Combined Analysis of Clustering Results with and without Stemming:

Introduction: The analysis compares the clustering results obtained from the program data before and after applying stemming. Stemming is a preprocessing technique used to normalize text data by reducing words to their root or base form. The objective is to evaluate the impact of stemming on clustering effectiveness and identify any differences in the results.

Data Preprocessing:

- **Non-Stemmed Data:** The original program data was used without stemming.
- **Stemmed Data:** Stemming was applied to the program data to normalize the opcodes.

Clustering Algorithms:

- Dot Product Kernel Clustering
- Polynomial Kernel (Degree 2) Clustering

Combined Clustering Results:

1. **Non-Stemmed Data:**
    - Dot Product Kernel:
        - Cluster Assignments: [4 1 4 1 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
        - Average Silhouette Score: 0.9468
    - Polynomial Kernel:
        - Cluster Assignments: [4 1 4 1 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
        - Average Silhouette Score: 0.9796
2. **Stemmed Data:**
    - Dot Product Kernel:
        - Cluster Assignments: [4 1 4 1 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
        - Average Silhouette Score: 0.9468
    - Polynomial Kernel:
        - Cluster Assignments: [4 1 4 1 3 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
        - Average Silhouette Score: 0.9796

Difference Analysis:

- **Cluster Assignments:** The cluster assignments remain unchanged after stemming, indicating that stemming did not affect the grouping of documents into clusters.
- **Average Silhouette Scores:** The average silhouette scores are identical between non-stemmed and stemmed data for both dot product and polynomial kernels.

While performing stemming we noticed that the analysis took a longer time to computer and cluster as the major computation is done for the stemming apart form that we observed enhanced results and feature map reduction when using stemming.

If you map all similar words to a single token and execute the same clustering analysis, this will be the result

1. We observed a reduced feature space as most of the words are grouped by a same token.
2. As the feature space is reduced there will be a improvement in the cluster quality and performance.
3. The clusters will be more interpretable than before and the analysis takes lesser time to process.

Conclusion: The clustering results using dot product and polynomial kernels yielded similar cluster assignments. However, the average silhouette scores indicate that polynomial kernel clustering produced higher-quality clusters with better separation between clusters. This suggests that the polynomial kernel method may be more effective for clustering the program data based on similarity.