FOODBERT REPORT

Group-08

Gunjan Panda	Shubhangi Agrawal	Hardik Kumar
MT22099	MT22126	2020506

Objective:

Creating Bert embeddings to capture the culinary knowledge encapsulated in recipe instructions and the ingredients section.

1. *Creating average Bert Embeddings for all the ingredients in the Recipe Instructions & Ingredients Section.*
2. *Finding the cosine similarity among all pairs of ingredients.*
3. *Manually evaluate ingredient pairs with Top X & Bottom X pairs based on a reproducible protocol.*

Methodology:

1. **Data Compilation**
   Combine data for 'South-American' Cuisine based on the common field 'Recipe-ID' field, getting 'RecipeID', 'Ingredients' and 'Recipe Instructions' as the final set of features.

Input Data:

1. Top5_region.csv - Having details regarding Recipe ID and Regions for Top-5 Cuisines: Italian, Mexican, South American, Canadian and Indian Subcontinent.
2. Top5_region_Ingredients.csv - Having details regarding Ingredients & Recipe Instructions corresponding to each Recipe.
3. tdata - data having recipe IDs, Ingredients and Ingredient Phrases of all Top-5 cuisines.

	recipe_no	ingredient_Phrase	Ingredient
0	3697	2 fresh Dungeness crabs , cleaned and with the...	dungeness_crab
1	3697	2 teaspoons ground turmeric	turmeric
2	3697	1/2 teaspoon salt	salt
3	3697	1 tablespoon mustard seed	mustard_seed
4	3697	1 tablespoon hot water	water

2. **Data Pre-Processing**
    - Clean Ingredients by converting multi-word ingredients into single word ingredients.

ingredient	Ingredient
dungeness crab	dungeness_crab
turmeric	turmeric
salt	salt
mustard seed	mustard_seed
water	water

    - Converting ingredients of each recipe into meaningful sentences making it easier for BERT to capture context against each ingredient. Ex: This recipe contains egg milk as ingredients.

```
This recipe contains plantain butter and mozzarella_cheese as ingredients
This recipe contains oil plantain egg white_sugar vanilla_extract purpose_flour baking_powder salt and queso_fresco as ingredients
This recipe contains olive_oil onion celery carrot quinoa water bay_leaf lemon_zest lemon_juice green_pea salt and black_pepper as ingredients
This recipe contains prune bacon_strip and bamboo_skewer as ingredients
```

    - Clean Recipe Instructions by removing digits, extra spaces, punctuations and special characters.
    - Converting all multi-word ingredients present in cleaned instructions to single-word ingredients to later count occurrence of each ingredient throughout all the cleaned instructions.
      Ex: all "mozzarella cheese" occuring in Instructions, gets replaced with "mozzarella_cheese".

```
'plantains ends cut off and peels removed tablespoon butter melted ounces mozzarella_cheese cut into inch slices',
'cup canola oil cup water teaspoon salt cup tapioca starch egg cup plain yogurt cup grated Parmesan cheese cup grated mozzarella_cheese',
```

    - Divide Instructions into a list of sentences, finally getting 7697 clean instructions.

After Cleaning Instructions:

```
'guavas peeled quarts water cups white sugar cup fresh orange juice plantain leaves for wrapping optional'
```

3. **Count Ingredient Occurrences per Recipe Instruction**
   Tally occurrences of unique ingredients in all recipes, creating a dictionary with ingredients as keys and counts as values. Use this dictionary to calculate Bert embeddings.

```
'salt': 3295, 'onion': 2755, 'beef': 2552, 'pepper': 1989, 'water': 1889, 'olive_oil': 1443, 'black_pepper': 1364, 'butter': 1042, 'garlic': 980
```

4. **Calculate Bert Embeddings**
    - As Bert has high computation power, randomly sample 100 recipes for ingredients with over 100 occurrences. For ingredients having occurrences less than 100, consider all recipe instructions using the ingredient to calculate the Bert Embeddings.
    - Calculate BERT Embeddings using the Embedding4BERT Library for each key (i.e., 'Ingredient') in the generated dictionary. Multiple embeddings are computed

for each ingredient, taking into account its corresponding context from the sampled instructions.

- The final embedding is obtained by averaging out the BERT embeddings and is saved in a dictionary containing all ingredients along with their respective average BERT embeddings.
  For Example:

```
Processing Ingredients:   0%|          | 0/1 [00:00<?, ?it/s]*****Current ingredient is***** yucca_root

Batches:   0%|          | 0/4 [00:00<?, ?batch/s]359 tokens: [CLS] lbs boneless skinless chicken_breasts
259 tokens: [CLS] lb white_fish tilapia or halibut recommended lb shrimp container lump_crab salt to add
381 tokens: [CLS] small red_onions cut in half and sliced very thin limes juice only tablespoon sunflower
112 tokens: [CLS] quarts chicken_stock yucca_root peeled and cut into inch cubes plantains peeled halved

Batches: 100%|██████████| 4/4 [00:09<00:00,  2.26s/batch]
Processing Ingredients: 100%|██████████| 1/1 [00:09<00:00,  9.08s/it]Total Number of words in sampled ser
lbs boneless skinless ['chicken', '_', 'breasts'] cubed medium potatoes halved carrots sliced into thick
Total Number of embeddings in sampled sentences:  872
Total Number of occurences for yucca_root is: 13
Shape of Overall word embedding for yucca_root is:  (768,)
Ingredients with IndexError: []
```

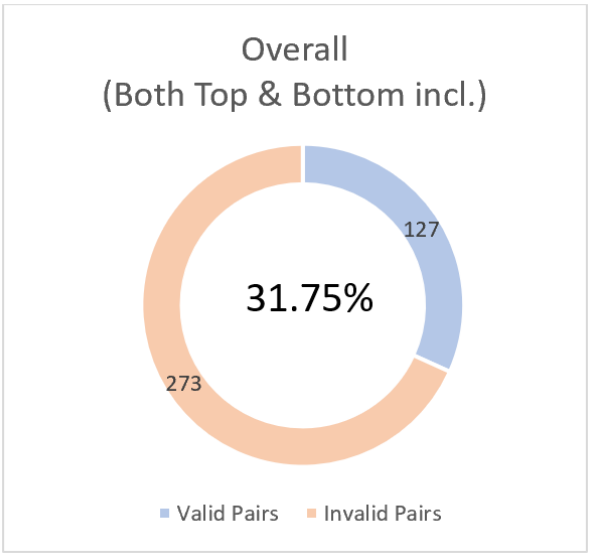
- BERT Embeddings generated stored as pickle file named as "sa_all_ing_embeddings.pkl"

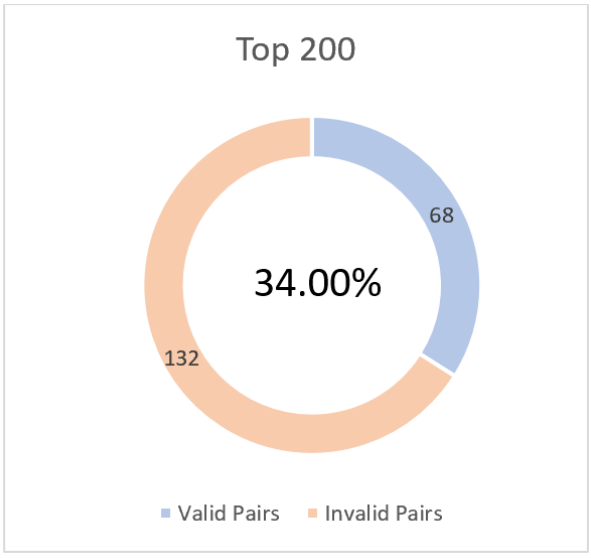
5. **Finding the cosine similarity among all pairs of ingredients**
   Cosine Similarity to determine the similarity between all pairs of ingredients based on their BERT embeddings. Store the most similar ingredient pairs together in a csv named "most_similar_df_all_occ.csv".

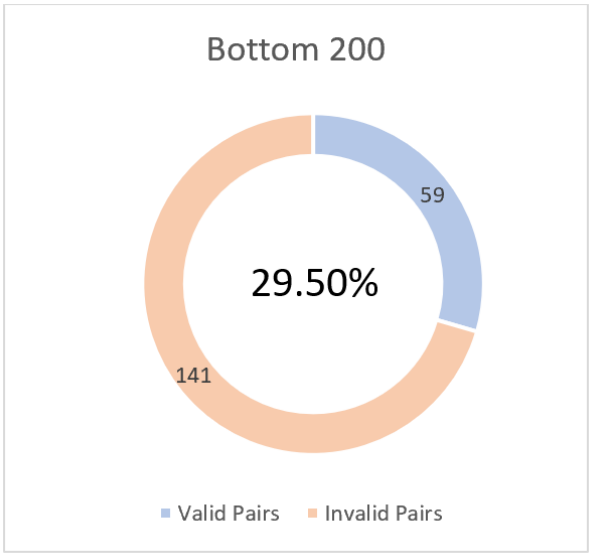
6. **Manual Verification of Bert Embeddings**
   Upon analyzing the Top 200 and Bottom 200 ingredient pairs manually, the following points were adhered to:
   - The South American cuisine so studied follows a Positive Food Pairing due to which, apart from consciously mapping ingredients, their flavor pairing was also checked on FlavorDB to see if a pair of ingredients share a good amount of flavor molecules or not (for our experiment, a count >=90 was chosen to qualify this). E.g. (Oregano, Marjoram): 151, (Red Grape, Great Northern Bean): 138, (Coconut Milk, Canola Oil): 100.
   - Many ingredients were listed in their processed or derivative form, and FlavorDB does not consist of such an exhaustive list of ingredients, due to which their respective generic ingredients were considered. E.g., Brown Sugar-> Sugar, Tomato Puree-> Tomato, Onion Powder-> Onion, Chuck Roast-> Beef Processed.
   - Online resources (Google Search) were also referred to look for substitutes or other names for an ingredient. E.g. (Kumara-> Sweet Potato), (Lime Juice, Rice Vinegar).

- Also, ingredient pairs belonging to the same category were considered valid for the experiment. E.g.: (Stewing Lamb, Beef Bacon): Meat, (Champagne, Vermouth): Beverage.

Results:







Conclusion:

The above-followed approach adds a new angle to the ongoing experiment and can be considered further.

- The pre-processing was done, keeping the context-capturing nature of the underlying Bert Model which can be experimented further.
- The sampling of recipes was conducted for 100 and 500 recipes, whose results didn't deviate much, although the computation time was comparable, and so finalized with the sampling of 100 recipes. The experiment can even be tested for higher sampling numbers given the computational resources, resulting in better results.
- Also, the approach used for Manual Verification can be translated to the main experiment (Cuisine Fusion), with Top_5 cuisines, where even the ingredient pairs with less shared flavor molecules(can be argued to contribute as a part of a Contrasting blend of ingredients which is seen in Negative Food Pairing, given the set of cuisines considered, consists of both Positive and Negative Food Pairing.

OUTPUT files submitted:

1. Bert Embeddings calculated for Ingredients: **"sa_all_ing_embeddings.pkl"**
2. CSV generated having most similar ingredient pairs for all above Ingredients: **"most_similar_df_all_occ.xlsx" , Containing manually verified Top-20 & Bottom-20 Ingredient Pairs.**