In this task, I worked on refining a list of candidate dish names for the Italian cuisine using an automatic labeling process. The goal was to improve the initial list by removing false positives, correcting false negatives, and potentially adding new dish names. Additionally, I expanded the list of dish names by applying pattern mining and word association techniques, including SpaCy for text processing and word2vec for semantic analysis.

Task 3.1

The initial dataset provided in manualAnnotationTask/Italian.label consisted of dish names with associated labels. Some of the dish names were incorrect, either being false positives or false negatives. To refine the list, I performed the following steps:

1. Removed False Positives: Phrases that were clearly not dish names, such as "light rail" and "standard italian," were removed from the list.
2. Fixed False Negatives: Missing dish names like "carbonara" and "gelato" were manually added with positive labels to ensure they were recognized as valid dish names.

I applied basic filtering and manual inspection to identify non-dish names and missing dishes. This step helped ensure that the dataset better represented actual Italian dishes.

After refining the labels, I saved the updated list in the Refined_Italian.label file.

Task 3.2

To expand the list of dish names further, I used pattern mining and word association techniques:

1. Pattern Mining with SegPhrase: I applied the SegPhrase framework to extract potential dish names from the reviews dataset (yelp_academic_dataset_review.json). SegPhrase merges consecutive words based on statistical significance and provides quality scores for phrase candidates. Using the refined labels as input, I applied this technique to identify and classify potential new dish names.
2. Word Association with Word2Vec: To complement SegPhrase, I used word association methods like word2vec, which captures the semantic relationships between words. Using word embeddings, I calculated the cosine similarity between words and phrases to identify new dish names that are semantically similar to existing ones. This technique helped uncover additional dish names that may not have appeared in the pattern mining phase.
3. Data Subsampling: Due to the large size of the reviews dataset, I sampled 1% of the review data to ensure that the processing time remained manageable while still providing meaningful results.

Parameters and Methodology

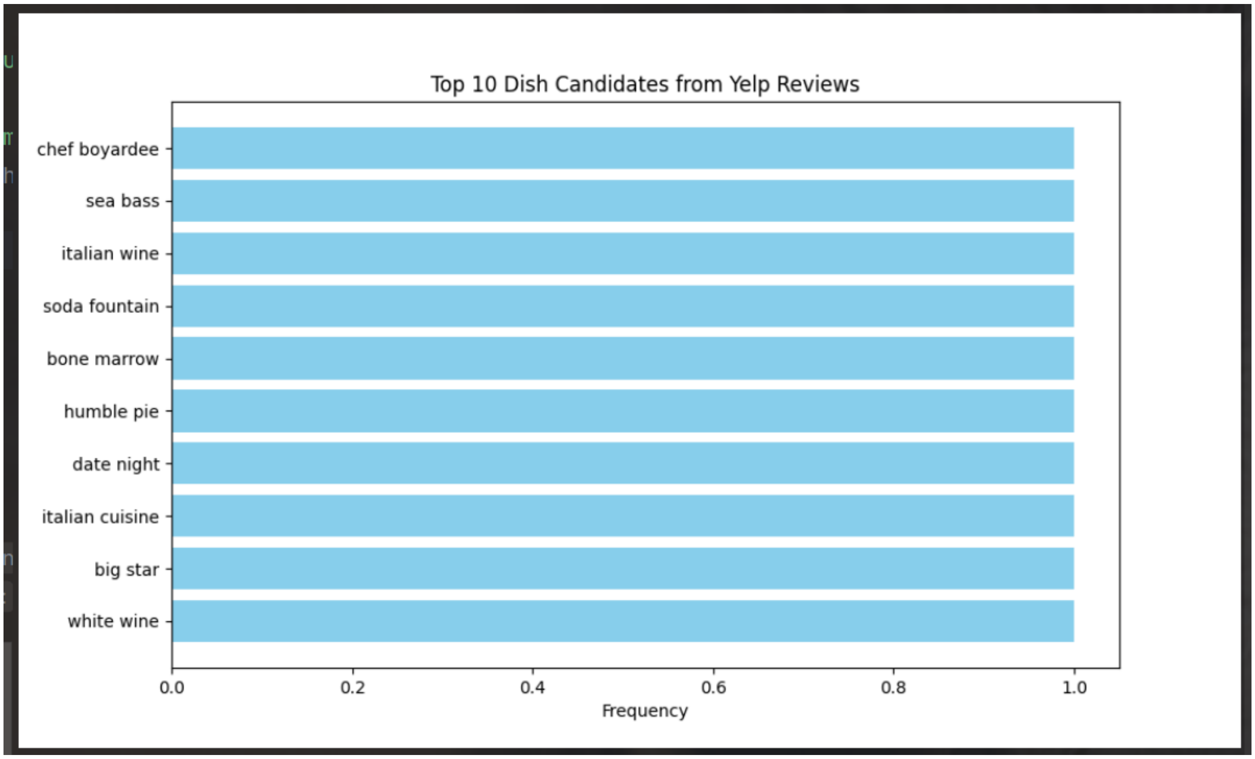
- SegPhrase: The classifier was applied with the refined dish names as the initial list. Statistical features like frequency, co-occurrence, and segmentation were used to score potential dish names.
- Word2Vec: I used the pre-trained word2vec model to compute the similarity between phrases and other dish-related words. The model helped expand the list with semantically similar terms.
- Data Sampling: 1% of the reviews were randomly sampled to make the processing more efficient and feasible for the given task.

Results sample

chef Boyardee, sea bass, italian wine, soda fountain, bone marrow, humble pie, date night

Conclusion and Evaluation

While the results were generally satisfactory, further tuning of parameters or exploring other advanced word association methods might lead to even more accurate and comprehensive results. Additionally, applying these techniques to other cuisines or datasets could further validate their effectiveness.


Top 10 Dish Candidates from Yelp Reviews
chef boyardee
sea bass
italian wine
soda fountain
bone marrow
humble pie
date night
italian cuisine
big star
white wine
Frequency