Semantic Food Search Engine

1. Project Overview

The Semantic Food Search Engine is designed to improve the search experience within a food delivery platform by moving beyond traditional keyword or regex-based search to a context-aware, semantic search system. It enables users to find dishes and menus based on the meaning and similarity of their queries rather than exact keyword matches.

The system leverages advanced Natural Language Processing (NLP) techniques, including custom text cleaning pipelines, Word2Vec and Sentence-BERT (SBERT) embeddings, and integrates with MongoDB Atlas Vector Search for scalable, real-time semantic retrieval.

2. Motivation and Objectives

- Problem: Legacy search systems relying on regex or keyword matching often fail to capture user intent, resulting in poor search relevance and user experience.
- Goal: Build a semantic search engine that understands the meaning behind user queries and retrieves relevant dishes even when exact keywords are missing or phrased differently.
- Scope: Handle a large dataset of 30,000+ dishes with multilingual descriptions, supporting efficient, scalable search with low latency.

3. Data Preparation and Cleaning Pipeline

3.1 Data Sources

- Datasets include dish names, descriptions, and categories, often containing noisy text such as nutrition terms, measurement units, and temporary placeholder words.

3.2 Cleaning Steps

- Non-ASCII removal: Strip out non-ASCII characters to standardize text.

- Number and unit removal: Remove numeric values and measurement units (e.g., "100g", "200ml") to reduce noise.
- Lowercasing and punctuation cleaning: Normalize text to lowercase and remove unwanted punctuation.
- Temporary word removal: Remove placeholder or irrelevant words that do not contribute to semantic meaning.
- Deduplication: Within each row, remove duplicate tokens while preserving the original order and underscore-separated phrase structure.
- Lemmatization: Use spaCy to lemmatize tokens, applying custom lemma exceptions and removing stopwords and nutrition-related words.
- Phrase joining: Join common multi-word phrases with underscores to preserve semantic units (e.g., "paneer_tikka").

This pipeline ensures high-quality, normalized input for embedding models.

4. Embedding Models

4.1 Word2Vec

- Trained a domain-specific Word2Vec model on the cleaned dish data to capture fine-grained word-level semantic relationships.
- Used parameters optimized for the dataset size and sentence structure (e.g., vector size 200, window size 4, min_count 2, skip-gram).
- Evaluated embeddings with custom similarity and analogy tests relevant to food items.

4.2 Sentence-BERT (SBERT)

- Used SBERT to generate dense, context-aware sentence embeddings for entire dish descriptions.
- SBERT embeddings capture broader semantic context beyond individual words, improving retrieval accuracy for complex queries.

5. Backend Integration with MongoDB Atlas Vector Search

- Stored SBERT embeddings in MongoDB Atlas using its native vector search capabilities.
- Configured vector indexes optimized for cosine similarity and Approximate Nearest Neighbor (ANN) search.
- Enabled efficient, scalable, and low-latency semantic search queries directly within the database.
- Built RESTful API endpoints to serve search requests, combining hybrid keyword and semantic search.

6. Evaluation and Testing

- Developed custom evaluation scripts including:
    - Food domain analogy tests (e.g., "idli" is to "sambar" as "poori" is to "bhaji").
    - Similarity scoring between related dishes (e.g., "paneer_tikka" and "paneer_butter_masala").
    - Nearest neighbor queries to inspect embedding quality.
- Monitored embedding coverage and vocabulary statistics to ensure robust model performance.

7. Results and Impact

- Semantic search significantly improved user query understanding and retrieval relevance compared to legacy regex search.
- Enabled discovery of related dishes even with varied user phrasing or synonyms.
- Backend vector search integration ensured scalable, production-ready deployment with low latency.

8. Future Work

- Explore hybrid embeddings combining Word2Vec and SBERT vectors for richer representations.
- Expand evaluation with user feedback and A/B testing.
- Incorporate multilingual embeddings to better handle regional languages.

9. Summary

This project demonstrates the end-to-end development of a semantic search engine tailored for the food delivery domain. It combines rigorous data cleaning, domain-specific embedding training, and modern vector search infrastructure to deliver a powerful, user-friendly search experience.