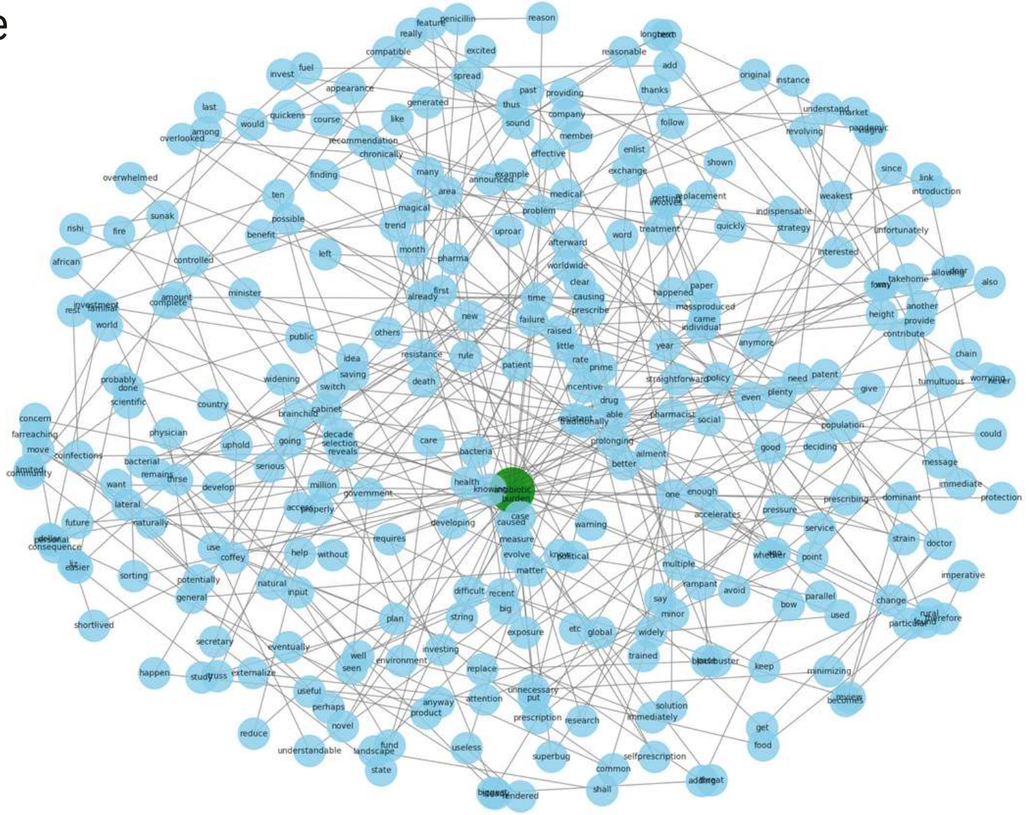
METHODOLOGY

- Node Creation:
  1)Each **term (word)** in the document becomes a node in the graph.
  2) Stop words are excluded from node.

- Node Labeling:
  1) Each node represents a **unique** word.

- Edge Labeling:
  1) Edges are labeled according to the **type of content** they connect (e.g., title, link, text).

- Graph Structure:
  1) Cycles are allowed in the graph.
  2) The graph represents the document's content.



Implementation of Method

METHODOLOGY

- Graph-Based k-Nearest Neighbors (k-NN) Extension
- K=5
- Graph-Based Model
- Distance Computation

Results

```
Class c1 corresponds to 'LifeStyle_and_Hobbies'
Class c2 corresponds to 'Disease_and_Symptoms'
Class c3 corresponds to 'Travel'
```

All Testing Files (names)	Predicted class	Actual class
D13_DS_T.txt	c2	c2
D13_TR_T.txt	c3	c3
D14_DS_T.txt	c2	c2
D14_TR_T.txt	c3	c3
D15_DS_T.txt	c2	c2
D15_LH_T.txt	c1	c1
D15_TR_T.txt	c3	c3
D4_LH_T.txt	c3	c1
D6_LH_T.txt	c3	c1

```
Time of execution: 0.848 seconds
```

Confusion Matrix

1	0	2
0	3	0
0	0	3

- For Class 1, there is 1 correct prediction and 2 misclassifications.
- For Class 2, all predictions are correct.
- For Class 3, all predictions are correct.

Future Work

- **Experiment with Alternative Algorithms:**
  1) Explore **other classification algorithms** that can effectively utilize graph data beyond k-NN.

- **Optimal Graph Node Selection:**
  1) Conduct experiments to determine the **ideal number of nodes** for each graph to optimize performance.

- **Optimize Graph Structures:**
  1) Evaluate different graph representations to enhance performance.

Future Work

- **Ensemble Learning:**
  1) Combine multiple classifiers to improve classification accuracy. Like Random Forest, Gradient Boosting.

- **Semi-Supervised Learning:**
  1) Combine a small amount of labeled data with a larger amount of unlabeled data to train classifiers. Techniques like self-training and co-training.

Q & A Session