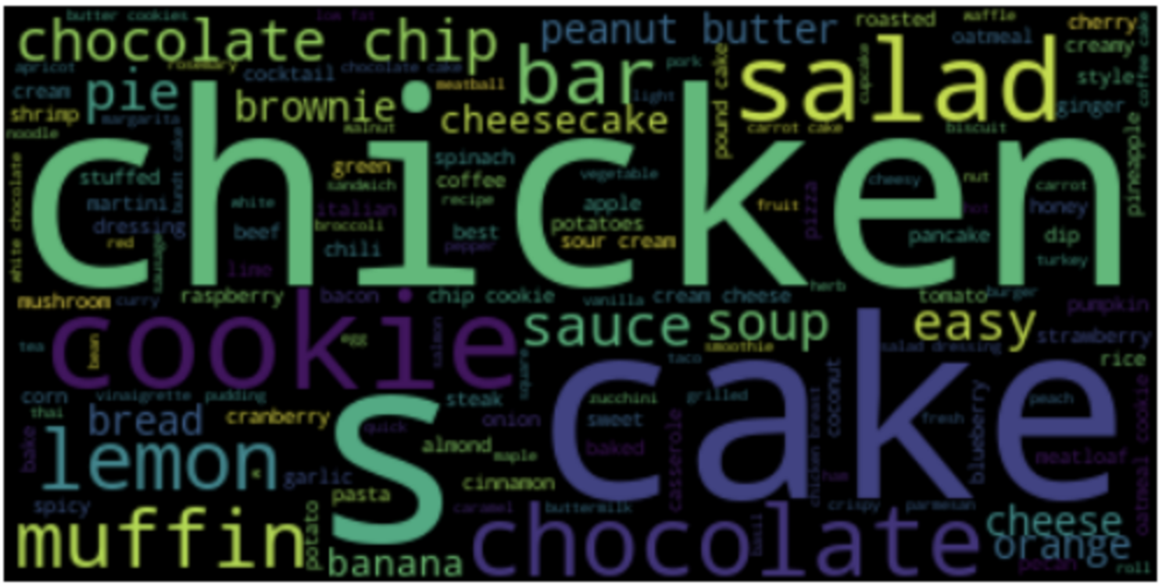
Let Machine Learning Categorize for you

One of the most interesting things I did during my four-month machine learning adventure was categorizing recipes into groups just by their names. It started as a simple homework assignment but quickly became a deep dive into data analysis. Instead of just ticking off a task, it turned into an exciting project that had me exploring different strategies and enjoying every bit of it!

Background on the project:
Cooking is a fun hobby often guided by recipe books that categorize dishes based on ingredients or types. However, when dealing with unsorted recipes, machine learning becomes a powerful tool, magically organizing the chaos for a smoother sorting experience.

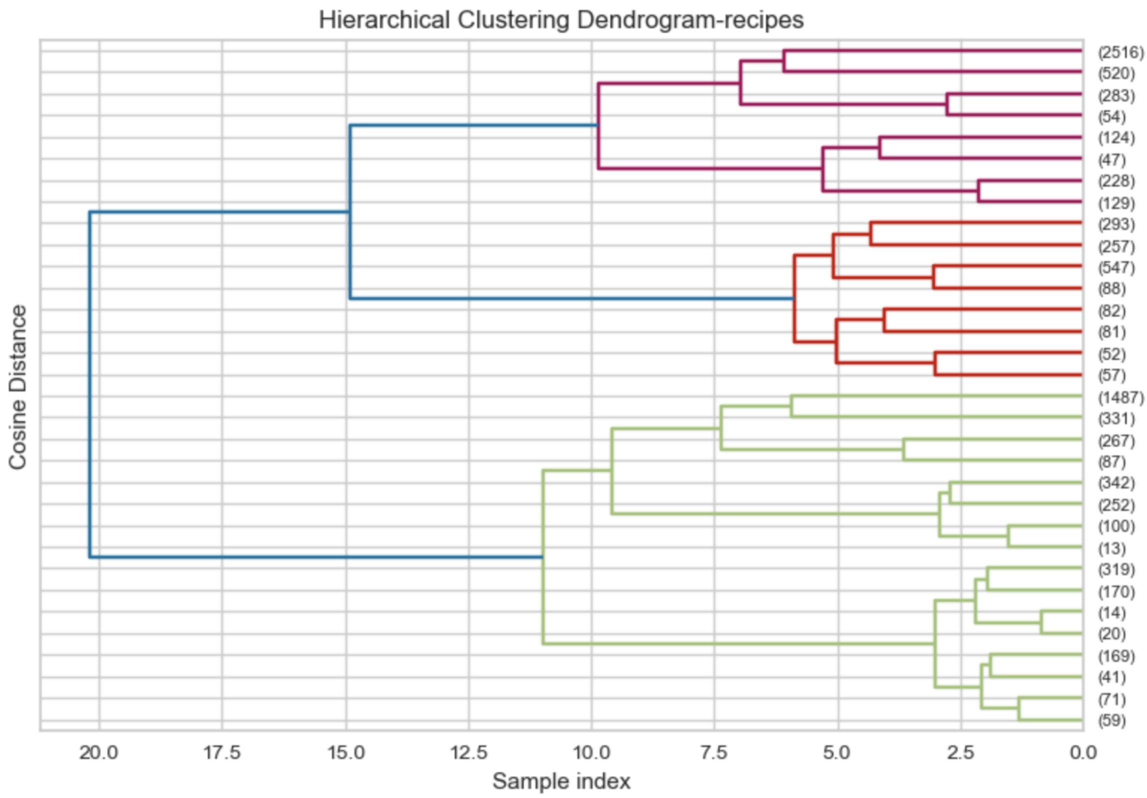
Dataset & Preprocess:
I utilized Kaggle's Food.com (https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions) recipes corpus as the example dataset for this project. Think of it as an unsorted recipe book, brimming with diverse recipes, steps, ingredients, and even reviews. However, my focus in this project was solely on the textual names of each recipe, ranging from the shortest "bread" to the longest "baked tomatoes with a parmesan cheese crust and balsamic drizzle." Additionally, I created a Word Cloud as below to visually depict the word frequencies within the recipe corpus.



Machine learning algorithms and models primarily operate on numeric representations of data for compatibility and computational efficiency. This is particularly essential when calculating similarities between data points, such as determining the resemblance of recipe names based on their distances. Thus, I used sentence embedding representation with the sentence transformer package, which takes into account the context of words and the semantic meaning of sentences to encode the recipe names for each example.

Machine Learning Model:
I employed the Hierarchical Clustering algorithm to address this problem, opting for its ease and flexibility in selecting the number of clusters for this specific scenario. To calculate the distance between recipes, I utilized the cosine distance metric because it is not influenced by the magnitude of each recipe numeric representation, making it more suitable for textural examples compared to Euclidean distance metrics.

The visualization of Hierarchical Clustering was achieved through a dendrogram, plotted with four levels. Additionally, I used ward linkage, a clustering criterion that combines clusters based on within-cluster variance, ensuring equally sized clusters at each level.



Result:

I printed the recipes in each cluster and I chose to flatten the dendrogram with 14 clusters (14 categories as result). (The print function is provided from CPSC330 Course Stuff from HW6 as a helper).

```
Cluster 1
----------
strawberry pound cake
shortcake with strawberries and banana
butter pecan bundt cake  marry me cake
mocha fudge layer cake
lemon lavender layer cake
bisquick pineapple coffee cake
lemon butter cake
a honey of a honey cake
courgette chocolate cake
marvelous marble cake   parve
```

```
Cluster 2
----------
natalie s chocolate chip cookies  easy and good
fabulous cut out cookies
old fashioned raisin filled cookies
easy chocolate oatmeal cake
big  fat  chewy chocolate chip cookies
martha s soft baked chocolate chip cookies
hallie s death by chocolate cookies
chocolate espresso cookies
best ever oatmeal cookies  land o lakes
rice krispie cookies
```

```
Cluster 3
----------
milk bread  bread machine
banana foster martini
zucchini banana bread
moist cornbread with cheese
easy strawberry bread
banana bread latte
banana maple nut coffee cake
rhubarb brown sugar loaf
creamy cornbread
crosby s orange marmalade gingerbread
```

```
Cluster 4
----------
banana chip muffins
chocolate chip devils food muffins
maple nut muffins
fat free blueberry bran muffins
banana protein muffins
corn bacon muffins
blueberry oat muffins
eggnog muffins with nutmeg streusel topping
          nut muffins
          t date nut muffins
```

```
Cluster 5
----------
chocolate chewy brownies
southern living basic yellow cake recipe   or chocolate
cherry swirl brownies
broccoli brownies
gooey one bowl brownies
these chocolate brownies are nuts
blueberry brownies
best ever brownies 6 ways
scrumptious brownies
maple frosted pumpkin blondies
```

```
Cluster 6
----------
belafonte bars
glazed honey bars
blue ridge blackberry lemon bars
low fat peanut butter s more bars
fudgy chocolate oatmeal bars
candied ginger   cardamom bars
chocolate   caramel oatmeal bars
date nut bars
razz ma tazz bars
delicious marble bars
```

```
Cluster 7
----------
wholesome peanut butter cookies
best ever cream puffs with vanilla filling
english toffee bars
bacon surprise cupcakes with maple frosting
lower fat peanut butter oatmeal cookies
stove top biscuits
chocolate snickerdoodles
skillet shepherd s pie
chocolate sour cream coffee cake w topping and glaze
cornmeal cookies
```

```
Cluster 8
----------
cherry gin gria
vodka creme brulee martini
morta de chocolata gelatini   death by chocolate martini
dreamsicle cocktail
wrong island iced tea
mike s candy apple martini
melting life savers cocktail
negroni cocktail
electric iced tea
cafe roma martini
```

```
Cluster 9
----------
whiskey peach smash
mocha coffee cooler
paula deen s broccoli coleslaw
garlic free ketchup
jack daniel s cedar plank salmon
blue lagoon
pork diane
blondies with variations
cocktail a la louisiane
mexican sunset
```

```
Cluster 10
----------
aloha  albacore tuna salad
salt and vinegar potato salad
ciro and sal s salad dressing
rsc 11   salad with a lime dressing
spinach pear salad w bacon and honey dijon dressing
ranch salad dressing
moroccan orange and carrot salad
st  louis salad
penny s french dressing
greek garden salad
```

```
Cluster 11
----------
lattice top chicken
keith moore s king ranch chicken
thai chicken meatballs with dipping sauce
sauteed chicken breast with clover honey and chili
barbecued chicken hash
easy italian chicken sandwich
crispy chicken with peanut dipping sauce
grilled raspberry chicken
balsamic glazed chicken
naan chicken sandwich
```

```
Cluster 12
----------
pepperoni pizza pancakes
spinach and artichoke dip   vegan
no tomato meatloaf and mushroom gravy
cream cheese braids
turkish style pizza
sausage   pepper mac n  cheese
creamy morel mushroom sauce
white chocolate boysenberry cappuccino
pancetta wrapped fish with grain mustard sauce
brazilian cheese puffs   pao de queijo  gluten free
```

```
Cluster 13
----------
hungarian three coin spinach potato soup
chickpea soup a la provencale
chicken   roasted red potatoes
meg o malley s irish parliament bean soup
ww loaded baked potatoes
blt soup
pear  celeriac and stilton soup
creamy white bean and chorizo soup
smooshed potatoes
mark bittman s chicken and rice soup
```

```
Cluster 14
----------
cranberry nut rolls
mediterranean shrimp  n  pasta
jalapeno chicken pita crisps
italian sausage skillet
bacon horseradish sauce
german black walnut balls
spicy moroccan chicken skewers
emeril s meatball soup
basic beef stew  very hearty
ranch roasted carrots
```

Next, I assign labels to each category based on the prevalent recipes within them, such as cake, cookie, bread, muffin, brownie, bar, alcohol drink, salad, chicken dishes, soup, and main dish. However, I omit cluster 9 from the graph above as I encounter difficulty assigning general names to the recipes within this cluster. The content in this cluster appears to be somewhat blended, making it challenging to provide meaningful categorizations. Additionally, I combined cluster 12 and cluster 13 into the same category as main dishes. I have tried to flatten the dendrogram with different numbers of clusters (e.g., starting from 6 to 15), and I found that numbers around 10 will give us reasonable clusters. There are some clusters that Hierarchical Clustering consistently identifies across these different numbers of categories, such as cake and cookies.

Caveats for Result:
I didn’t get a chance to evaluate the model and obtain a score since I lack the correct category labels for all the recipes in the dataset. Thus, manual inspection and analysis are crucial for this project.

The first aspect that can be improved is the need for further exploration of the number of final categories. Some models, like DBSCAN, automatically choose the number of clusters for you, but they also involve other hyperparameter tuning. I observed that by increasing the number of clusters using Hierarchical Clustering, more specific clusters may appear (e.g., muffin and brownie are categorized into the same cluster if the number of clusters is 10 instead of 14). However, it may also generate more scattered clusters that are challenging to generalize with a label. The trade-off in the number of clusters is worth exploring depending on how specific the clusters need to be.

Through manual inspection, I also discovered that not all recipes were successfully categorized into their expected clusters. For example, "Mexican Sunset" and "Blue Lagoon" are both cocktails, but due to their less descriptive names compared to "Negroni Cocktail," the model may consider them outliers and fail to categorize them into alcohol drinks. This issue arises because I only used the names of recipes from the raw dataset. Including steps and ingredients from raw dataset into this model training process might mitigate this problem.

Conclusion:
In conclusion, the clustering problem differs from other supervised machine learning problems, making it trickier to find a uniform scoring method or pipeline guideline to follow. I thoroughly enjoyed exploring the various decisions that can be made throughout the entire process. Further exploration could involve trying different combinations of linkage metrics, distance metrics, and the chosen number of clusters.

Reference:
The guide for the project is from CPSC330 course content HW6.
The word cloud code is adapted from: https://github.com/amueller/word_cloud