Semantic Keywords Extraction System

Project Overview

This project implements a comprehensive semantic keyword extraction system that combines three different methodologies: TF-IDF statistical analysis, semantic embedding-based extraction, and frequency analysis. The system processes knowledge base documents and provides intelligent keyword identification for RAG (Retrieval-Augmented Generation) applications.

Project Objectives

The primary goal of this research is to develop an advanced keyword extraction pipeline that can:

- **Extract meaningful keywords** from technical documentation using multiple approaches
- **Compare effectiveness** of traditional statistical methods versus modern semantic techniques
- **Implement RAG capabilities** for intelligent document retrieval and query processing
- **Evaluate system performance** across different query types and extraction methodologies

Problem Statement

Traditional keyword extraction methods often fail to capture contextual relationships and semantic meanings in technical documents. This project addresses the challenge by implementing a multi-method approach that combines:

1. Statistical importance through TF-IDF analysis
2. Contextual understanding through semantic embeddings
3. Basic vocabulary mapping through frequency analysis

Methodology and Architecture

Technical Approach

The system employs a hybrid architecture that processes knowledge items through multiple extraction pipelines simultaneously, enabling comparative analysis and complementary keyword identification.

Core Components

- **Multi-Method Keyword Extractor**: Implements three distinct algorithms for comprehensive keyword coverage
- **RAG System**: Provides semantic search and document retrieval with confidence scoring
- **Evaluation Framework**: Measures and compares performance across different methodologies

Tools and Technologies Used

Programming Environment

- **Python 3.11**: Primary programming language
- **Jupyter Notebook**: Development and execution environment
- **Sentence-Transformers 4.1.0**: Semantic embeddings using 'all-MiniLM-L6-v2'
- **Scikit-learn 1.2.2**: TF-IDF vectorization and cosine similarity
- **PyTorch 2.6.0**: Deep learning framework backend

Dataset Information

- **Source**: Sample RAG Knowledge Item Dataset
- **Total Articles**: 10 knowledge items
- **Generated Chunks**: 30 text segments
- **Processing Strategy**: 200-word chunks with 50-word overlap for context preservation

Method	**Top Keyword**	**Score**	**Rank 2**	**Score**	**Rank 3**	**Score**
TF-IDF	email	0.0888	step	0.0653	company	0.0600
SEMANTIC	connect company email	0.7078	email account mobile	0.6976	synchronize company email	0.6677
FREQUENCY	your	140	step	89	email	71

Table 1: Keyword Extraction Methods Comparison

Caption: Direct comparison of three keyword extraction methods showing actual results from the system

Table 2: RAG System Performance Results

Query ID	Confidence Score	RAG Keywords (Top 3)	Performance Status
Test Query	0.678	connect company email (0.708), email account mobile (0.698), use company email (0.674)	Best Performance
Query 1	0.678	-	Excellent
Query 2	0.481	-	Fair
Query 3	0.583	-	Good
Query 4	0.529	-	Acceptable
Query 5	0.537	-	Acceptable
Average	**0.562**	-	**Overall Good**

Caption: RAG system confidence scores across different queries with actual performance metrics

Applications and Use Cases

Primary Applications

- **Technical Documentation Search**: Automated keyword identification for troubleshooting guides
- **Knowledge Management Systems**: Enterprise document classification and retrieval

- **Customer Support Automation**: Intelligent query processing and response generation
- **Content Analysis**: Semantic understanding of technical documentation patterns

Research Contributions

This project demonstrates the effectiveness of combining traditional statistical methods with modern semantic approaches for keyword extraction in technical domains. The research provides valuable insights into the comparative performance of different extraction methodologies.

Key Results Summary

- **Best Semantic Score**: 0.7078 (connect company email)
- **Highest TF-IDF**: 0.0888 (email)
- **Most Frequent Term**: "your" (140 occurrences)
- **RAG Average Confidence**: 0.562
- **Best Query Performance**: 0.678
- **System Status**: All methods working successfully

Conclusion

The system demonstrates successful implementation of multi-method keyword extraction with the semantic approach achieving the highest relevance scores, while the RAG system maintains consistent performance across different query types. This comprehensive approach provides a robust foundation for intelligent document processing and retrieval applications in technical domains.

The integration of multiple extraction methodologies offers complementary insights: statistical importance through TF-IDF, contextual understanding through semantic embeddings, and basic vocabulary analysis through frequency counting. This multi-faceted approach ensures comprehensive keyword coverage suitable for various application requirements.