Summiva: An Enterprise-Scale NLP System for Content Summarization, Tagging, and Search

Saikumar Bollam

University of Arizona, College of Information Science

Mentor: Dr. Liangming Pan
Instructor: Dr. Greg Chism

Keywords: NLP, Text Summarization, Content Tagging, Enterprise Search, Topic Modeling, Grouping, Scalable AI

1. Abstract

The exponential growth of **unstructured text data** presents challenges in **information retrieval, summarization, structured tagging, and content grouping at an enterprise scale**. In the **digital era**, the **sheer volume of online information** can be overwhelming. **Professionals, researchers, and casual readers** often need **quick insights** from web pages or articles **without reading every word**.

We introduce **Summiva**, a modular NLP system designed for **enterprise-level document summarization, intelligent tagging, structured grouping, and scalable search**. Summiva integrates **state-of-the-art summarization algorithms**, **adaptive topic modeling**, and **high-performance search indexing** to process and store large volumes of web content. Unlike conventional consumer-oriented summarization tools, **Summiva focuses on structured storage and retrieval**, ensuring **scalability, efficiency, and local data privacy**.

Our approach evaluates multiple **summarization, tagging, and grouping techniques**—including **transformer-based models (T5, BART, PEGASUS), graph-based clustering, and deep-learning-driven entity recognition**—to determine the most effective solutions for **large-scale deployment**. The system is **optimized for enterprise search**, allowing rapid retrieval of processed text using **modern indexing solutions (Elasticsearch, FAISS, MeiliSearch)**.

This proposal outlines **Summiva's system architecture, backend and frontend technical implementation, experimental evaluation, real-world enterprise applications, and a structured 3-month project timeline with backlog flexibility**.

---

2. Introduction

2.1 Problem Statement

The **explosive growth of digital text data** has made it increasingly difficult for enterprises to **extract, store, and retrieve meaningful insights** from **unstructured text sources**. **Efficient summarization, tagging, grouping, and search** capabilities are **critical** for organizations managing **large-scale text repositories**.

Traditional summarization models primarily focus on **extractive or abstractive techniques**, but they **lack structured tagging, advanced grouping methods, and enterprise-ready search capabilities**. Additionally, enterprise environments require **modular, scalable, and locally deployable** architectures to ensure **privacy, efficiency, and compliance**.

2.2 Contributions

Summiva addresses these challenges by providing an **intelligent NLP-driven framework** that:

1. **Extracts meaningful content** from any **URL**.
2. **Generates a concise summary** using **multi-algorithm summarization**.
3. **Tags and groups key concepts** using **state-of-the-art entity recognition, clustering, and topic modeling**.

4. **Provides enterprise-scale search capabilities**, leveraging **high-performance indexing**.
5. **Offers a flexible, locally deployable system** to **ensure privacy and efficiency**.

Summiva is designed to be **modular, scalable, and enterprise-ready**, making **large-scale text processing fast, structured, and accessible**.

---

3. Project Plan & Timeline

3.1 Development Schedule (3-Months, 4-Hours per Day, Backlog Flexibility)

To efficiently build **Summiva**, we define **a structured development plan** that prioritizes **core functionalities** while maintaining **flexibility for backlog tasks**.

Week	Task	Hours (Weekly)	Notes & Backlog Considerations
Week 1-2	**Project Setup & Core Infrastructure**	20	Set up **GitHub repo, database (PostgreSQL)**, define API architecture. If delays occur, backlog shifts to Week 3.
Week 3-4	**Content Fetching & Cleaning**	20	Implement web scraping with **BeautifulSoup & Newspaper3k**. Store cleaned text in database.
Week 5-6	**Summarization Engine**	20	Implement **extractive (TextRank, BERT) & abstractive (T5, PEGASUS)** summarization. Adjust models based on efficiency.

Week 7-8	**Tagging & Grouping**	20	Apply **NER (BERT, SpaCy), topic modeling (LDA, BERTopic), clustering (K-Means, HDBSCAN)** for structured metadata.
Week 9-10	**Enterprise Search & Indexing**	20	Build **keyword search (Elasticsearch) & semantic search (FAISS)**. Prioritize keyword search if time is tight.
Week 11-12	**Frontend & UI (Optional, If Time Permits)**	20	If ahead of schedule, create a **React-based UI**. If delayed, focus on API improvements.
Week 13-14	**Testing & Optimization**	20	Optimize search performance, fix latency issues. Stress test with **large datasets**.
Week 15-16	**Deployment & Documentation**	20	Prepare final **report, GitHub documentation, and conference submission**.

---

4. System Architecture

4.1 Overview

Summiva consists of **four core components**:

1. **Summarization Engine** (extractive + abstractive approaches)
2. **Tagging & Grouping** (NER, clustering-based grouping)

3. **Enterprise-Ready Search** (keyword + semantic indexing)
4. **Web-Based User Interface** (optional frontend)

4.2 Grouping Implementation

Grouping in Summiva is implemented through **three primary methods**:

- **Topic Modeling**: **LDA, BERTopic, NMF** for organizing documents into meaningful clusters.
- **Clustering Similar Content**: **K-Means, DBSCAN, Spectral Clustering** for grouping related documents.
- **Named Entity-Based Grouping**: **NER (BERT, BiLSTM-CRF, SpaCy)** to categorize documents by key entities.

Grouping enhances **search retrieval, structured indexing, and contextual relevance** within enterprise applications.

---

5. Experimental Setup & Evaluation

5.1 Datasets

Summiva is evaluated on:

- **CNN/DailyMail dataset** for summarization.
- **Wikipedia Named Entity Recognition Dataset** for tagging.
- **BBC News & Reuters Dataset** for topic modeling and grouping.

5.2 Evaluation Metrics

- **Summarization Metrics**: ROUGE-1, ROUGE-2, ROUGE-L.
- **Tagging Metrics**: F1-score, precision-recall.
- **Grouping & Clustering Metrics**: Silhouette Coefficient, Adjusted Rand Index (ARI).
- **Search & Retrieval Metrics**: NDCG (Normalized Discounted Cumulative Gain).

5.3 Results & Discussion

---

6. Conclusion & Future Work

Summiva demonstrates the feasibility of a **scalable, enterprise-ready NLP system** that integrates **summarization, tagging, grouping, and search** into a unified framework.

Future work includes:

- **Scaling Summiva to multi-document summarization**.
- **Optimizing for real-time processing** in large-scale corporate environments.
- **Exploring reinforcement learning for search optimization**.

Date: 01/31/2025

Student Signature :

B. SAI KUMAR

Mentor Signature :

Instructor Signature: