Clustering Recipes: Methodologies and Insights

Abstract

Clustering recipes into meaningful categories is a challenging task due to the complex interplay of ingredients, preparation methods, and cuisines. This research aims to explore and refine clustering techniques to group recipes effectively. Starting with a dataset of 6,000 recipes, preprocessing and purification steps reduced the dataset to 300, and a series of experiments using advanced vectorization, dimensionality reduction, and clustering algorithms were conducted. The initial ground truth of 34 clusters yielded low Adjusted Rand Index (ARI) scores, prompting ground truth revision to 15 clusters. The study highlights the role of iterative refinement in achieving improved ARI scores, with the highest score of 1.0 achieved using Doc2Vec with PCA on the revised ground truth.

1 Introduction

Categorizing recipes into well-defined groups can enhance applications like personalized recipe recommendations and culinary analysis. This study investigates the clustering of recipes using machine learning techniques and evaluates clustering accuracy against manually created ground truths.

2 Dataset Overview and Preprocessing

2.1 Initial Dataset

- **Source:** The dataset consisted of 6,000 recipes collected from diverse sources, including online recipe platforms and culinary resources.
- **Attributes:**
  - Recipe titles and descriptions
  - List of ingredients
  - Cooking methods

2.2 Data Purification

To ensure the quality of clustering results, extensive preprocessing steps were applied:

1. **Duplicate Removal:** Recipes with identical attributes were removed.
2. **Null Values:** Entries with missing or incomplete data were discarded.

3. **Text Cleaning:** Ingredients and instructions were cleaned by removing:
    - Stopwords
    - Special characters
    - Irrelevant text patterns (e.g., ads or promotional content).

2.3 Dataset Reduction

- From **6,000 to 3,000 Recipes:** Recipes were filtered based on completeness and unique attributes, focusing on diversity in ingredients and preparation methods.
- From **3,000 to 300 Recipes:** The dataset was further reduced for computational efficiency and manual validation by selecting representative recipes from each preliminary category.

3 Knowledge-Based Categorization

Before applying clustering algorithms, recipes were preliminarily grouped using **Wikipedia** and other culinary resources for reference. This step provided insights into the commonalities among recipes, which informed the creation of the initial ground truth.

4 Initial Ground Truth Creation

4.1 Old Ground Truth

Recipes were manually categorized into **34 clusters** based on features such as primary ingredients, cuisine type, and preparation methods. Examples included Curry, Bread, Desserts, Beverages, and Snacks.

5 Methodology

5.1 Vectorization Methods

To convert textual recipe data into numerical vectors suitable for machine learning, three methods were applied:

1. **TF-IDF (Term Frequency-Inverse Document Frequency):** Captures the importance of terms in a recipe relative to the dataset.
2. **SBERT (Sentence-BERT):** Encodes recipes into contextual embeddings using transformer models.
3. **Doc2Vec:** Learns dense vector representations for entire documents, enabling semantic analysis of recipes.

5.2 Dimensionality Reduction

To reduce the dimensionality of vectorized data, the following techniques were tested:

1. **PCA (Principal Component Analysis):** Linear dimensionality reduction focusing on variance preservation.
2. **UMAP (Uniform Manifold Approximation and Projection):** Non-linear dimensionality reduction that preserves local and global data structure.

5.3 Clustering Algorithms

- **K-Means Clustering:** Partitioned recipes into 34 clusters, aligning with the old ground truth.
- **DBSCAN and HDBSCAN:** Density-based methods for identifying clusters of arbitrary shapes.

6 Initial Results and Observations

Using the old ground truth with 34 clusters, clustering performance was evaluated using the **Adjusted Rand Index (ARI):**

Vectorization	Dimensionality Reduction	ARI Score
TF-IDF	PCA	0.144
TF-IDF	UMAP	0.292
SBERT	PCA	0.258
SBERT	UMAP	0.270
Doc2Vec	PCA	0.612
Doc2Vec	UMAP	0.264

Table 1: Initial ARI Scores

Despite extensive parameter tuning, the ARI scores remained unsatisfactory, with marginal improvements through modifications in:

- Vectorization parameters (e.g., max features, n-grams).
- Dimensionality reduction parameters (e.g., UMAP's `n_neighbors`, PCA's `n_components`).

7 Ground Truth Revision

To address clustering inconsistencies, the ground truth was revised:

7.1 New Ground Truth

Recipes were reorganized into **15 broader clusters** based on key characteristics (e.g., primary ingredients, preparation styles). Examples of new clusters include Curry, Desserts, Beverages, and Rice.

Cluster	Category	Description
1	Curry	Spiced gravy dishes common in South Asian cuisines.
2	Beverages	Hot and cold drinkable recipes.
3	Dessert	Sweet recipes typically served as treats or at the end of meals.
⋮	⋮	⋮

Table 2: Revised Ground Truth Categories

8 Revised Experiments

With the new ground truth, experiments were repeated. Results showed significant improvement in clustering performance:

Vectorization	Dimensionality Reduction	ARI Score (Old Ground Truth)	ARI Score
TF-IDF	PCA	0.144	0.072
TF-IDF	UMAP	0.292	0.146
SBERT	PCA	0.258	0.162
SBERT	UMAP	0.270	0.277
Doc2Vec	PCA	**0.612**	**1.000**

Table 3: Revised ARI Scores

9 Analysis and Insights

1. **Impact of Revised Ground Truth:** The new ground truth simplified cluster definitions, improving semantic alignment.
2. **Best Performing Combination: Doc2Vec with PCA** achieved an ARI of **1.0**, validating the importance of tailored vectorization and dimensionality reduction.
3. **Dimensionality Reduction Performance:** UMAP captured non-linear relationships better than PCA but struggled with small clusters.

10 Conclusion

This study demonstrated the importance of iterative refinement in recipe clustering:

1. **Revised Ground Truth:** Simplified cluster definitions provided better alignment with algorithmic outputs.
2. **Optimal Methodology:** The combination of **Doc2Vec with PCA** and the new ground truth yielded the best results, achieving an ARI of **1.0**.
3. **Future Directions:**
   - Incorporating additional metadata like cooking time and calorie count.
   - Exploring ensemble clustering techniques for further improvements.

This framework provides a robust approach for clustering recipes, enabling practical applications in culinary recommendation systems and food research.