Subject: Information Retrieval and Text Mining

Approach and Reasoning

Project Approach:

Problem: One new paper company wants to organize the recipes presented on Reddit and upload them on their website. The problem is that all the recipes on the internet were mixed with ingredients, recipes and equipment. The company wanted output in a detailed and organized manner where all the recipe comments need to be checked and divided into different entities like separating dishes, recipes and equipment from each other from all the recipe documents.
The company wants to employ a machine learning model for doing that task as doing it manually requires a lot of staff and effort as comments were in huge numbers.

Input: Three datasets were received as input. One dataset is full of unlabelled data with 800 examples. ChatGPT LLM annotated another dataset and the last dataset was annotated by a group of members annotated at a Python conference. Whole freedom was given to use any single dataset or use a hybrid dataset of two or three datasets.

Output: Output must have a model trained and annotated along with the output as a zip file with a write-up including annotation guidelines and project approach.

Procedure: 1. Firstly we selected an unlabelled dataset and decided to work on it as we did not trust GPT annotations. So we went with annotating around 375 unlabelled data. It exceeded 50% data of the dataset.
2. Then we saved those annotations and exported them into a JSONL file “annotated data.jsonl”. We splitted the dataset into two sets with 70:30 for training dataset and target dataset. Then we converted the datasets into json files using a python program.
3. We converted training_data.json and target_data.json files into spacy models using a python program. Then we used them to construct a basic model by passing them to model via config file. We got 50% accuracy.
4. We thought about improving the model further, so we used the pydata.json dataset too. In the same way, we divided data into training and target datasets with 70% and 30% data from pydata dataset.
5. Now we repeated the same steps by converting those datasets into training_newdata.json and target_newdata.json by using the same python program.
6. Using python program, then we converted json files into spacy model files and passed them as new data sets for training and improving the model which was already trained on the labelled dataset.
7. With this model was iterated and now having knowledge of two datasets. Its accuracy improved from 50% to 57%. There was only significant difference between the output as we did not corrected the pydata dataset. We just trained it directly due to lack of time.

Annotation Guidelines

Definitions of Entities: The first guideline of annotation is to define all entities. Defining is the most vital and basic step as it helps and guides the persons with no prior knowledge to understand and perform annotations by giving them a basic understanding of what to do.

There are three entities in our project. They are:

1. Ingredients: These are the basic components or substances used to prepare a dish or a recipe. Ingredients keep varying depending on the dish we choose to cook.
2. Dish: A dish is a food item specifically served as a single item during a meal. It is basically like a final product that we aim to cook using ingredients and equipment and the user consumes it at last.
3. Equipment: Equipment refers to the tools and machinery used in food processing and preparing activities. Most of the equipment includes electric devices and will help us convert raw materials into dishes.

Example: Let us take an example. I would like to make a chicken biryani. I need chicken, rice, oil, tomatoes, cooker, masalas. chicken biryani is a dish and the cooker is the equipment. Rice, oil, tomatoes, and masalas are ingredients.

Problems faced/ Difficult scenarios:

- **Mac and Cheese -**Unfamiliar persons will make these mistakes: Mac and Cheese two ingredients ❌ Mac and Cheese - one ingredient ✅
- **Dish-** Sometimes, the dish also plays a vital role in deciding. If you are making fried rice, Ketchup is an ingredient but if you are making ketchup then ketchup is a dish.
- **Combination-** Sometimes having a combination is very confusing. Salted meat - ingredient or salt, meat- ingredients. Salted meat does not exist, only meat exists.
- **Context-** In some instances, the same item is considered an ingredient in some cases but also a dish in other cases. Rice is an ingredient for biryanis but fried rice is a dish.
- **New Cusines-** If you are not familiar with recipes from global cuisines then you cannot decide whether it is a dish or an ingredient until you google it. Bibimbap is a Korean dish, but others cannot decide until they browse it.
- **Different Names-** The same substances will have different names in different languages. Coriander is widely known in Asian countries but the majority of them do not know what cilantro is.
- **Ambiguous Terms-** Rolls are food but may be anything like Swiss rolls, bread rolls etc. If you include just rolls then there is a doubt on whether to take it as a dish or an ingredient.
- **Ingredients as Equipment**-While mortar and pestle are primarily used for grinding ingredients, they can also be considered equipment in some contexts. Determining how to categorize was challenging
- **Preparation methods-** If you encounter grilled chicken or baked/ fried. It would be confusing what to take as an ingredient. Ingredient- chicken or grilled chicken?