Data Clustering Project: Food.com Recipes with K-Means

by Aiden Seo, Dabin Im

December 7th, 2023

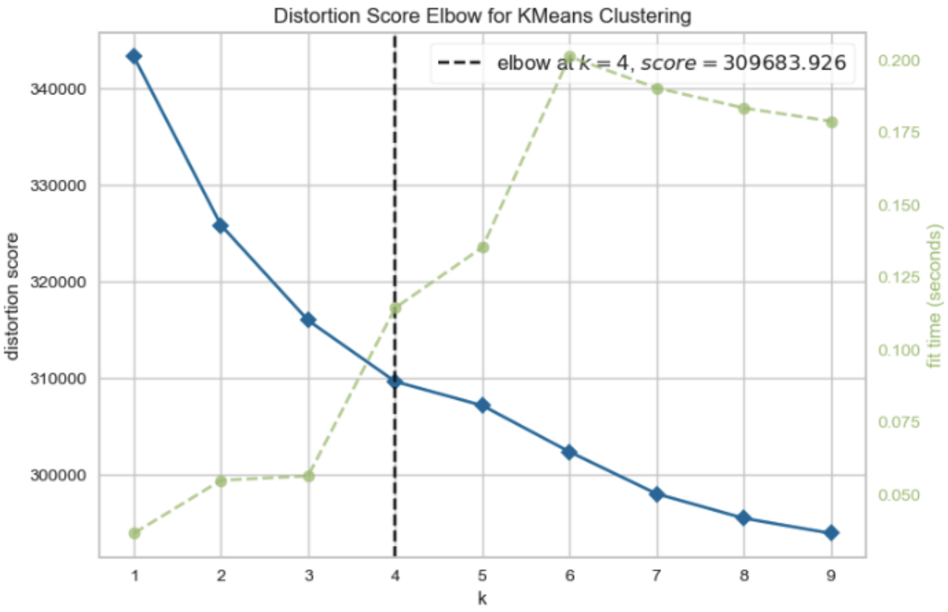
While studying the machine learning course, one of the most fundamental insights I gained is the ability to swiftly process massive amounts of data that would be challenging for humans to handle and even make decisions based on it. My team had the experience of building a model that classifies documents of Kaggle's Food.com recipes corpus, among various machine learning assignments. Natural Language Processing techniques apply machine learning in diverse language-based applications. Through constructing this document clustering model, we were able to group data without explicitly specifying representative labels for each cluster, enabling us to comprehend topics or content efficiently.

Dataset

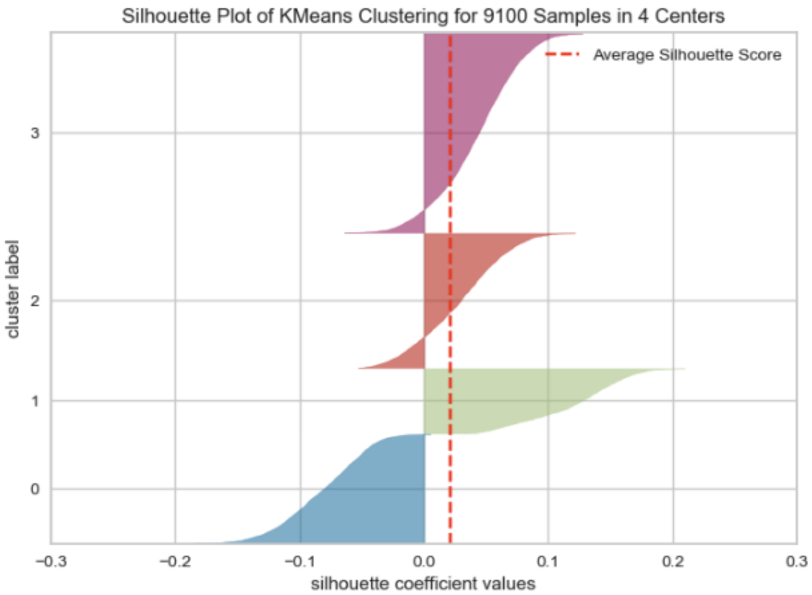
The dataset is available for download on Kaggle and consists of a substantial amount of recipes and recipe reviews uploaded to Food.com. Among the various files in this archive, only the file named 'RAW_recipes.csv' was utilized, representing the raw data of food recipes. Since there are over 230,000 examples in this file, a subset of 9,100 samples was selected for convenience in this modeling. Additionally, only one feature, the recipe name, was chosen from the 12 available features for use in this analysis.

K-Means Modeling

We didn't know the actual number of clusters. So we experimented with various models, including DBSCAN and Hierarchical clustering which are known to be more effective in such situations. However, in this particular problem, the K-Means clustering model demonstrated the best performance. To better represent the text data, we utilized the sentence embedding representation from a deep learning model. Since the K-Means algorithm requires specifying the number of clusters as a hyperparameter, we visualized an elbow plot to find the inflection point through visualization and used the yellowbrick library for this. We experimented with a diverse range of values for K, and the results from the model visualizer, specified in the range from 1 to 10, are as follows.



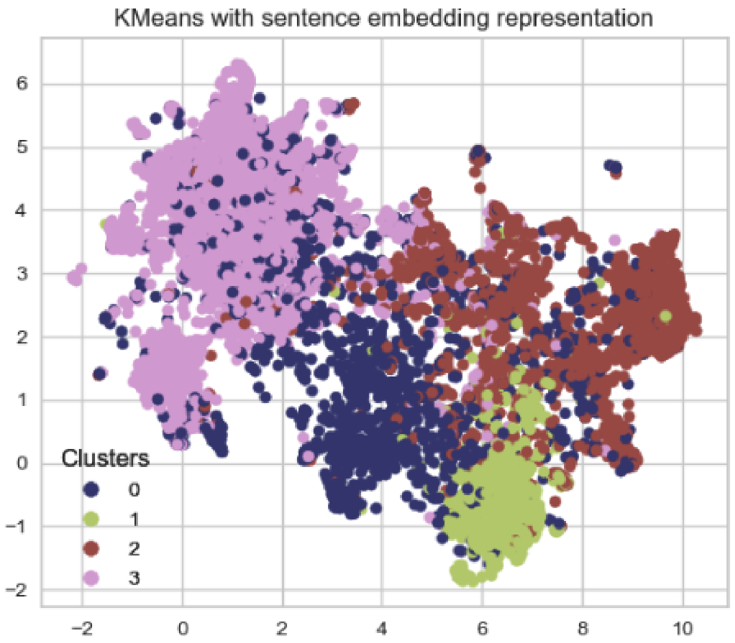
Unlike supervised machine learning, objectively measuring the quality of clustering is challenging due to the absence of a target. Observing the plotted data, considering 4 as the number of clusters appeared reasonable, as it signifies a point with relatively low inertia. The distortion score at this point was 309,683.926. To validate this decision, we tried another visualization method, specifically silhouette score plotting, as an alternative approach to finding hyperparameter K.



As seen in the above figure, there are noticeable drastic drop-offs, and lots of data points have silhouette scores less than 0, indicating that the results are not ideal. However, we can see that the number of samples representing the thickness of each cluster appears balanced. Moreover, as the number of clusters increases, examples become closer to neighboring clusters, leading to lower silhouette scores. Ultimately, the number of clusters in our model was determined to be 4. The average silhouette score was 0.02148306, which did not turn out as high as anticipated.

Results

This time, let's examine the results of the K-Means model with 4 clusters based on sentence embedding representations through visualization.



In the figure above, each data point is distinguished by its cluster label. The crucial aspect here is not which cluster label each point belongs to but rather which points share the same label. Due to the abundance of data, some points may overlap or intersect, but in general, we can say that close points were assigned the same label. Now, based on this, the output of the clustering is as follows.

Cluster 0	**Cluster 1**
royal tea room scones oriental almond ponzu slaw wuollet s brownie enormous minnesota brownies peposo alla fornacina pink prozac cocktail cafe vandermint gingerbread scones chocolate rose baked rigatoni dallas morning news buffett blueberry blaster margarita	key lime cupcakes grand marnier poppy seed cake light and fluffy lemon jello cheesecake cinnamon tea cake women s weekly marx brothers cheesecake island plunder rum cake cream cheese coffee cake muffins moon cake i double chocolate chip banana cake light easy mix butter cake

Cluster 2	**Cluster 3**
fruity tea loaf peanut butter pie semi homemaker recipe sandra lee super simple latkes potato pancakes sweet corn muffins cake mix chocolate peanut butter cookies cinnamon chip scone easy strawberries cream hot tea pecan puffs blueberry oatmeal cookies to your health muffins	smoky waldorf salad pear celeriac and stilton soup sunday dinner mashed potatoes spiral macaroni salad roasted beet salad with orange dressing shiitake mushrooms with scallions rotel quick beef sour cream casserole renee s meatloaf corn and tomato salad colorado style beef enchiladas

Despite a relatively low silhouette score of 0.021, our K-Means achieved meaningful clustering. With these results, we are now able to assign labels based on the information observed within each cluster. While there may be some inconsistency in some data within the clusters, we find it acceptable as the K-Means algorithm assigns cluster labels to outliers as well. The labels we obtained using the K-Means clustering model are as follows.

- Cluster 0: Desserts and cocktails
- Cluster 1: Cakes
- Cluster 2: Baking-related (cookies or muffins)
- Cluster 3: Salads and other main dishes

Things to ponder

Is our hyperparameter decision appropriate:
We utilized the K-Means clustering model, which lacks a precise quantitative method for finding the optimal hyperparameter. This implies the lack of a clear evaluation metric for the number of clusters, making it challenging to determine which cluster number is optimal.

Is the clustering result objective:
Even if we achieved effective clustering with a good model using an appropriate hyperparameter K, the results may not be entirely objective. While there might be a logical or optimal clustering, there is no definitive correct answer.

Text representation's limitations:
To achieve a good text representation, we leveraged the representation of a deep learning model. Text data lacks structure, and normalizing it for model utilization is a task that, while relatively easy for humans, is ambiguous and challenging for computers. Therefore, we cannot claim that the sentence embedding representation we used is 100% perfect.