Milestone -2 Report

Priya Poosaala

Rollno: 26

1.Project Overview

The AI Skill Gap Analyzer (Milestone 2) is a web application designed to automate the extraction, analysis, and visualization of both technical and soft skills from unstructured text sources such as resumes and job descriptions. Building on the foundation of Milestone 1, this version leverages state-of-the-art Natural Language Processing (NLP) and Machine Learning (ML) techniques, including spaCy pipelines, custom Named Entity Recognition (NER), and semantic embeddings via Sentence-BERT. The system provides users with a comprehensive, interactive platform to identify skill gaps, analyze skill distributions, and support data-driven decisions in recruitment, upskilling, and workforce planning.

2.Objectives

- Automate skill extraction from resumes, job descriptions, or any unstructured text.
- Support multiple extraction methods: keyword matching, POS patterns, context-based, NER, and noun chunking.
- Enable semantic skill analysis using BERT embeddings for similarity and matching.
- Allow users to annotate and train custom NER models for domain-specific skill detection.
- Visualize skill distributions and relationships with interactive charts (pie, bar, radar, treemap, sunburst).
- Export results in CSV, JSON, and text formats for further analysis or reporting.
- Provide a user-friendly, interactive UI with Streamlit, supporting real-time feedback and multi-tab navigation.

3.Data for Skill Gap Analyzer

The system processes the following data:

- **Resume:** Uploaded by the user (candidate) as plain text (TXT) or **Job Description (JD):** Provided by employers or users in text format.
- **Skill Dictionary:** A comprehensive, categorized list of technical and soft skills, including abbreviations and synonyms, is used for matching and extraction.
- **Annotations:** User-generated labeled data for training custom NER models.

The application matches extracted skills from the input documents against the predefined skill dictionary, supporting both direct matches and expanded forms (e.g., abbreviations).

4. Implementation Details

The AI Skill Gap Analyzer operates as a multi-tab Streamlit web application, guiding the user through the following workflow:

1. **Skill Extraction Tab**
    - Users input or upload text (resume or job description).
    - The system preprocesses the text using spaCy, removing noise and tokenizing sentences.
    - Multiple extraction methods are applied in parallel:
        - Keyword Matching: Direct lookup of skills from a curated database.
        - POS Pattern Extraction: Identifies skills based on grammatical patterns (e.g., adjective + noun).
        - Context-Based Extraction: Uses regular expressions to find skills in context (e.g., "experience in Python").
        - NER-Based Extraction: Leverages spaCy's NER to detect skill entities.
        - Noun Chunking: Extracts potential skills from noun phrases.
    - Extracted skills are normalized (abbreviations expanded, duplicates removed) and categorized.
    - Confidence scores are assigned based on the number of methods that detected each skill.
    - Results are displayed as metrics, categorized skill lists, and extraction method statistics.
2. **BERT Embeddings Tab**
    - Generates semantic embeddings for each extracted skill using Sentence-BERT.
    - Allows users to compute similarity between any two skills.
    - Provides a tool to find the most semantically similar skills to a selected target.
    - Displays a similarity heatmap for the top-N skills.
3. **Custom NER Training Tab**
    - Users can load annotated data (from the annotation tab or upload).
    - Configure training parameters (number of epochs).
    - Train a custom spaCy NER model to recognize skills in text.
    - Visualizes training loss and allows testing the trained model on new text.
4. **Annotation Tab**
    - Provides an interface for users to annotate skills in sample text.
    - Supports adding, editing, and removing skill annotations.
    - Allows exporting annotations in JSON or spaCy format for NER training.
5. **Visualizations Tab**
    - Presents interactive charts:
        - Pie Chart: Skill distribution by category.
        - Bar Chart: Top skills by confidence.
        - Radar Chart: Skill counts across major categories.
        - Treemap: Hierarchical view of categories and sample skills.
        - Sunburst: Interactive drill-down from categories to skills.
    - Includes a detailed, filterable skill table.
6. **Export Tab**
    - Users can export extracted skills and statistics as CSV, JSON, or formatted text reports.

5. Text Processing Flow

1. **Text Input:** Users paste or upload resume and job description text.
2. **Preprocessing:** Text is cleaned, tokenized, and processed with spaCy, including custom stop word handling.
3. **Skill Extraction:** Multiple methods are applied in parallel to maximize recall and precision.
4. **Skill Normalization**: Abbreviations and synonyms are expanded for consistent matching.
5. **Categorization:** Extracted skills are mapped to predefined categories.
6. **Confidence Scoring:** Each skill is assigned a confidence score based on detection methods.
7. **Visualization:** Results are presented with interactive charts and tables.
8. **Export:** Users can download results in various formats.

6. Features Implemented

- **Multi-method skill extraction** (keyword, POS, context, NER, noun chunks).
- **Customizable skill database** with categories and abbreviations.
- **Advanced text preprocessing** with spaCy.
- **Semantic skill embeddings** and similarity analysis using Sentence-BERT.
- **Custom NER annotation and training interface.**
- **Interactive visualizations:** Pie, bar, radar, treemap, and sunburst charts.
- **Export options:** CSV, JSON, and text reports.
- **User interface:** Multi-tab Streamlit app with real-time feedback and session state management.

7.Statistics Computation

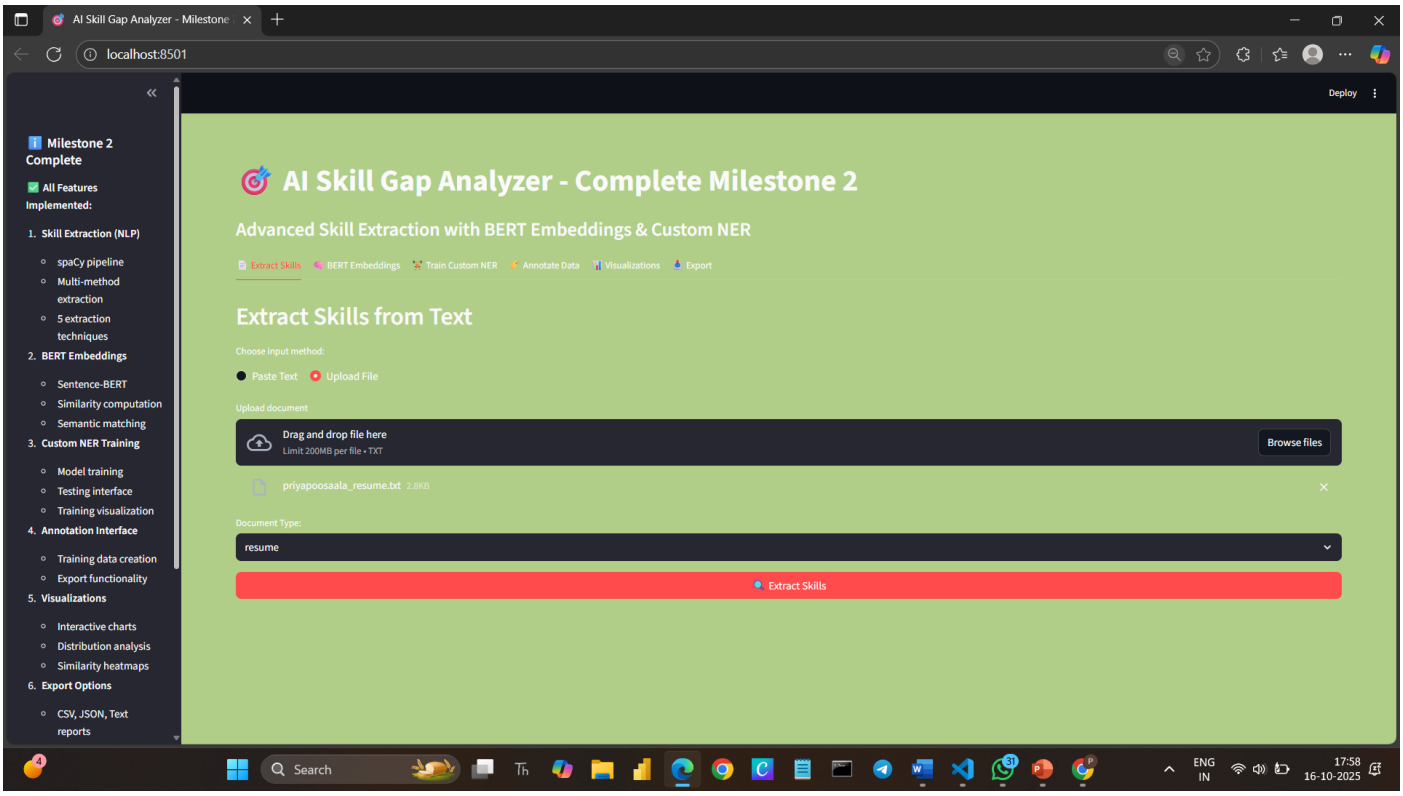
The analyzer computes and displays:

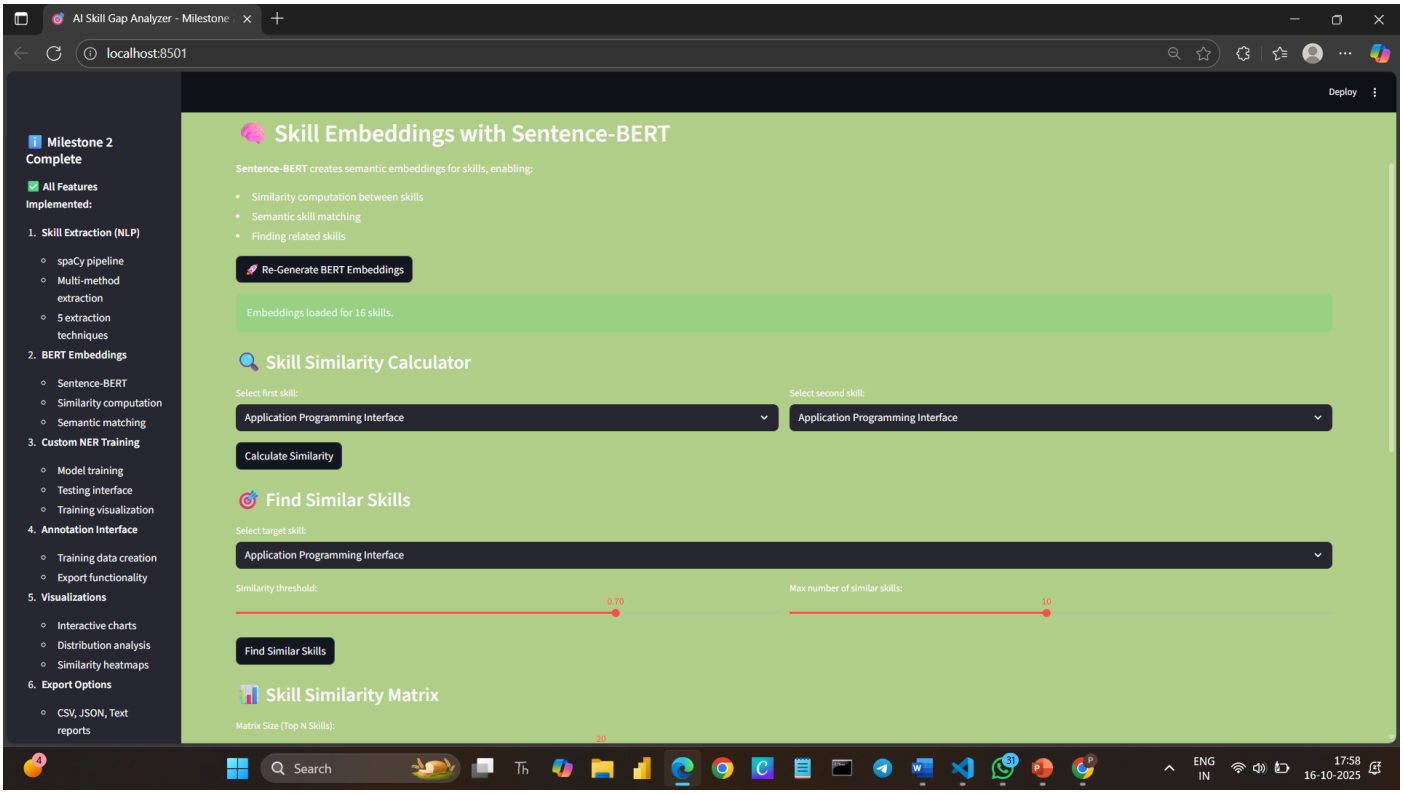
- **Total Skills Extracted**
- **Technical Skills Count**
- **Soft Skills Count**
- **Average Confidence Score**
- **Skill Distribution by Category**

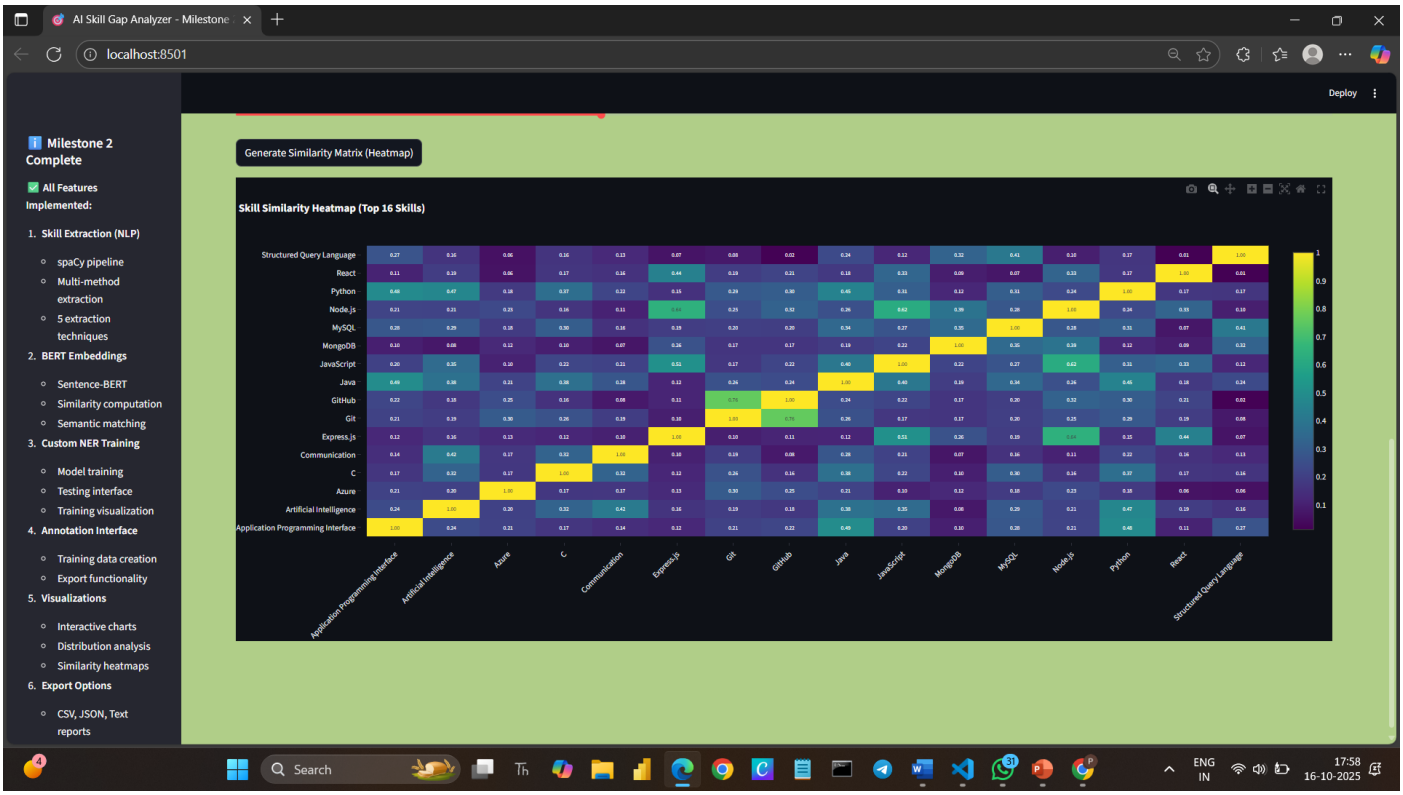
- **Extraction Method Statistics** (number of skills found by each method)
- **Similarity Scores** (between skills, using BERT)
- **Skill Gap Analysis** (planned for future milestones)

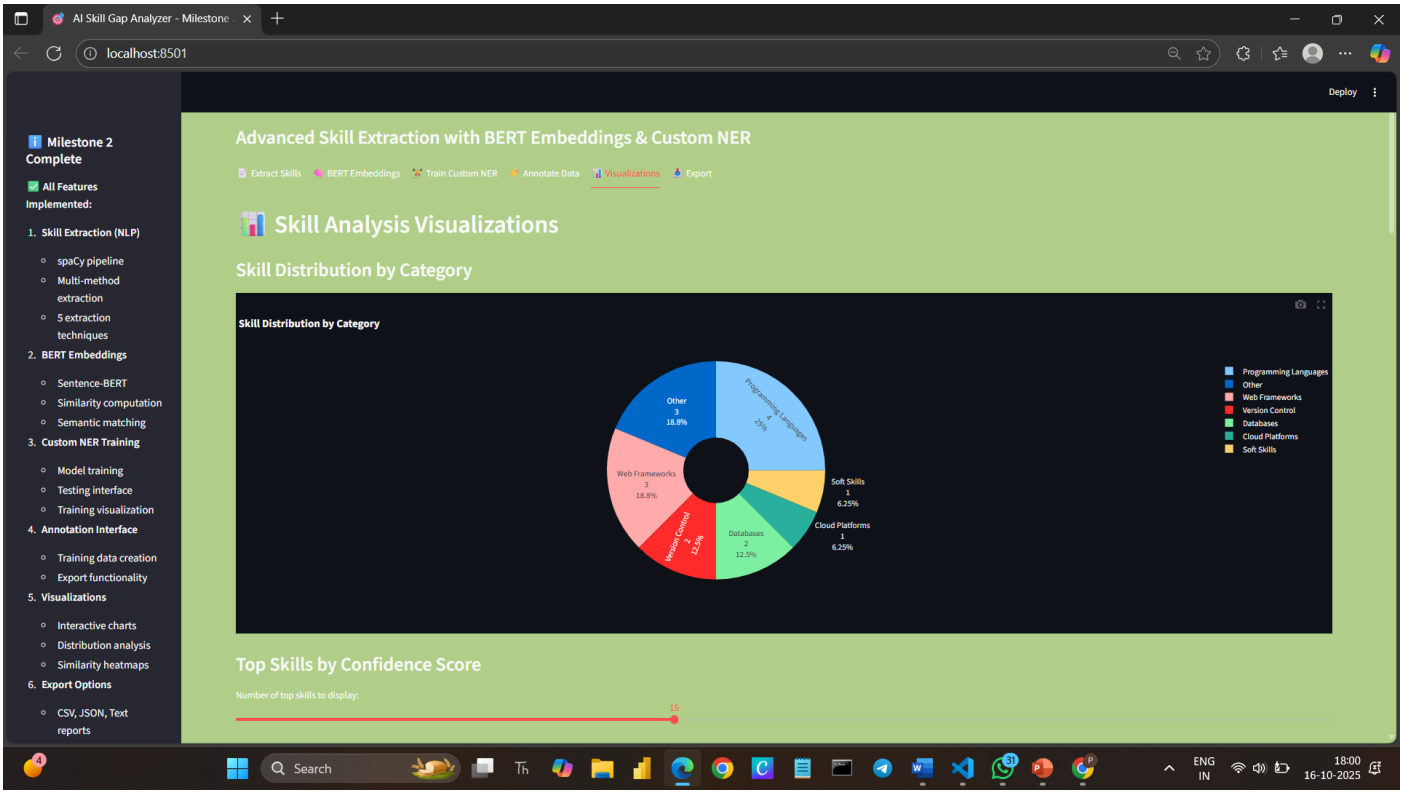
These statistics are presented as metrics, tables, and visualizations for easy interpretation.

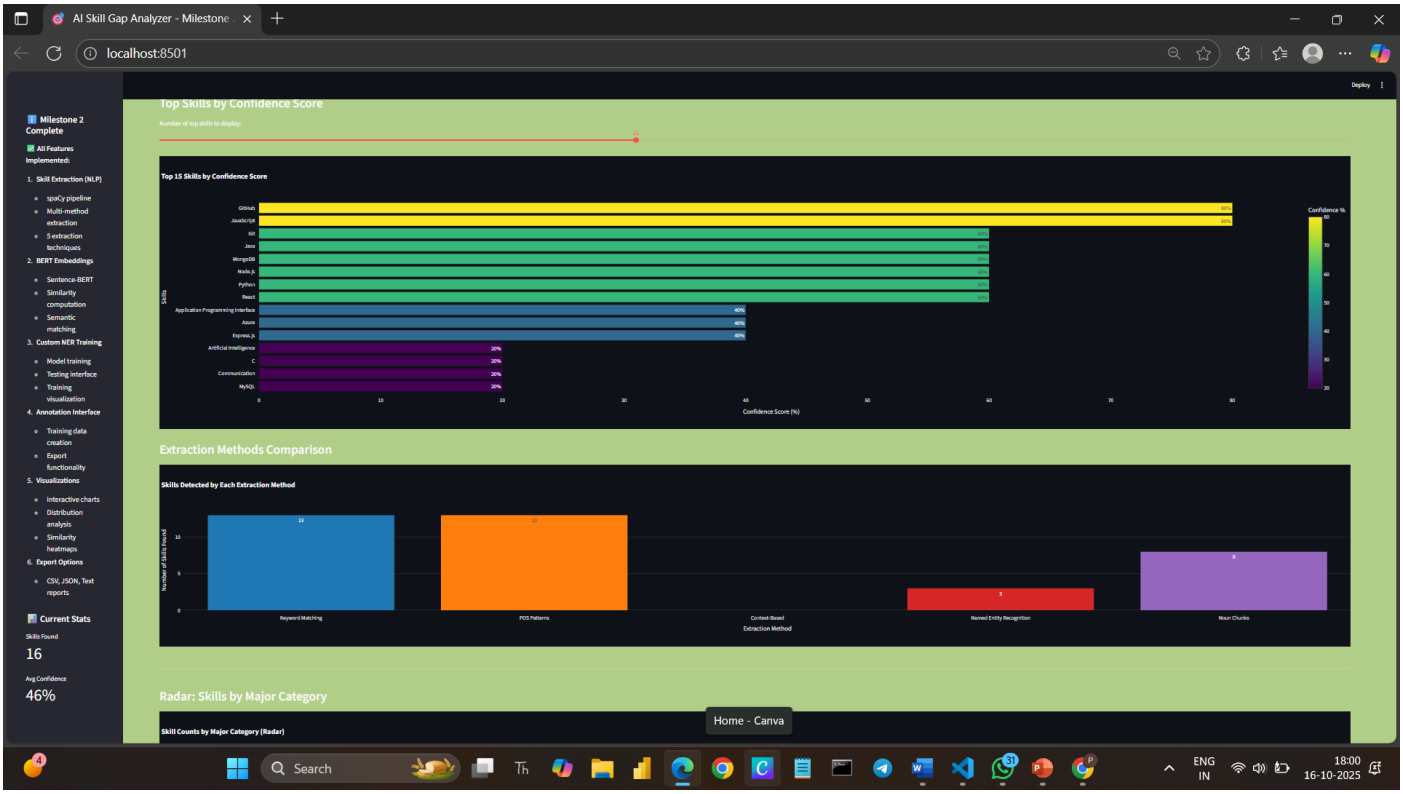
8.screenshots

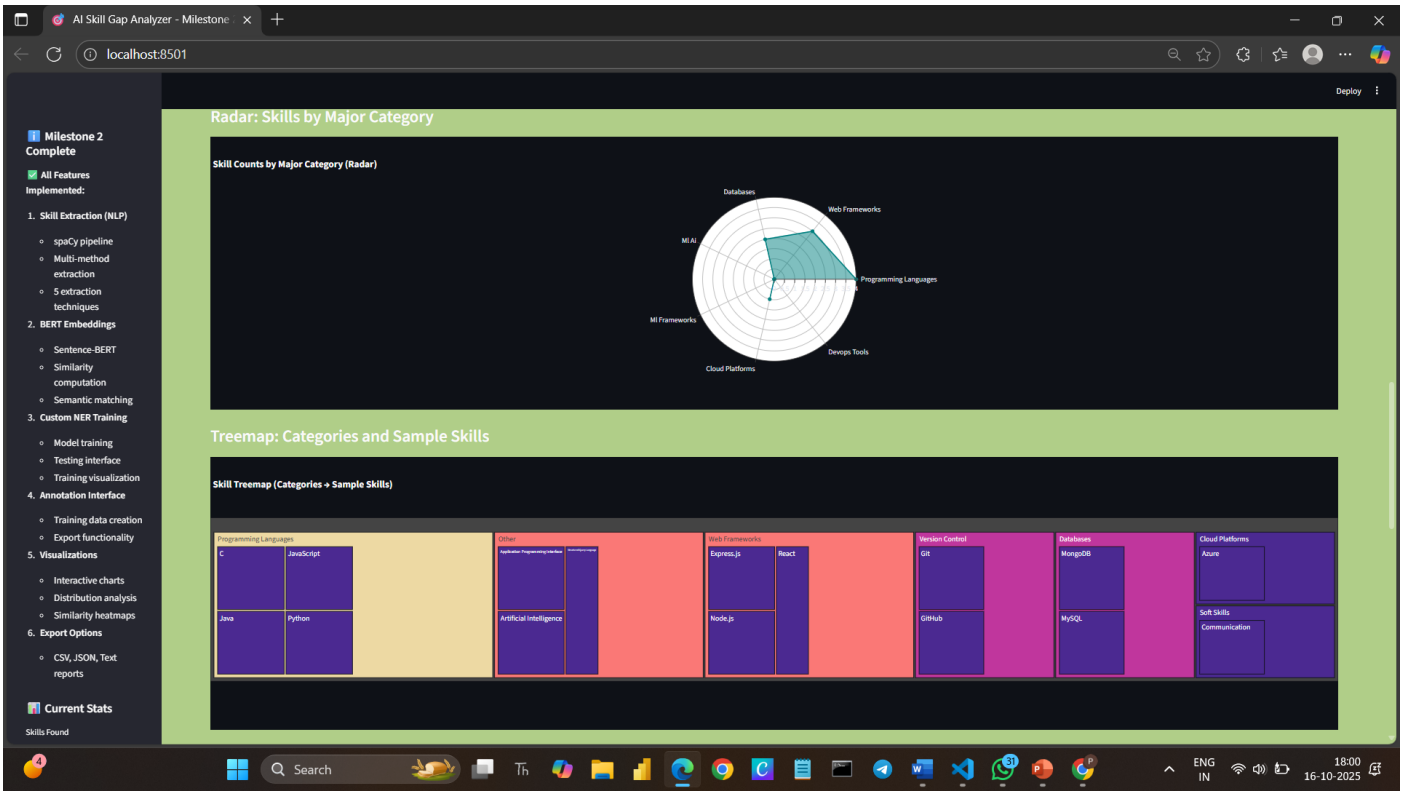


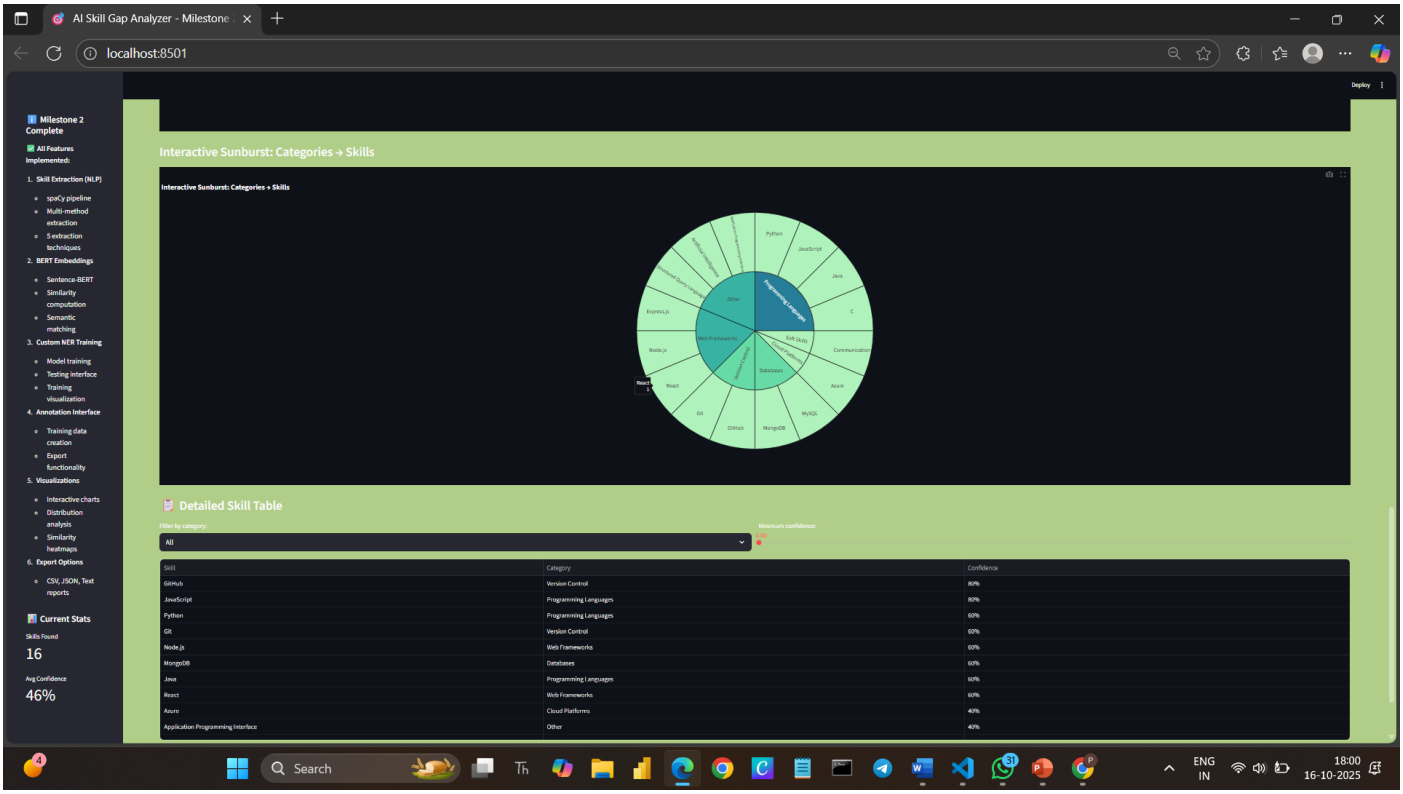

AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
1. Skill Extraction (NLP)
spaCy pipeline
Multi-method extraction
5 extraction techniques
2. BERT Embeddings
Sentence-BERT
Similarity computation
Semantic matching
3. Custom NER Training
Model training
Testing interface
Training visualization
4. Annotation Interface
Training data creation
Export functionality
5. Visualizations
Interactive charts
Distribution analysis
Similarity heatmaps
6. Export Options
CSV, JSON, Text reports
Skill Embeddings with Sentence-BERT
Sentence-BERT creates semantic embeddings for skills, enabling:
Similarity computation between skills
Semantic skill matching
Finding related skills
Re-Generate BERT Embeddings
Embeddings loaded for 16 skills.
Skill Similarity Calculator
Select first skill:
Application Programming Interface
Select second skill:
Application Programming Interface
Calculate Similarity
Find Similar Skills
Select target skill:
Application Programming Interface
Similarity threshold:
0.70
Max number of similar skills:
10
Find Similar Skills
Skill Similarity Matrix
Matrix Size (Top N Skills):
20
Search
ENG IN
17:58
16-10-2025


AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
1. Skill Extraction (NLP)
spaCy pipeline
Multi-method extraction
5 extraction techniques
2. BERT Embeddings
Sentence-BERT
Similarity computation
Semantic matching
3. Custom NER Training
Model training
Testing interface
Training visualization
4. Annotation Interface
Training data creation
Export functionality
5. Visualizations
Interactive charts
Distribution analysis
Similarity heatmaps
6. Export Options
CSV, JSON, Text reports
Generate Similarity Matrix (Heatmap)
Skill Similarity Heatmap (Top 16 Skills)
Structured Query Language
React
Python
Node.js
MySQL
MongoDB
JavaScript
Java
GitHub
Git
Express.js
Communication
C
Azure
Artificial Intelligence
Application Programming Interface
Search
ENG IN
17:58
16-10-2025


AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
1. Skill Extraction (NLP)
spaCy pipeline
Multi-method extraction
5 extraction techniques
2. BERT Embeddings
Sentence-BERT
Similarity computation
Semantic matching
3. Custom NER Training
Model training
Testing interface
Training visualization
4. Annotation Interface
Training data creation
Export functionality
5. Visualizations
Interactive charts
Distribution analysis
Similarity heatmaps
6. Export Options
CSV, JSON, Text reports
Advanced Skill Extraction with BERT Embeddings & Custom NER
Extract Skills
BERT Embeddings
Train Custom NER
Annotate Data
Visualizations
Export
Skill Analysis Visualizations
Skill Distribution by Category
Skill Distribution by Category
Other 3 18.8%
Programming Languages 4 25%
Soft Skills 1 6.25%
Cloud Platforms 1 6.25%
Databases 2 12.5%
Version Control 2 12.5%
Web Frameworks 3 18.8%
Programming Languages
Other
Web Frameworks
Version Control
Databases
Cloud Platforms
Soft Skills
Top Skills by Confidence Score
Number of top skills to display:
15
Search
ENG IN
18:00 16-10-2025


AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
1. Skill Extraction (NLP)
spaCy pipeline
Multi-method extraction
5 extraction techniques
2. BERT Embeddings
Sentence-BERT
Similarity computation
Semantic matching
3. Custom NER Training
Model training
Testing interface
Training visualization
4. Annotation Interface
Training data creation
Export functionality
5. Visualizations
Interactive charts
Distribution analysis
Similarity heatmaps
6. Export Options
CSV, JSON, Text reports
Current Stats
Skills Found
16
Avg Confidence
46%
Top Skills by Confidence Score
Number of top skills to display:
15
Top 15 Skills by Confidence Score
GitHub
JavaScript
Git
Java
MongoDB
Node.js
Python
React
Application Programming Interface
Azure
Express.js
Artificial Intelligence
C
Communication
MySQL
Skills
Confidence Score (%)
Confidence %
Extraction Methods Comparison
Skills Detected by Each Extraction Method
Keyword Matching
POS Patterns
Context-Based
Named Entity Recognition
Noun Chunks
Extraction Method
Radar: Skills by Major Category
Skill Counts by Major Category (Radar)
Home - Canva
Search
ENG IN
18:00 16-10-2025


AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
1. Skill Extraction (NLP)
spaCy pipeline
Multi-method extraction
5 extraction techniques
2. BERT Embeddings
Sentence-BERT
Similarity computation
Semantic matching
3. Custom NER Training
Model training
Testing interface
Training visualization
4. Annotation Interface
Training data creation
Export functionality
5. Visualizations
Interactive charts
Distribution analysis
Similarity heatmaps
6. Export Options
CSV, JSON, Text reports
Current Stats
Skills Found
Radar: Skills by Major Category
Skill Counts by Major Category (Radar)
Databases
Web Frameworks
ML/AI
Programming Languages
ML Frameworks
Devops Tools
Cloud Platforms
Treemap: Categories and Sample Skills
Skill Treemap (Categories → Sample Skills)
Programming Languages
C
JavaScript
Java
Python
Other
Artificial Intelligence
Web Frameworks
Express.js
React
Node.js
Version Control
Git
GitHub
Databases
MongoDB
MySQL
Cloud Platforms
Azure
Soft Skills
Communication
Search
ENG
IN
18:00
16-10-2025


AI Skill Gap Analyzer - Milestone
localhost:8501
Deploy
Milestone 2 Complete
All Features Implemented:
Current Stats
Skills Found
16
Avg Confidence
46%
Interactive Sunburst: Categories → Skills
Interactive Sunburst: Categories → Skills
Detailed Skill Table
Filter by category:
All
Minimum confidence:
Skill
Category
Confidence
GitHub
Version Control
80%
JavaScript
Programming Languages
80%
Python
Programming Languages
60%
Git
Version Control
60%
Node.js
Web Frameworks
60%
MongoDB
Databases
60%
Java
Programming Languages
60%
React
Web Frameworks
60%
Azure
Cloud Platforms
40%
Application Programming Interface
Other
40%
Search
ENG
IN
18:00
16-10-2025

9.Project Architecture

- **SkillDatabase:** Skill categories, abbreviations, and patterns.
- **TextPreprocessor:** spaCy-based text cleaning and NLP.
- **SkillExtractor:** Multi-method extraction and confidence scoring.
- **SentenceBERTEmbedder:** Semantic embeddings and similarity.
- **CustomNERTrainer:** Annotation, training, and prediction for NER.
- **AnnotationInterface:** UI for annotation management.
- **SkillVisualizer:** All interactive charts.
- **CompleteSkillExtractionApp:** Main Streamlit app, session state, and tab management.

10.Milestone Completed

Milestone 2 delivers a fully functional, extensible Streamlit application that supports advanced skill extraction, semantic analysis, custom NER training, and rich visualization. The system is modular, user-friendly, and ready for further enhancements such as synonym handling, OCR, multi-format file support, and personalized recommendations..