COOKR HACKATHON

SUBMITTED BY

1.MENAKA A(CSE) -717721173O9
2.HASHINI B A (IT) -717721181I5

PROBLEM STATEMENT

Statement - Kitchen Review Summary

Description:

Build algorithms to automate the generation of summaries for customer reviews. We receive numerous reviews from different customers for various orders, and we need to generate summaries from the hundreds of reviews collected.

For instance:

Review 1: Praises about the quality

Review 2: Praises about the Timeliness

Review 3: Complaints about Quantity So, the summary review can be - People usually praise this kitchen's quality, but they have an issue with the quantity of the food provided. They also appreciate the proper timely delivery from this kitchen.

OUR IDEA

Each Kitchen will get overall review summary , considering many aspects from the customer.

The reviews from the customer not only focus on some constraint, but also some additional points that customer shares. Basic Characteristics like **Quantity , Quality , Timeliness** are given points using which each Kitchen will get into ranking system. We also take note of additional data from the customer , that would be

1. Innovation in Recipe
2. Presentation of Dishes.
3. Ease of Ordering
4. Communication and Customer Service
5. Flexibility in Customization
6. Transparency in Ingredients

These additional features were also given points, and that will enhance Kitchen's Overall Rating.

Suppose if a kitchen does not make into these additional features , it will not affect the Kitchen's overall impression. The idea behind it is each kitchen will get rating and also short review for each basic characteristics mentioned above. As mentioned , each will receive a over all summary based on their reviews. Even if the kitchen's basic characteristic rating is nil , the overall summary will be provided. Ranking system helps in identifying the kitchen that well versed in all the aspects.

FUTURE PROGRESS

From the idea , lets dig into the future progress and how are idea can be used in a large scale

- **Deep Learning Models**:

  1. Recurrent Neural Networks (RNNs): Capturing temporal dynamics, handling variable-length sequences

  2. Convolutional Neural Networks (CNNs): CNNs excels at capturing spatial information and local patterns in data.

  3.Generative Pre-trained Transformer(GPT) comprehensive analysis and understanding of large-scale text data.

- **Batch Processing :**Batch processing for processing large volumes of historical text data in batches. This allows you to analyse and summarize data in chunks.

  Benefits : Saves time instead of looking all the reviews

- **Data Pre-processing:** Remove noise, and normalize text for analysis. Extract relevant features from text data, such as word embeddings or TF-IDF vectors Benefits: Removes noise , errors and inconsistencies

CHALLENGES FACED

- Training models on large datasets requires substantial computational resources, including CPU, memory, and GPU resources.
- Training large datasets can be time-consuming, sometimes taking days, weeks, or even months to converge
- Large datasets increase the risk of overfitting, where the model learns to memorize the training data rather than generalize to unseen data.
- Labelling large datasets for supervised learning tasks can be time-consuming and expensive
- Ensuring data quality becomes more critical as the size of the dataset increases.

OUR IMPLEMENTATION

- A Simple model for the problem statement considering the basic characteristics like Quantity, Quality , Timeliness is made into account.
- Each reviews of the Kitchen is analysed , based on which short review for each basic characteristics is provided.

Quantity Rating	Quantity Review
Quality Rating	Quality Review
Timeliness Rating	Timeliness Review

- The algorithm we implemented here is,

Sentiment Analysis: Used the VADER(Valence Aware Dictionary and Sentimental Reasoner

Text Summarization: Used the TextRank algorithm, which is an unsupervised graph-based algorithm for extractive text summarization.

DEMO

PROBLEM STATEMENT -2

Statement - Item Categorization.

Description:

Create a model or research the necessary steps to create a model for categorizing items. When the cook adds an item to their kitchen, it should be automatically categorized into multiple categories. We can provide the sample data for this to train the model.

For instance:

- Idly - South Indian, Protein Rich, Breakfast, Baked Items etc.
- Chicken Vindaloo - North India, Punjabi, Non-Veg, Chicken, Protein Rich etc.
- Ragi Dosa - South Indian, Diabetic Friendly, Millet Based, Pregnancy friendly etc.

OUR IDEA

- Given a dish name as input , the software will give out the further details regarding to it
- When a user gives a dish name as a input , it will fetch out the dataset from where it directs the preparation method and key ingredients used up.
- Once checking at the ingredients , we can categorize it's cuisine.
- Looking at the preparation style we can find out it's style of preparation i.e: baked, fried
- Now three key ingredients that is used in large quantity is taken out .
- According to the key ingredients , the nutrient value of the dish is calculated i.e: Carbohydrate, protein, fiber etc.
- Each age group of people follow different diet.
- Taking note of the nutrient value , specific dish can be suggested for different users

  i.e :

  Steamed Broccoli , boiled eggs – Pregnency-Friendly

PREPROCESSING

Extract relevant information from the dataset (XLS file) such as ingredients, preparation method, key ingredients, nutrient values, etc.

- Text processing for dish names (removing stop words, stemming, etc.).

Algorithms:

- For text processing: Natural Language Processing (NLP) techniques, such as tokenization, can be used. Libraries like NLTK or SpaCy can help.

PRE PROCESSING

1. Tokenization:

Definition: Tokenization is the process of breaking down text into smaller units, usually words or phrases (tokens)

2. Text Similarity:

Definition: Measures the similarity between two texts.

3. Text Classification:

Definition: Assigns predefined categories to text based on its content.

4. Named Entity Recognition (NER):

Definition: NER identifies and classifies entities (e.g., names of people, locations, organizations) in text.

CUISINE CATEGORIZATION

- Use the extracted ingredients to categorize the cuisine of the dish.
- Algorithms that work well for text classification tasks, such as Naive Bayes or Support Vector Machines, can be effective.

Algorithms:

Naive Bayes Classifier or Support Vector Machines for text classification.

NAIVE BAYES

TRAINING: The Naive Bayes Classifier works well for text classification. For training, you provide a dataset with food names and their corresponding categories.

Example:

Training Data

Food Name	Category
Idly	South Indian
Chicken Vindaloo	North Indian
Ragi Dosa	South Indian

Training the Model: The algorithm learns the probability of each category given the food name.

Prediction When a new food name is given, the algorithm calculates the probability of each category and assigns the category with the highest probability.

Example

Test Data:

Food Name	Predicted Category
Masala Dosa	South Indian

STYLE OF PREPARATION

- Analyze the preparation method to determine the style (baked, fried, etc.).
- Decision Trees or Random Forests can be suitable for this, as they handle both categorical and numerical features well.

Algorithms:

- Decision Trees or Random Forests for classification based on mixed features.

DECISION TREE

Decision Trees learn by splitting the data based on features. For training, you provide a dataset with food features (e.g., spice level, cuisine) and their corresponding categories.

Example:

Training Data

Food Name	Spice Level	Cuisine	Category
Idly	Low	South Indian	Breakfast
Chicken Vindaloo.	High	North Indian	Non-Veg
Ragi Dosa	Medium	South Indian	Diabetic Friendly

Training the Model:* The tree splits based on features to classify each food item.

Prediction: When a new food item is given, it traverses the tree to predict the category based on features.

Example

Food Name	Spice Level.	Cuisine	Predicted Category
Masala Dosa	Medium	South Indian.	Breakfast

RANDOM FOREST

Training: Random Forest is an ensemble of decision trees. You provide a dataset similar to Decision Trees.

Used Prediction:

It predicts based on the majority vote of multiple decision trees.

Example:

Food Name	Spice Level	Cuisine	Predicted Category
Masala Dosa	Medium	South Indian.	Breakfast

NUTRIENT VALUE CLASSIFICATION

- Use the key ingredients to calculate the nutrient values.
- Regression algorithms, such as Linear Regression, can be applied for this numerical prediction task.

Algorithms:

- Linear Regression for predicting nutrient values based on key ingredients.

AGE GROUP DIET SUGESTION

- Based on the calculated nutrient values, categorize the dish for different age groups.
- Clustering algorithms, such as K-Means, can help group dishes based on nutrient profiles.

Algorithms:

- K-Means Clustering for grouping dishes based on nutrient values.

K CLUSTERING

Training: KNN doesn't explicitly train. It memorizes the training data.

Example:

Training Data:

Food Name	Feature 1	Feature 2	...	Category
Idly	Value 1	Value 2	...	South Indian
Chicken Vindaloo	Value 1	Value 2	...	North Indian
Ragi Dosa	Value 1	Value 2	...	South Indian

Prediction: It finds the k-nearest items in the training set and predicts based on majority voting.

Example:

Test Data:

Food Name	Feature 1.	Feature 2	...	Predicted Category
Masala Dosa	Value 1	Value 2	...	South Indian

RECOMMENDATION SYSTEM

- Suggest specific dishes for users based on their age group and dietary preferences.
- Collaborative Filtering or Content-Based Recommendation systems can be used.

Algorithms:

- Collaborative Filtering or Content-Based Recommendation systems for suggesting dishes.

CHALLENGES FACED

- Both Naive Bayes and Decision Trees can be sensitive to irrelevant features in the dataset. Including irrelevant features can lead to overfitting in Decision Trees
- When dealing with continuous features, discretization may be required, which can lead to information loss and reduced model performance
- Categorizing the based on the cuisine would be tricky
- Handling huge volume of data and training it is a big task
- Training machine learning algorithm is time consuming
- Training algorithm is expensive

THANK YOU