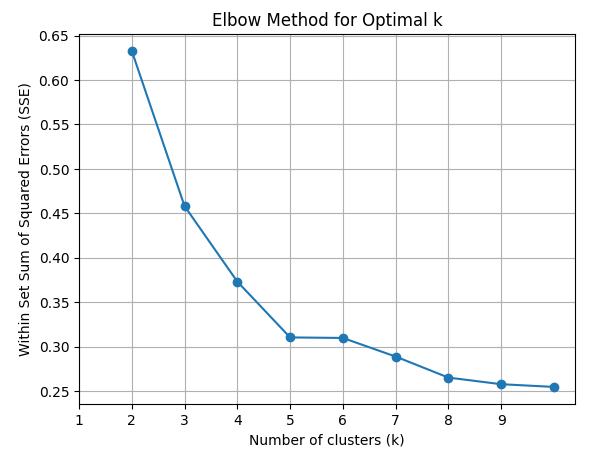
**Tyson Brown CS777 Term Project**

**Introduction**

The goal of this project is to develop a beer recommender based upon a beer review dataset of approximately 1.5 million records. Reviews consisted of a user’s perception of a beer’s aroma, appearance, palate, taste, and an overall review score. Based upon these criteria a user should be able to enter a beer name and obtain a list of beers that have related characteristics to enjoy.

**Methodology**

Two different methods will be used to develop this recommender. The first method will use Kmeans clustering to group the user reviews of each beer into clusters. Based upon these clusters, beers that fall within the same cluster can then be assumed to have similar characteristics and can be used in the recommender. The first step in implementing the Kmeans clustering algorithm was to use the “Elbow Method” to select the optimal number of clusters. Within Set Sum of Squared Error was used to select the best number for k. After running the optimization code, five clusters were determined to be the optimal amount.



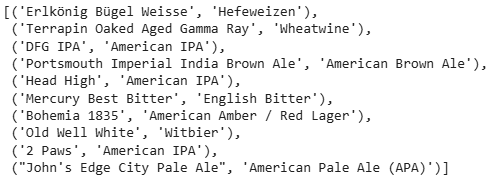
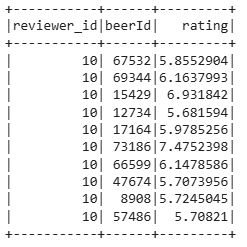
After the beers were separated into five clusters, then those cluster predictions were used as recommendations based upon other beers in the cluster. This returned a list of similar beers that a user could try that should have similar characteristics of other beers in the cluster based upon user reviews.

The second method Alternating Least Squares will be used in collaborative filtering of the beer dataset to develop a recommender system based upon historical reviews of users. The dataset had to be modified slightly to create an index of reviewers since the original dataset contained only a string value reviewer id. Only the overall review score was used to generate the model. Then the new indexed dataset was split into train/test sets with an 80/20 percentage split. The pyspark ALS method was then used to fit the model. The test data was used to obtain predictions which were evaluated using root mean squared error. Ten recommendations for all users were then generated via the recommendForAllUsers method. Then entering an individual reviewers id, the recommendations for that user were obtained.

**Results**

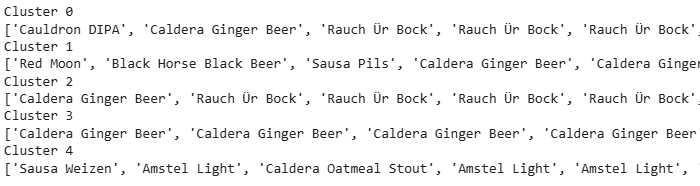
The Kmeans clustering results were problematic due to different reviewer’s perceptions of beer characteristics. This resulted in beers being assigned to multiple clusters. Beer recommendations were based upon the first cluster assignment for the beer submitted for recommendations.

The ALS model generated the top ten results for each user based upon the reviews of other users by computing a ratings score. These recommendations were then used to generate the beer names and styles.



**Discussion**

The inconsistency in reviews for the Kmeans clustering model makes relying on this type of model for recommendations unreliable considering beers could belong to multiple clusters and the cluster selected for recommendations may differ from the reviewer’s taste preferences. It seems like this type of recommender system would only be useful in situations where items can only be in one cluster. In the below results it can be seen that Caldera Ginger Beer is found in clusters 1-3.



The limitations of ALS discussed in lecture also hold true for this implementation. Users may not have reviewed every beer. Beers with a higher overall rating may be recommended more.

**Conclusion**

This relatively simple implementation of a recommender system exposes the complexities of such models. Clustering for this dataset may not be appropriate as items are grouped based upon subjective reviews resulting in assignment to multiple clusters. The ALS model worked as intended but also has its own limitations.