Jose forno

16 December 2024

Recipes System

Information Retrieval

Project Proposal

This project focuses on building an advanced information retrieval system for recipe datasets, utilizing a publicly available Kaggle dataset containing over 2 million recipes. The system is designed with two primary objectives: document classification and semantic search through natural language queries. The classification component groups recipes into meaningful categories like "vegan" or "seafood" using clustering techniques on semantic embeddings generated by pre-trained models. Meanwhile, the query system transforms user inputs into vector representations, enabling the retrieval of semantically relevant recipes ranked by similarity. Both functionalities leverage state-of-the-art NLP technologies, ensuring intuitive and efficient exploration of the dataset.

The development process involves several key technologies and methodologies. Python serves as the main programming language, incorporating libraries like Sentence-Transformers for embedding generation, Scikit-learn for clustering and dimensionality reduction, and Matplotlib for visualizations. PostgreSQL, with vector extensions, is used for storing embeddings and conducting similarity searches. The application will be fully containerized using Docker, ensuring ease of deployment. By integrating these tools, the project demonstrates the potential of modern information retrieval systems to enhance access to large, unstructured datasets while providing users with meaningful and accurate results based on natural language inputs.

Project Objective

Project Architecture

Text Embedding

Dataset

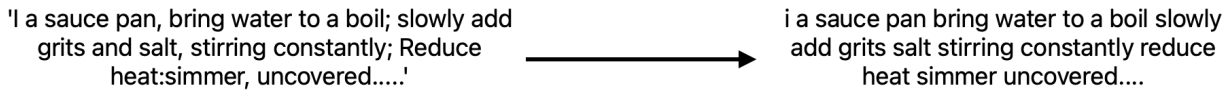
Explain the dataset, columns and what values are going to be use during the system

Pre Processing

The Idea of this step is to prepare the text to be embedded into a numerical format to then be processed by the IR system.

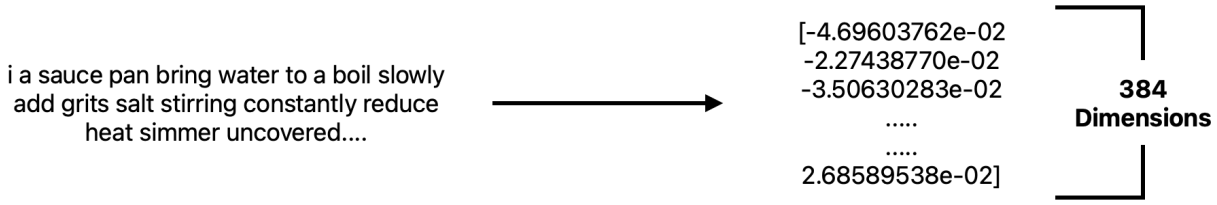
During this step:

- Sentences are separated into words
- All characters are converted into lower case characters
- Stop words are removed



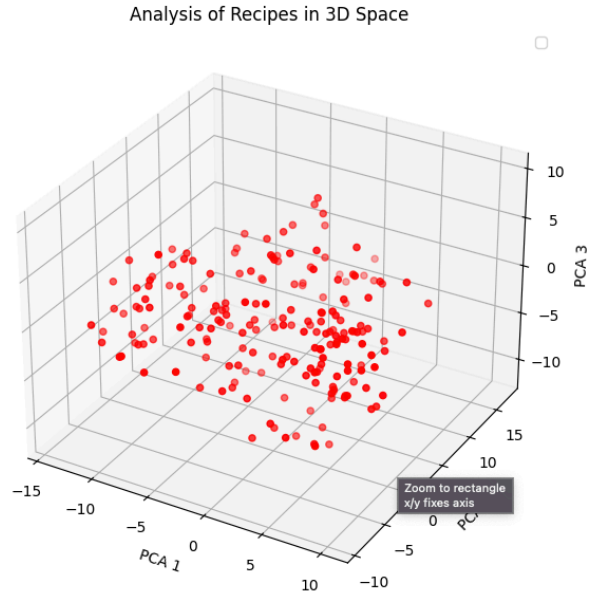
Embedding

For the embedding of the sentences the system uses a transformer it captures the relation between words and their context. It maps sentences and paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search. The following example illustrates the idea of the embedding of a sentence.



Sentence-Transformer Model
https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

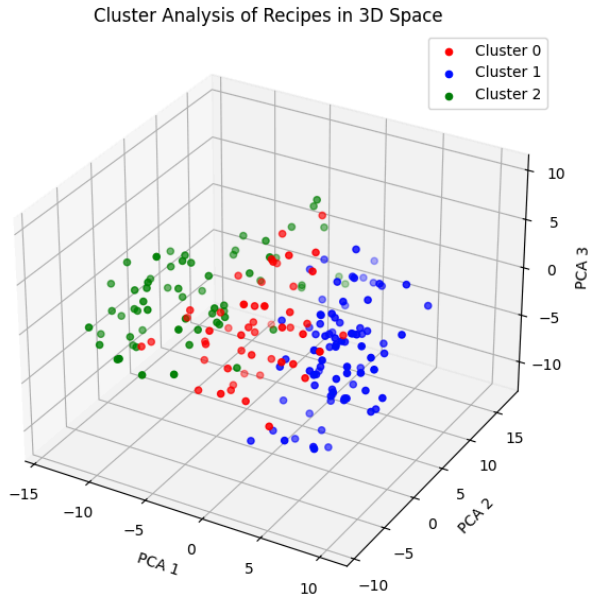
This vector can be represented as a point on that 384 embedded dimension. Expanding this procedure to all of the recipes on the dataset we can imagine a high dimensional space with multiple points represented with those embedded vectors. To have a better idea of this concept we can apply a dimensionality reduction technique (PCA) to map all vectors into a 3D space (Easier to understand for us humans). The next figure shows this concept on the first 200 recipes.



This 384 dimensions embedding allows for documents to be clustered, classify and even allows the system to use metrics of similarity between documents, enabling it to accomplishing the Information retrieval task.

Clustering

Using ML algorithms like **kmeans** the system can generate clusters of documents that represent a similar meaning, for this application, this means, similar recipes. Following the previews example, when applying a **Kmeans** with **K=3** we obtain the following visualization.



As can be seen, now the cloud of points is clustered into three groups, each one having recipes that contain similar content between them. This clusters then can be use to generate **classes** of recipes, for example like, Mexican food, Italian food, spicy food, etc. This process is usually done by human experts on the field, for the propose of this project some dummy labels will be generated.

Analyzing some recipes from broth cluster one and cluster two the similarities can be observed. Similar ingredients are being used from recipes on the same cluster, and also similar coking utensils and procedures. From that the expert can define a name for each cluster, like Salty food and sweets. One important remark es that the more clusters the algorithms generates (increasing K), more similar will be the recipes from each cluster and the name can be define more specifically. Having only three clusters the names are very general.

Cluster 1 - Salty Food

Recipe 1 - Grilled Garlic Cheese Grits

…, bring water to a boil…add grits and salt…Add cheese and garlic; stir until cheese is melted… non stick cooking spray… Brush both sides with **olive oil**… medium heat for 4 to 6 minutes on each side c until lightly browned.,

Recipe 2 - Asparagus Omelette Wraps

…Add the milk, sage, thyme, garlic, **pecorino** and season with cracked black pepper….enough salted boiling water to cover…. a large non stick flat pan with a little **olive** oil…. grated pecorino….

Cluster 2 - Sweets

Recipe 1- Hot Sweet Almond Brittle

Preheat oven to 375°F Place **almonds** in single layer in a 15 x 10 jelly roll pan… Remove from oven and place **nuts**.. Set aside. While the **almonds** cool, mix the salt, cumin coriander, cayenne and 1 T. **sugar** in a small bowl….

Recipe 2 - Chip Chocolate Chip Cookies

Position racks in the top and bottom thirds of the oven; preheat over to 350°. In a bowl, whisk the flour, baking soda, and **nutmeg** set aside...big bowl …Add **sugars…**