2023 CAPSTONE PROJECT: FINAL REPORT

DeepChef

A Modern Recipe Recommender

Amirhossein Kiani
Data Science Diploma Program @ BrainStation

Problem Statement

How can we use natural language processing (NLP) and unsupervised machine learning techniques to build a personalized recipe recommendation system that utilizes users' specific recipe preferences, such as ingredients, instructions and themes, and provides more curated recipe suggestions?

Background

In recent years, the significance of recommendation systems has grown substantially, offering an efficient means to filter and suggest personalized content to users according to their preferences. While content-based and collaborative filtering methods are prevalent in recommendation systems, they do have drawbacks, like over-specialization and cold start issues. To overcome these limitations, hybrid techniques that blend multiple approaches have displayed promising outcomes. Moreover, advancements in NLP and unsupervised machine learning present fresh prospects for personalized recommendation systems.

Value Added

DeepChef aims to provide a unique, user-centric, and engaging recipe recommendation experience by incorporating state-of-the-art natural language processing and unsupervised machine learning techniques. The recommendation system utilizes an OpenAI's advanced embedding model, which allows for comprehensive analysis of recipe plot summaries and other textual data. DeepChef has the potential to offer personalized and meaningful recipe recommendations that cater to specific cuisine preferences and tastes, which may enhance user engagement and satisfaction, proposing an alternative to the typical "phrase-based" search of recipes that different websites offer.

Potential Use Cases

- **Personalized recipe recommendations**: Users could input their desired recipe ingredients, instructions or themes to receive recipe recommendations tailored to their tastes. This could help users discover new recipes that they might not have otherwise found and could enhance their overall culinary experience.
- **Cooking with Leftovers:** Users can input leftover ingredients, and the recommender can suggest creative ways to use them in new recipes.
- **Cuisine Exploration:** Users can input ingredients and flavors from different cuisines they want to explore, and the recommender can suggest authentic recipes from those cuisines.
- **Content recommendation for cooking websites**: cooking websites could integrate and fully develop DeepChef to complement their recommendation systems, offering users recipe suggestions based on specific recipe preferences or cuisines. This could enhance user satisfaction and engagement, and ultimately drive more views and revenue for these websites.

About the Data

For this project, I used and enhanced two publicly available datasets (on kaggle.com): a recipes dataset, containing over 520,000 recipe information scraped from food.com, and a dataset containing over 1,400,000 user reviews on a large portion of the recipes. Two of the crucial columns missed lots of important contextual information, so I re-scraped the whole dataset to address th issue.

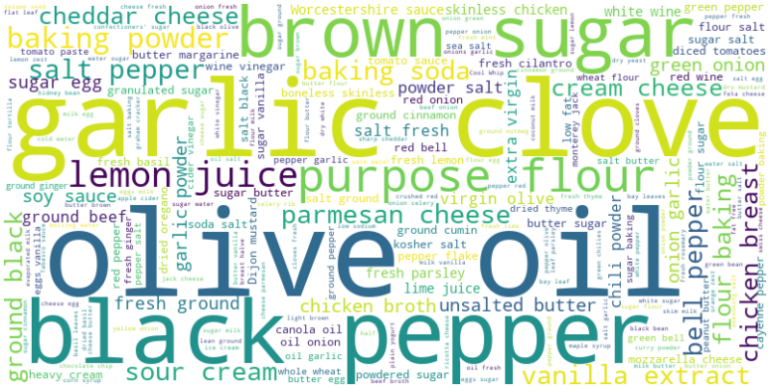


Figure 1] The 'word-cloud' of the entire set of ingredients used in the recipes dataset. The larger a phrase is, to more prevalent it is. In this case, some of the most common ingredients are olive oil, garlic glove, black pepper and brown sugar.

DeepChef can be extended in various ways. Currently, I plan to incorporate other features into the recommender system. For instance, all the recipes have associated data on their nutritional factors and it would be ideal to give the users the option to choose recipes within ranges of nutritional values, and find recipes that are similar in those regards.

So far, I have provided the codes for such extensions (available on the project's GitHub repository) but have't implemented this feature into the Streamlit app (see the last section).

Data Cleaning, EDA, & Text Preprocessing

During the process of data cleaning, I successfully created the links to the recipes on food.com's domain and scraped almost all of them to retrieve a large number of missing crucial ingredient informations associated with recipes. These were later on used in preprocessing the recipes text data in performing topic modeling, extracting semantic embeddings, and powering the Streamlit app.

Aside from removing and feature-engineering several columns from the original recipes dataset, I added some new features (such as recipe 'topics' and semantic embeddings) and fixed a few (such as aggregated ratings associated with over 80,000 recipes).

As for text preprocessing, I curated a function that performs several NLP tasks, such as lemmatizing, stemming and tokenizing. This function was then applied to most of the information-rich text data of recipes and used for topic modeling and semantic embedding.

Feature Extraction, Feature Engineering, and Analysis

In my project, I utilized various feature extraction and engineering techniques to transform raw data into features suitable for machine learning algorithms. To convert ingredients' textual data into numerical vectors, I used open-source sentence transformers through the Bertopic library.

Bertopic leverages BERT's contextual word embeddings to represent the documents and clusters them based on their semantic similarity. It aims to improve upon traditional topic modeling algorithms by utilizing the contextual information captured by BERT, which can lead to more accurate and meaningful topic representations.

I incorporated the extracted topics into the original recipes dataset to bring insights on the distribution of the recipes with respect to them, a task that is done in the Jupyter notebook on EDA, which is available on the project's GitHub repository.

In my analysis, I also used Bertopic to identify underlying themes among the recipes and their ingredients in the dataset. To visualize the topics extracted using of BERT, I used Bertopic's built-in visualization tools create interactive visualizations of the identified topics.

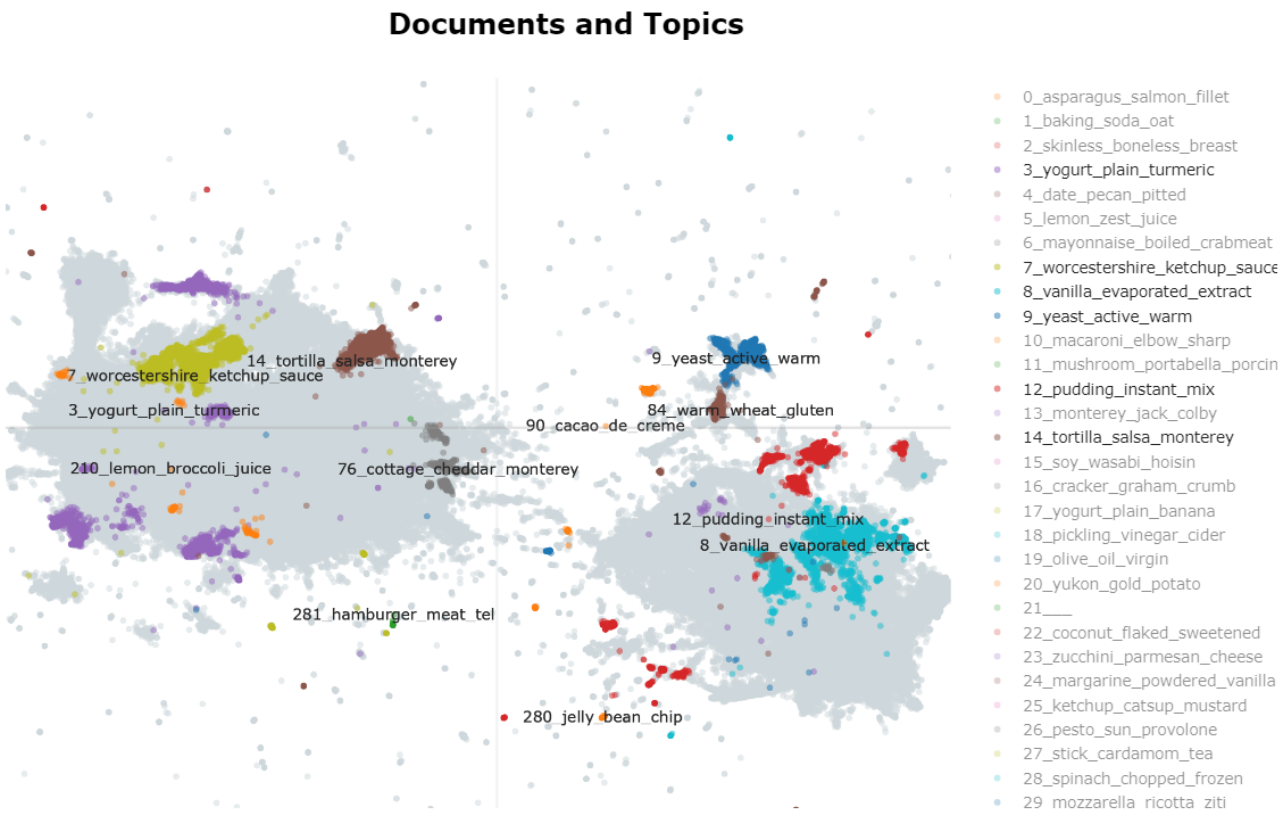


Figure 2] A screenshot of an interactive Bertopic visualization with the number of topics set as 300. (The majority of topics have been turned 'off' to bring more visibility for this image.) For instance, as we can see, a topic (topic 9) has the words 'yeast', 'active' and 'warm' in it as the prevalent terms. This topic presumably comprises of many, if not all, recipes for baking bread.

Modeling & Evaluation

To extract the semantic embeddings associated with recipes' text data, which were later used to power our main recommendation system, I used OpenAI's state of the art embedding model, ada-002. Ada-002 is a contextual embedding model, meaning it generates embeddings based on the context of the input text. It can understand the meaning of words based on the surrounding words and sentences, allowing it to capture complex linguistic relationships.

To assess the quality of the embeddings that were going to be used for our recommender system, I decided to first experiment with it at a smaller scale. I first fed to the model only recipe ingredients, but the results weren't very intriguing: the recommender system couldn't pick up well on ingredient measures and other contextual information associated with recipes. In the next iteration, I preprocessed and incorporated much more contextual information to the embedding model, and the results became significantly more satisfactory.

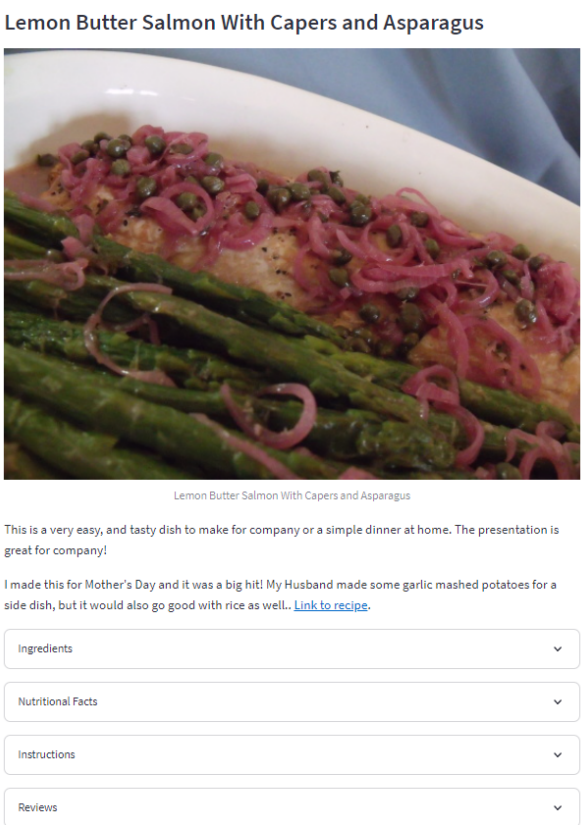
Streamlit Application

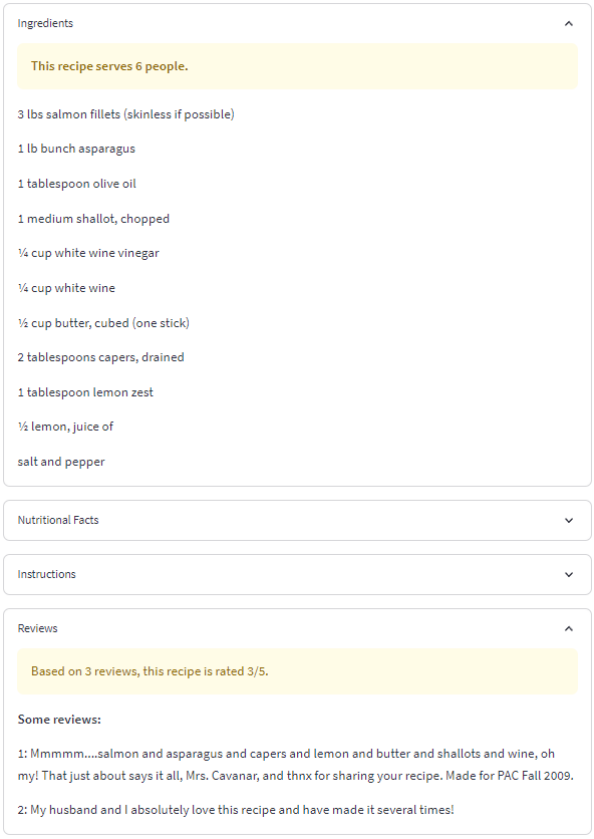
To demonstrate my recipe recommendation system, I created a simple yet functional application using Streamlit. This app is currently available at this address: https://deepchef.streamlit.app

Streamlit is an open-source Python library that enables data scientists and developers to create interactive web applications for data exploration, visualization, and machine learning models with ease. It is designed to bridge the gap between data scripting and web development, allowing users to turn their data analysis code into shareable and interactive web apps with minimal effort.

Based on my experience with the app, I recommend users include recipe-specific keywords in their input, particularly ingredients. The model accepts a user prompt of any length, turns the input into vectors (using OpenAI's ada-002, for which I have created a subscription to make sure the app is up and running), and runs the vector against the embeddings that were formerly derived. The model uses cosine similarity in its comparisons and returns the top-five most similar recipes to the user query.

The output contains an image of the recipe (crated by the author or any other user), a description (written by the recipe author) and an expandable list of ingredients and their quantities. It also retrieves top-two reviews (in the sense of publishing date), the average user rating and a link to the recipe on the website itself.





DeepChef

A Modern Recipe Recommender

DeepChef is a recommender system created by Amir Kiani that uses user prompts consisting of ingredients, recipe instructions and even themes, and retrieves recipes from food.com that are most similar to the prompt. This app uses OpenAI's semantics embdedding models on the text data of over 520,000 recipes scraped from food.com. Learn more about DeepChef, its motivations, limitations and codes at its GitHub repository.

Sample Prompts

What are you craving?

A delicious and healthy dish featuring grilled salmon fillets served with roasted asparagus.

Please click only ONCE and wait until the results show up before making a new query

Submit

Lemon Butter Salmon With Capers and Asparagus

Figures 3, 4, 5] My Streamlit app recommends **Lemon Butter Salmon With Capers and Asparagus** as the top choice to the user input of "A delicious and healthy dish featuring grilled salmon fillets served with roasted asparagus." The retrieved image, as well as other recipe info are showcased above.