Project 19: Food Review Analysis **[Mehrdad]**

The main objective of this project is to build a personalized food recommendation system that incorporates user reviews, ratings, and food content information, using advanced NLP techniques to analyze reviews and improve recommendation quality. Consider the Dataset of Food Recommendation Systems which is recommender system dataset collected from Food.com website. In this project you will use these data from this Kaggle repository:

- **RAW_recipes.csv**: Contains the raw dataset with detailed recipe information including name, id, minutes, contributor_id, submitted, tags, nutrition, n_steps, steps, description, ingredients, and n_ingredients.
- **RAW_interactions.csv**: Contains the raw dataset with detailed recipe information including user_id, recipe_id, date, rating, and review.

1. Initial Data Exploration: Explore the dataset by analyzing key features such as the number of ingredients, number of steps, and tags associated with each recipe. Visualize the distributions of these features using histograms, bar charts, or word clouds to gain insights into the overall structure of the dataset and identify patterns in the recipe metadata and reviews.

2. Data Preprocessing: Clean and preprocess the text data by removing special characters, punctuation, and stopwords. Convert all text to lowercase for uniformity and apply tokenization to split the reviews into individual words or tokens.

3. Review length Analysis: Compute basic statistics for the length of the reviews, including mean, median, minimum, and maximum values, measured by both the number of words and characters. Visualize the distribution of review lengths using histograms or box plots to understand the variation across the dataset. Next, compare the review lengths across different rating groups (0-5) to identify any potential relationships between review length and rating. Calculate the correlation between rating groups (0-5) and the length of reviews to determine if longer or shorter reviews are associated with higher or lower ratings. Finally, plot the average review length for different rating groups to explore and visualize any interesting patterns.

4. Sentiment Analysis on User Reviews: Write a script using pre-trained sentiment analysis models (e.g., VADER, TextBlob, or BERT) to transform each user review into a 0-5 rating scale, where 0 indicates completely negative and 5 indicates completely positive sentiment. Compare the sentiment scores with the corresponding user ratings and analyze the alignment between them. Perform an analysis for each rating group (0-5) and calculate the correlation between sentiment scores and ratings for each group.

5. Topic Modeling on Recipe Descriptions and Reviews: Use Latent Dirichlet Allocation (LDA) or Latent Semantic Analysis (LSA) to perform topic modeling on recipe descriptions and user reviews. Identify the main topics discussed in the dataset, both in the recipe descriptions and user review. Explore how these topics correlate with user sentiments and ratings, and identify which topics generate more positive or negative feedback.

6. Named Entity Recognition (NER) for Ingredient Extraction: Apply Named Entity Recognition (NER) techniques to extract key ingredients and food items from the recipe descriptions. This step aims to identify important food components that might influence user preferences and recommendation decisions. Evaluate how different food content mentions affect user ratings and sentiments in the reviews.

7. Sentiment Trends by Cuisine and Ingredient Groups: Analyze the sentiment polarity of user reviews for different cuisine types or ingredient groups. To this aim suggest an approach to group the recipes by cuisines (e.g., Italian, Mexican) or main ingredients (e.g., chicken, pasta) and calculate the average sentiment score for each group using the sentiment analysis models from Task 4. Visualize the sentiment trends to identify which cuisine types or ingredients tend to receive more positive or negative reviews. Discuss any significant patterns or trends and explore how these insights can be used to improve recipe recommendations for users with specific preferences.

8. Review-Based Recipe Similarity Calculation: Calculate recipe similarity based on the semantic content of the user reviews and recipe description using methods of TF-IDF, Word2Vec, or BERT embeddings.

9. Cluster Analysis and Sentiment Trends: Use clustering techniques such as K-Means or Agglomerative Clustering to group recipes based on their descriptions and user reviews. Then, for each identified recipe cluster, analyze the most common tags, ingredients, and sentiment trends. Visualize the average sentiment and rating distribution across clusters to identify popular or highly-rated recipe groups. Discuss any significant patterns or trends related to user preferences within different clusters.

10- Rating Prediction using User Review: Write a script to use a machine learning model that predict user’s rating based on the provided user reviews from the available dataset (can use 80% for training the model and 20% for testing). Feel free to use any feature extraction model to extract features from recipes and also machine learning model for final prediction. Report and visualize the result in terms of F1 score and confusion matrix and discuss the accuracy between different rating groups.

11- Average Rating Prediction based on recipe feature: In this task, the goal is to calculate the average rating of each recipe based on ratings provided by previous users. First, calculate the average recorded rating for each recipe using the RAW_interactions dataset. Next, select and apply feature extraction methods to represent various parts of the recipe data, such as the recorded reviews, tags, description, title, ingredients, and other relevant features. Using these extracted features, train a machine learning model to predict the average rating for each recipe. After building the model (can use 80% for training the model and 20% for testing), report and discuss the results. Also, analyze the importance of each feature in the prediction, discussing which features have the most significant impact on the model's performance.

12- Advanced NLP Techniques and Suggestions: Feel free to explore other NLP techniques or state-of-the-art machine learning models to enhance the rating prediction in Task 11. You can experiment with transformer-based models such as RoBERTa, DistilBERT, or GPT, for feature extraction and comparing their performance to the previously used methods like BERT or traditional word embeddings.

13-Identify appropriate literature to discuss the findings and comment on the strength and weakness of the data processing pipeline.