An

Industry Oriented Mini Project Report on

# AUTOMATED PHRASE MINING FROM MASSIVE TEXT

Submitted in partial fulfillment of the requirements for the award of

degree

**BACHELOR OF TECHNOLOGY IN**

# COMPUTER SCIENCE AND ENGINEERING

Submitted By

**CH.MADHUMATHI (23TQ1A05D1)**

**CH.DEVAJI (23TQ1A05G2)**

**G.GOWTHAM RAJ (23TQ1A05F3)**

Under the Guidance of

# MS.VIJAYATA RAMTEKE

(Asst.Professor)



**(UGC AUTONOMOUS) (Affiliated to JNTUH, Approved by AICTE,**

**Accredited by NBA & NAAC with A Grade, nirf**

**Ranked & An ISO Certified Institution) Narapally (V),**

**Gatkesar(M), Medchal(D),TS-500088 2024-2025**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

CERTIFICATE

This is to certify that the project report entitled AUTOMATED PHRASE MINING FROM MASSIVE TEXT being submitted by

CH.MADHUMATHI (23TQ1A05D1)

CH.DEVAJI (23TQ1A05G2)

G.GOWTHAM RAJ (23TQ1A05F3)

In partial fulfillment for the award of the degree of Bachelor of

Technology in Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad, is a record of Bonafide work carried out under my guidance and supervision. The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree or Diploma.

# Guide Head of the Department

VIJAYATA RAMTEKE B.RAVINDAR REDDY

(Asst.professor)

Internal Examiner External Examiner

PRINCIPAL

# DECLARATION

We declare that this project report titled ONSCREEN MARKING ANSWER SHEET EVALUATION submitted in partial fulfillment of the

degree of B. Tech in CSE is a record of original work carried out by us under the supervision of MS.VIJAYATA RAMTEKE and has not formed

the basis for the award of any other degree or diploma, in this or any other Institute or University. In keeping with the ethical practice in reporting scientific information, due acknowledgments have been made wherever the findings of others have been cited.

**DATE :**

## SIGNATURE : 4

**CH.MADHUMATHI (23TQ1A05D1)**

**CH.DEVAJI (23TQ1A05G2)**

**G.GOWTHAM RAJ (23TQ1A05F3)**

# 

# ACKNOWLEDGEMENT

We express our sincere gratitude to beloved and highly esteemed institute SIDDHARTHA INSTITUTE OFTECHNOLOGY AND SCIENCES for grooming

us into Computer Science and Engineering graduate, we wish to thank PRINCIPAL DR. M. JANARDHAN for Providing a great learning environment.

We profoundly express our sincere thanks to B.RAVINDAR REDDY , Head Department of COMPUTER SCIENCE AND ENGINEERING , SITS, for his cooperation and encouragement in completing the project successfully.

We wish to record our deep sense of gratitude to our Project In-charge

MS.VIJAYATA RAMTEKE, Assistant Professor, in COMPUTER SCIENCE AND ENGINEERING , SITS, for giving her insight, advice provided during the review sessions and providing constant monitoring from time to time in completing our project and for giving us this opportunity to present the project work.We express our sincere gratitude to our guide MS.VIJAYATA RAMTEKE, Associate Professor, In COMPUTER SCIENCE AND ENGINEERING DEPARTMENT, SITS, who

motivated throughout the period of the project and also for her valuable and intellectual suggestions apart from her adequate guidance, constant encouragement right throughout our work.

Finally, we would like to thank Project Coordinator, Project Review Committee

(PRC) members, all the faculty members and supporting staff, Department of Computer Science and Engineering, SITS for extending their help in all circumstances.

CH.MADHUMADHI (23TQ1A05D1)

CH.DEVAJI (23TQ1A05G2)

G.GOWTHAM RAJ (23TQ1A05F3)

## 

## TABLE OF CONTENTS

DESCRIPTION PAGE NUMBER

**PROJECT TITLE 1**

**CERTIFICATE 2**

**DECLARATION 3**

**ACKNOWLEDGEMENT 4**

**ABSTRACT 7**

## CHAPTER 1:INTRODUCTION 8-11

1.1 PROBLEM STATEMEN

1.2 OBJECTIVE OF PROJECT

1.3 SCOPE OF THE PROJECT

1.4 MOTIVATION

## CHAPTER 2: LITERATURE SURVEY 12-15

2.1 INTRODUCTION

2.2 AUTO PHRASE

2.3 METHODOLOGY

## CHAPTER 3: EXISTING SYSTEM 16-19

3.1 AUTO PHRASE

3.2 PHRASE MINING TOOLS

3.3 NATURAL LANGUAGE PROCESSING(NLP)

3.4 INFORMATION EXTRACTION STSTEM

## CHAPTER 4: PROPOSED SYSTEM 20-22

4.1 SYSTEM COMPONENTS

4.2 SYSTEM ARCHITECTURE

4.3 TECHNICAL APPROACH

4.4 BENEFITS

## CHAPTER 5: SOFTWARE $ HARDWARE REQUIREMENTS

5.1 SOFTWARE REQUIREMENTS

5.2 HARDWARE REQUIREMENTS **23-24**

## CHAPTER 6: SYSTEM DESIGN 25-31

6.1 SOFTWARE DESIGN

6.2 CLASS DIAGRAME

6.3 SEQUENCE DIAGRAME

6.4 ACTIVITY DIAGRAME

6.5 MODULES

## CHAPTER 7: IMPLEMENTATION 32-35

7.1 PROJECT ARCHTECTURE

7.2 SAMPLE CODE

## CHAPTER 8: RESULT 36-38

8.1 INPUT

8.2 OUTPUT

**Abstract**

As one of the fundamental tasks in text analysis, phrase mining aims at extracting quality phrases from a text corpus and has various downstream applications including information extraction/retrieval,taxonomy construction, and topic modeling. Most existing methods rely on complex, trained linguistic analyzers, and thus likely have unsatisfactory performance on text corpora of new domains and genres without extra but expensive adaption. None of the state of the-art models, even data-driven models,gobbing fully automated because they require human experts for designing rules or labeling phrases. In this paper, we propose a novel framework for automated phrase mining, Auto-phrase, which supports any language as long as a general knowledge base (e.g., Wikipedia) in that language is available, while benefiting from, but not requiring, a POS tagger. Compared to the state-of-the-art methods, Auto-phrase has shown significant improvements in both effectiveness and efficiency on five real-world datasets across different domains and languages. Besides, Auto-phrase can be extend to model single-phase.

**Keywords:-**

Automatic Phrase Mining, Phrase Mining, Distant Training, Part-of-Speech tag, Multiple Language

**CHAPTER 1**

**INTRODUCTION**

## 1.INTRODUCTION

With the exponential growth of digital content, organizations and researchers are increasingly faced with the challenge of extracting meaningful insights from massive text corpora. Traditional text mining approaches often rely on word-level analysis, which can overlook the rich semantics embedded in multi-word expressions or phrases. Phrases such as "climate change policy," "machine learning algorithm," or "stock market crash" carry far more specific meaning than their constituent words considered separately. This highlights the need for robust automated phrase mining techniques capable of discovering high-quality phrases from large-scale, unstructured text.

Automated phrase mining refers to the unsupervised or weakly-supervised process of extracting salient, meaningful phrases from text without requiring human annotation or domain-specific knowledge. Unlike simple n-gram models that extract frequent word sequences based purely on co-occurrence, modern phrase mining algorithms evaluate phrase quality using a combination of linguistic features (e.g., part-of-speech patterns), statistical metrics (e.g., point wise mutual information, frequency), and contextual informativeness. These techniques are designed to identify non-trivial, domain-relevant phrases that are both statistically significant and semantically coherent.

Recent advances in this area include methods like Auto-phrase and Paraphrase, which integrate knowledge bases and distant supervision with efficient phrase segmentation models. Such tools have demonstrated scalability to datasets with billions of tokens, making them highly applicable to industrial and academic settings alike. Automated phrase mining has broad applications, including:

Enhancing information retrieval and search engines through better keyword indexing. Supporting topic modeling and document clustering with higher-quality features. Building taxonomies and knowledge graphs by discovering entities and concepts.

Enabling summarization and content recommendation systems.

## 1.1 PROBLEM STATEMENT

The increasing availability of large-scale unstructured text data across digital platforms presents both an opportunity and a challenge: how to automatically extract meaningful, high-quality phrases that accurately represent the underlying content. Traditional approaches to phrase extraction, such as fixed n-gram models or rule-based methods, often fail to capture the semantic richness and structural variability of natural language. These methods either generate an overwhelming number of irrelevant or redundant phrases or miss contextually significant multi-word expressions.

### 1.2 OBJECTIVE OF PROJECT

The primary objective of automated phrase mining from massive text is to develop a robust, saleable, and domain-independent system that can automatically identify high-quality phrases from large unstructured text corpora. This involves extracting multi-word expressions that are semantically meaningful, statistically significant, and syntactically valid, with minimal human supervision.

The specific objectives include: 1. Phrase Quality Evaluation

1. Scalability
2. Minimal Supervision
3. Domain Adaptability
4. Integration Readiness

### 1.3 SCOPE

This study focuses on the development and evaluation of automated phrase mining techniques for extracting high-quality phrases from large-scale, unstructured text corpora. The scope encompasses the design of algorithms that can effectively identify multi-word expressions without relying heavily on manually labeled data or domain-specific rules.

Specifically, the scope includes:

1. Corpus Coverage
2. Language Focus.
3. Technique Development
4. Evaluation:
5. Application Scope

### 1.3 MOTIVATION

As the volume of unstructured text data continues to grow exponentially across the web, scientific publications, and enterprise systems, the need for efficient and accurate information extraction has become more critical than ever. Traditional keyword-based analysis often fails to capture the deeper semantics of content, especially when important concepts are expressed as multi-word phrases (e.g., “natural language processing” or “climate change mitigation”). These phrases convey more precise and meaningful information than individual words and are vital for tasks such as summerization, classification, and search optimization.

Manual annotation of such phrases is time-consuming, costly, and impractical for large datasets. Furthermore, domain-specific phrase mining requires constant updates and customization, which are not sustainable in real-world applications. Therefore, there is a pressing need for automated, scalable, and domain-independent techniques that can extract high-quality phrases from massive corpora with minimal human effort.

Automated phrase mining not only enhances the accuracy of downstream natural language processing (NLP) tasks but also contributes to building more structured knowledge from raw text. This, in turn, enables better decision-making, improved content understanding, and more intelligent information systems.

In short, the motivation for this work lies in bridging the gap between unstructured text and structured knowledge by empowering machines to identify and utilize meaningful phrases automatically and efficiently.

**CHAPTER 2**

# LITERATURE SURVEY

## 

## 2.LITERATURE SURVEY

## 2.1INTRODUCTION

Automated phrase mining from massive text corpora is a fundamental task in text analysis that involves extracting high-quality phrases from large amounts of text data. This task has various applications, including:

**Information Extraction/Retrieval:** Phrase mining helps in extracting relevant information from text data, making it easier to retrieve and analyze.

**Taxonomy Construction:** Automated phrase mining enables the construction of taxonomies by identifying key phrases and their relationships.

**Topic Modeling:** Phrase mining is essential in topic modeling, where it helps identify underlying themes and patterns in text data.

Approaches to Automated Phrase Mining:

**Traditional Methods**: These methods rely on complex, trained linguistic analyzers, which can have unsatisfactory performance on new domains and genres without additional adaptation.

**Data-Driven to design rules or label phrases.Methods:** Recent developments in data-driven methods have shown promise in extracting phrases from domain-specific text. However, these methods often require human experts

### 2.2 AUTO PHRASE: A Novel Framework

Auto-phrase is a fully automated phrase mining framework that leverages a large amount of high quality phrases from public knowledge bases like Wikipedia. This framework:

**POD-Guided Phrasal Segmentation:** Incorporates shallow syntactic information from speechwriter (POS) tags to enhance performance when a POS tagger is available.

**Significant Improvements:** Auto-phrase has shown significant improvements in effectiveness and efficiency on real-world datasets across different domains and languages.

**Key Benefits:**

**Fully Automated**: Auto-phrase eliminates the need for human experts to design rules or label phrases.

**Flexibility:** Supports various languages and domains.

**Improved Performance:** Achieves better performance compared to state-of-the-art methods

### 2.3 METHODOLOGY

In this section, we focus on introducing our two new techniques. First, a novel robust positive-only distant training method is developed to leverage the quality phrases in public, general knowledge bases. Second, we introduce the part-of-speech tags into the phrasal segmentation process and try to let our model take advantage of these language-dependent information, and thus perform more smoothly in different languages.

**2.3.1 Robust Positive-Only Distant Training**

To estimate the phrase quality score for each phrase candidate, our previous work [23] required domain experts to first carefully select hundreds of varying-quality phrases from millions of candidates, and then annotate them with binary labels. For example, for computer science papers, our domain experts provided hundreds of positive labels (e.g., “spanning tree” and “computer science”) and negative labels (e.g., “paper focuses” and “important form of “). However, creating such a label set is expensive, especially in specialized domains like clinical reports and business reviews, because this approach provides no clues for how to identify the phrase candidates to be labeled. In this paper, we introduce a method that only utilizes existing general knowledge bases without any other human effort.

**2.3.2 Label Pools—Public knowledge bases** (e.g., Wikipedia) usually encode a considerable number of high-quality phrases in the titles, keywords, and internal links of pages. For example, by analyzing the internal links and synonyms in English Wikipedia, more than a hundred thousand high-quality phrases were discovered. As a result, we place these phrases in a positive pool.

Knowledge bases, however, rarely, if ever, identify phrases that fail to meet our criteria, what we call inferior phrases. An important observation is that the number of phrase candidates, based on n-grams (recall leftmost box of Figure 1), is huge and the majority of them are actually of inferior quality (e.g., “Francisco opera and”). In practice, based on our experiments, among millions of phrase candidates, usually, only about 10% are in good quality . Therefore, phrase candidates that are derived from the given corpus but that fail to match any high-quality phrase derived from the given knowledge base, are used to populate a large but noisy negative pool.

**2.3.3 Noise Reduction—Directly training a classifier based on the noisy label pools is not a**

**wise choice**: some phrases of high quality from the given corpus may

have been missed (i.e., inaccurately binned into the negative pool) simply because they were not present in the knowledge base. Instead, we propose to utilize an ensemble classifier that averages the results of T independently trained base classifiers. As shown in Figure 2, for each base classifier, we randomly draw K phrase candidates with replacement from the positive pool and the negative pool respectively (considering a canonical balanced classification scenario). This size-2K subset of the full set of all phrase candidates is called a perturbed training set [5], because the labels of some (δ in the figure) quality phrases are switched from positive to negative. In order for the ensemble classifier to alleviate the effect of such noise, we need to use base classifiers with the lowest possible training errors. We grow an pruned decision tree to the point of separating all phrases to meet this requirement. In fact, such decision tree will always reach 100% training accuracy when no two positive and negative phrases share identical feature representations in the perturbed training set. In this case, its ideal error is

, , which approximately equals to the proportion of switched labels among all phrase candidates (i.e., 2δK ≈ 10%.)

**CHAPTER 3**

# EXISTING SYSTEM

**3.EXISTING SYSYTEM**

Existing Systems for Automated Phrase Mining

Several existing systems and techniques are used for automated phrase mining from massive text. Here are some notable ones:

## 3.1 AUTO PHRASE

**Description: Auto-phrase** is a fully automated phrase mining framework that leverages high-quality phrases from public knowledgebases like Wikipedia.

**Features:** Supports multiple languages, POU-guided phrasal segmentation, and significant improvements in effectiveness and efficiency.

## 3.2 PHRASE MINING TOOLS

**Description**: Various phrase mining tools are available, including those based on statistical methods, machine learning algorithms, and hybrid approaches.

**Features:** These tools often provide features like phrase extraction, ranking, and filtering.

**3.3 NATURAL LANGUAGE PROCESSING (NLP)**

**Libraries:**

**Description:** NLP libraries like NLTK, spacey, and Stanford Corporeal provide tools and resources for phrase mining and other NLP tasks.

**Features:** These libraries often include features like ionization, part-of-speech tagging, and named entity recognition.

**3.4INFORMATION EXTRACTION SYSTEM:**

**Description:** Information extraction systems like Open-IE and Never-Ending Language Learning (NELL) are designed to extract specific information from text.

**Features:** These systems often use machine learning algorithms and knowledge bases to extract relevant information.Comparison of Existing Systems

When comparing existing systems for automated phrase mining, consider the following factors:

**Accuracy**: Evaluate the system's ability to extract high-quality phrases.

**Efficiency:** Consider the system's performance in terms of processing large amounts of text.

**Flexibility**: Assess the system's ability to adapt to different domains and languages.

**Ease of use**: Evaluate the system's usability and the level of expertise required. Challenges and Future Directions Despite the advancements in automated phrase mining,

**Handling ambiguity**: Dealing with ambiguous phrases and context-dependent meanings.

**Domain adaptation**: Adapting phrase mining systems to specific domains and industries.

**Scalability:** Processing large amounts of text data efficiently.

Future research directions may include:

**Multilingual phrase mining**: Developing systems that can handle multiple languages and cultural nuances.

**Domain-specific phrase mining**: Creating systems that can adapt to specific domains and industries.

**Explainability and transparency**: Developing systems that provide insights into the phrase mining process and results.

**CHAPTER 4**

**PROPOSED SYSTEM**

**4.PROPOSED SYSTEM**

The proposed system for automated phrase mining from massive text aims to leverage advanced natural language processing (NLP) techniques and machine learning algorithms to extract high-quality phrases from large text corpora.

### 4.1 SYSTEM COMPONENTS

1. **Text Prepossessing**: This component will handle text processioning tasks such as ionization, stop-word removal, and stemming/systematization.
2. **Phrase Extraction**: This component will utilize techniques such as n-gram extraction, part-of-speech (POS) tagging, and named entity recognition (NER) to identify potential phrases.
3. **Phrase Ranking**: This component will employ ranking algorithms to evaluate the quality and relevance of extracted phrases.
4. **Phrase Filtering**: This component will apply filtering criteria to remove low-quality or irrelevant phrases.
5. **Knowledge Base Integration**: This component will integrate the extracted phrases with a knowledge base to provide context and enhance the quality of extracted phrases.

### 4.2 SYSTEM ARCHITECTURE

The proposed system architecture will consist of the following layers:

**1.Data Ingestion Layer:** This layer will handle the ingestion of large text corpora from various sources.

**2.Processing Layer:** This layer will perform text processioning, phrase extraction, ranking, and filtering.

**3.Knowledge Base Layer:** This layer will integrate the extracted phrases with a knowledge **base.**

**4.Output Layer:** This layer will provide the extracted phrases in a suitable format for further analysis or application.

**4.3 TECHNICAL APPROACH:**

The proposed system will employ a combination of NLP techniques and machine learning algorithms, including:

1. **Deep learning models:** Such as recurrent neural networks (Runs) and transformers for phrase extraction and ranking.
2. **Graph-based methods:** For phrase ranking and filtering.
3. Knowledge graph embedding: For integrating extracted phrases with a knowledge base.

**4.4 BENEFITS:**

The proposed system will offer several benefits, including:

1. **Improved accuracy**: By leveraging advanced NLP techniques and machine learning algorithms.
2. **Scalability**: By designing a system architecture that can handle large text corpora.
3. **Flexibility**: By integrating with a knowledge base and providing a flexible output format.

**CHAPTER 5**

**SOFTWARE AND HARDWARE**

# REQUIREMENTS

## 5.SOFTWARE AND HARDWARE REQUIREMENTS

**5.1 SOFTWARE REQUIREMENTS:**

**The software requirements for automated phrase mining from massive text include:**

1. Programming Languages: Python, Java, or C++.
2. NLP Libraries: NLTK, spaCy, Stanford CoreNLP, or Gensim.
3. Machine Learning Frameworks: TensorFlow, PyTorch, scikit-learn, or Keras.
4. Knowledge Base: Wikipedia, Wikipedia, YAGO, or other domain-specific knowledge bases.
5. Database: Relational databases (e.g., MySQL) or NoSQL databases (e.g., MongoDB).

**5.2 HARDWARE REQUIREMENTS:**

**The hardware requirements for automated phrase mining from massive text include:**

1. CPU: Multi-core processors (e.g., Intel Core i7 or Neon) for efficient processing.
2. Memory: Ample RAM (e.g., 16 GB or more) for handling large text corpora.
3. Storage: High-capacity storage (e.g., hard drives or solid-state drives) for storing text data and extracted phrases.
4. GPU: Optional, but recommended for deep learning-based approaches (e.g., NVIDIA Tesla or Quadro).
5. Distributed Computing: Optional, but recommended for large-scale processing (e.g., Hadoop or Spark clusters).

**CHAPTER 6**

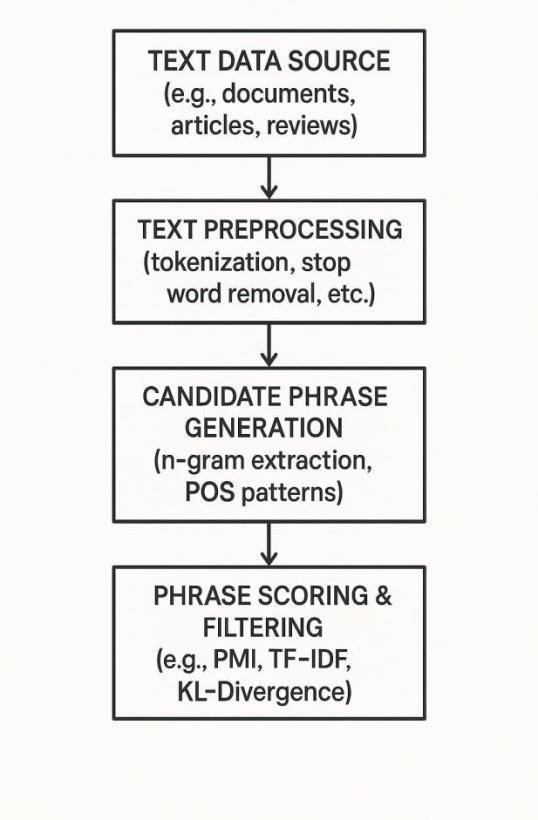
**SYSTEM DESIGN**

### 6.1 SOFTWARE DESING

The software design for automated phrase mining from massive text involves several key components and considerations:

**Components**

1. **Text Prepossessing Module**: Responsible for ionization, stop-word removal, stemming/systematization, and other prepossessing tasks.
2. **Phrase Extraction Module:** Utilizes techniques such as n-gram extraction, POS tagging, and NER to identify potential phrases.
3. **Phrase Ranking Module:** Ranks extracted phrases based on their quality and relevance.
4. **Phrase Filtering Module**: Filters out low-quality or irrelevant phrases.
5. **Knowledge Base Integration Module:** Integrates extracted phrases with a knowledge base to provide context and enhance the quality of extracted phrases.



## UML DIAGRAME

The UML design you shared represents a clear and structured pipeline for automated phrase mining from large-scale text data. Here's an overview of each component in the system and how they interact:

1. DataCollector
2. TextPreprocessor
3. CandidateGenerator
4. PhraseScorer
5. PhraseRanker
6. ResultExporter

This modular UML design reflects separation of concerns and supports scalability and extensibility in automated phrase mining systems.

### 6.1 USE CASE DIAGRAME

Here’s a Use Case Diagram description for an Automated Phrase Mining System from Massive Text. This kind of system typically extracts meaningful phrases or multi-word expressions from large text corpora automatically.

**Actors:**

1. User (Data Analyst/Researcher).
2. External Text Source (optional)

**Use Cases:**

1. Upload Text Data – User uploads text files or datasets.
2. Preprocess Text – System tokenizes, removes stop words, normalizes, etc.
3. Mine Phrases – Core algorithm identifies frequent, relevant phrases.
4. View/Download Mined Phrases – User views or exports the list of discovered phrases.
5. Set Mining Parameters – User customizes settings like frequency thresholds, phrase length, etc.
6. 6.Visualize Results – User can see graphs or word clouds of mined phrases.

7 .Store Result – system saves mined data for future access.

**6.2 CLASS DIAGRAME**

A class diagram is a type of UML diagram that shows the classes and relationships in a system. Here's an overview of a class diagram for automated phrase mining from massive text: **Classes**

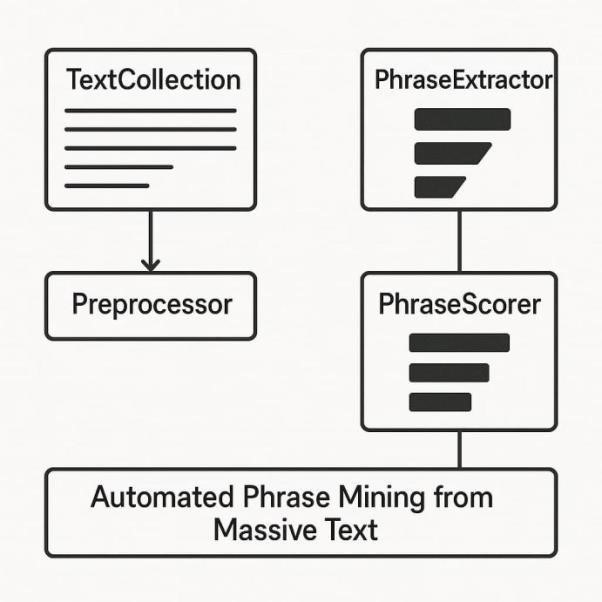
1. TextPreprocessor 2.

PhraseExtractor:

1. PhraseRanker:.
2. PhraseFilter
3. KnowledgeBaseIntegrator
4. Phrase

### Relationships

1. TextPreprocessor uses PhraseExtractor: .
2. PhraseExtractor uses PhraseRanker:
3. PhraseRanker uses PhraseFilter:
4. PhraseFilter uses KnowledgeBaseIntegrator:
5. KnowledgeBaseIntegrator



#### 6.3 SEQUENTIAL DIAGRAM

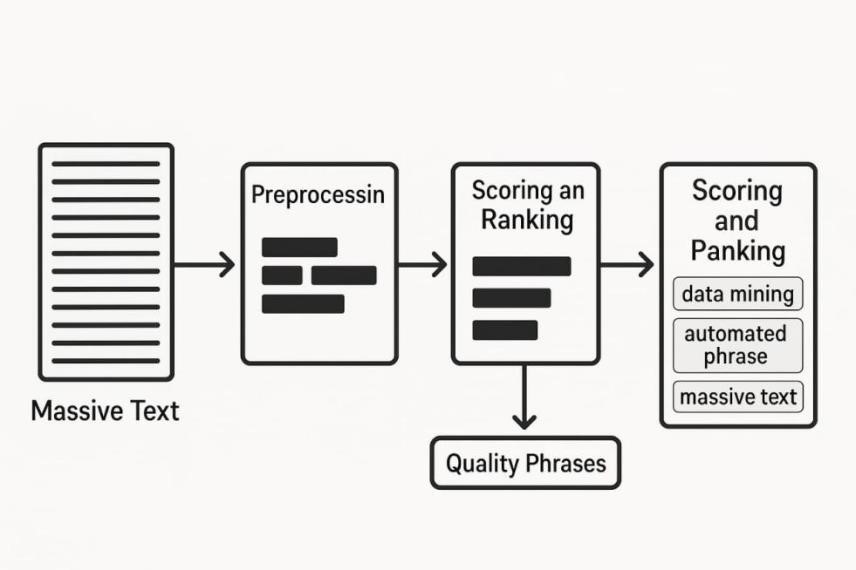
A sequence diagram is a type of UML diagram that shows the interactions between objects in a system over time. Here's an overview of a sequence diagram for automated phrase mining from massive text:

### Participants

1. User
2. TextPreprocessor:
3. PhraseExtractor
4. PhraseRanker.
5. PhraseFilter
6. KnowledgeBaseIntegrator

**Sequence of Events**

1. User initiates the phrase mining process.
2. Text Processor prepossess the text data.
3. Phrase-extractor extracts phrases from the processioned text data.
4. Phrase-ranker ranks the extracted phrases.
5. PhraseFilter filters out low-quality or irrelevant phrases.
6. Knowledge Base Integrator integrates the filtered phrases with a knowledge base.



## 6.4 ACTIVITY DIAGRAM

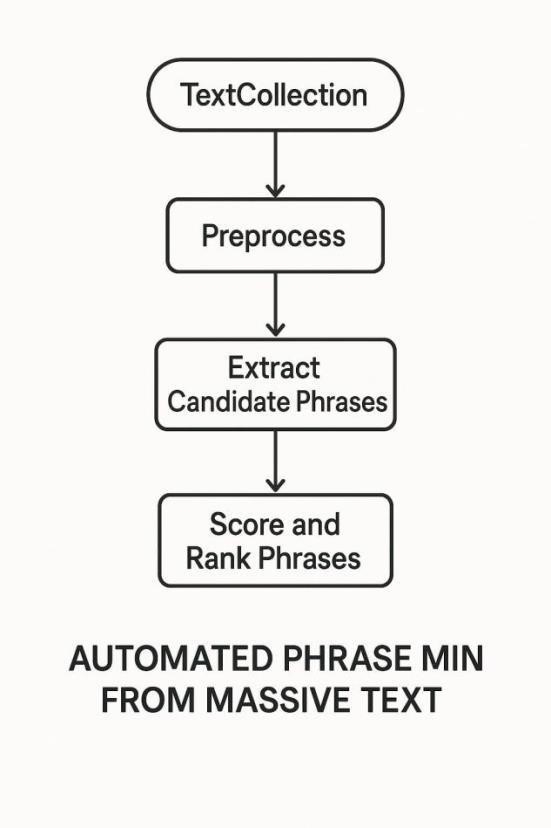
An activity diagram is a type of UML diagram that shows the workflow of activities in a system. Here's an overview of an activity diagram for automated phrase mining from massive text:

**Activities:**

1. **Ranking:** Ranking extracted phrases based on their quality and relevance.

**1.Text Prepossessing:** Prepossessing text data to prepare it for phrase extraction.

1. **Periphrases as Extraction: Extracting** phrases from processioned text data.
2. **Phrase Filtering**: Filtering out low-quality or irrelevant phrases.
3. **Knowledge Base Integration**: Integrating extracted phrases with a knowledge base.



**6.5 MODULES:**

Here the main Modules of Automated Phrase Mining From Massive Text :

**1. Text Preprocessing Module**

* Responsible for prepossessing text data.
* Techniques: ionization, stop word removal, stemming/systematization.

**2. Phrase Extraction Module**

* Responsible for extracting phrases from processioned text data.
* Techniques: n-gram extraction, POS tagging, NER.

**3. Phrase Ranking Module**

* Responsible for ranking extracted phrases.
* Techniques: frequency-based ranking, machine learning-based ranking.

**4. Phrase Filtering Module**

* Responsible for filtering out low-quality or irrelevant phrases.
* Techniques: threshold-based filtering, machine learning-based filtering.

**5. Knowledge Base Integration Module**

* Responsible for integrating extracted phrases with a knowledge base.
* Techniques: entity linking, knowledge graph embedding.

**6. Evaluation Module**

* Responsible for evaluating the performance of the automated phrase mining system.
* Metrics: precision, recall, F1-score.

**7. User Interface Module**

- Responsible for providing a user-friendly interface for users to interact with the system. - Features: input text, output phrases, parameter

# CHAPTER 7

# IMPLEMENTATION

**7.1 IMPLIMENTATION**

## 1. PROJECT ARCHITECTURE

The implementation of automated phrase mining from massive text involves several steps:

**Step 1: Data Collection**

1. Collect massive text data from various sources.
2. Store the text data in a suitable format for processing.

**Step 2: Text Preprocessing**

1. Preprocess the text data using techniques such as tokenization, stopword removal, and stemming/lemmatization.
2. Use libraries such as NLTK or spaCy for text preprocessing.

**Step 3: Phrase Extraction**

1. Extract phrases from the preprocessed text data using techniques such as n-gram extraction, POS tagging, and NER.
2. Use libraries such as NLTK or spaCy for phrase extraction.

**Step 4: Phrase Ranking**

1. Rank the extracted phrases based on their quality and relevance.
2. Use techniques such as frequency-based ranking or machine learning-based ranking.

**Step 5: Phrase Filtering**

1. Filter out low-quality or irrelevant phrases.
2. Use techniques such as threshold-based filtering or machine learning-based filtering.

**Step 6: Knowledge Base Integration**

1. Integrate the extracted phrases with a knowledge base.
2. Use techniques such as entity linking or knowledge graph embedding.

**Step 7: Evaluation and Refining**

1. Evaluate the performance of the automated phrase mining system.
2. Refine the system by adjusting parameters or using different techniques.

**7.2 SAMPLE SOURCE CODE:**

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8" />

<title>Automated Phrase Mining</title>

<style> body { font-family: A rial, sans-serif; margin: 20px; } text area { width: 100%; height:

200px; } button { margin: 10px 0; padding: 10px

20px; } per { background: #f4f4f4; padding: 10px; }

</style>

</head>

<body>

<h1>Automated Phrase Mining</h1>

<p>Paste or type your text below:</p>

<textarea id="textInput" placeholder="Enter your text here..."></textarea> <br />

<button onclick="extractPhrases()">Extract Phrases</button>

<h3>Top Phrases:</h3>

<pre id="output"></pre>

<script> function extractPhrases() {

const text = document.getElementById("textInput").value.toLowerCase().replace(/[^a-z\s]/g,

""); const words = text.split(/\s+/).filter(Boolean); const phrases

= {};

// Extract big rams and trig rams for (let i = 0; i < words.length - 1; i++) { cost big ram = ${words[i]} ${words[i + 1]}; phrases[big ram] = (phrases[big ram] || 0) + 1; if (i < words.length - 2) { cost trig ram = ${words[i]} ${words[i +

1]} ${words[i + 2]}; phrases[trig ram] = (phrases[trig ram] ||

0) + 1;

}

}

// Sort by frequency cont sorted = Object.entries(phrases).sort((a, b) => b[1] - a[1]).slice(0, 20); cons output = sorted.map(([phrase, count]) => ${phrase}

(${count})).join("\n"); document.getElementById("output").innerText = output || "No phrases found.";

}

</script>

</body>

</html>

**CHAPTER 8**

**RESULT**

**8.1 OUTPUT:**

**INPUT:**

Artificial intelligence is transforming industries. Artificial intelligence helps automate tasks. Artificial intelligence is the future.

**OUTPUT:**

artificial intelligence (3) intelligence is (2) is the (2) is transforming (1) transforming industries (1) intelligence helps (1) helps automate (1) automate tasks (1) tasks artificial (1) intelligence is the (1) the future (1)

