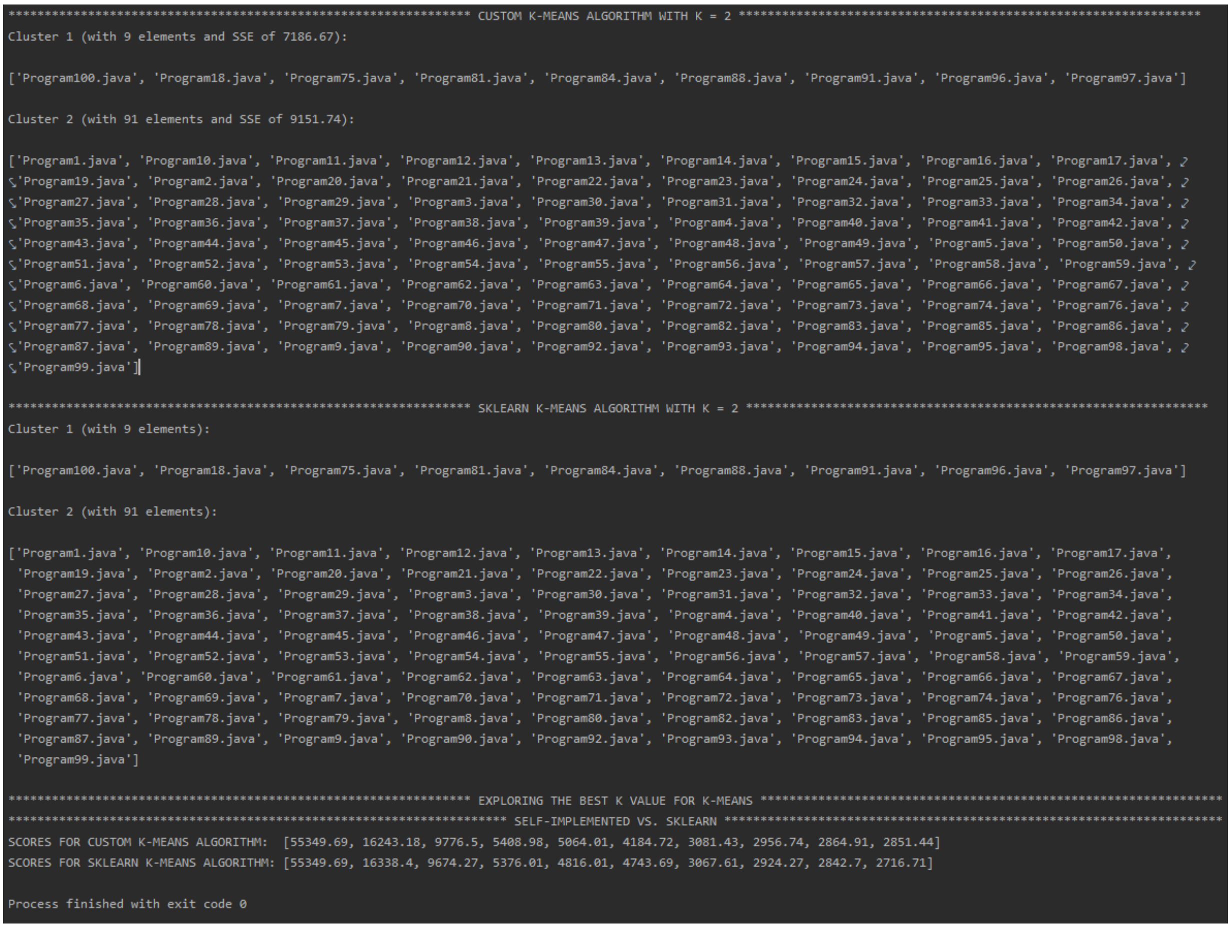
CSC-691 — Data Mining
Assignment 4

Esteban Murillo Burford

Thursday 24th October, 2019

Summary

For this assignment, everything related to the *k-Means* algorithm was implemented successfully. This time, a consensus has been reached between the two different versions of the algorithm (the one self-implemented and the one from the *sklearn* library), meaning that the generated clusters are almost always the same for both. However, it has been noted that in a small amount of scenarios, the elements belonging to the clusters vary slightly due to the nature of the algorithm. Either way, the results are satisfactory. Refer to Figure 1 for details about the program's output.



```
**************************************************************** CUSTOM K-MEANS ALGORITHM WITH K = 2 ****************************************************************
Cluster 1 (with 9 elements and SSE of 7186.67):

['Program100.java', 'Program18.java', 'Program75.java', 'Program81.java', 'Program84.java', 'Program88.java', 'Program91.java', 'Program96.java', 'Program97.java']

Cluster 2 (with 91 elements and SSE of 9151.74):

['Program1.java', 'Program10.java', 'Program11.java', 'Program12.java', 'Program13.java', 'Program14.java', 'Program15.java', 'Program16.java', 'Program17.java', 
'Program19.java', 'Program2.java', 'Program20.java', 'Program21.java', 'Program22.java', 'Program23.java', 'Program24.java', 'Program25.java', 'Program26.java', 
'Program27.java', 'Program28.java', 'Program29.java', 'Program3.java', 'Program30.java', 'Program31.java', 'Program32.java', 'Program33.java', 'Program34.java', 
'Program35.java', 'Program36.java', 'Program37.java', 'Program38.java', 'Program39.java', 'Program4.java', 'Program40.java', 'Program41.java', 'Program42.java', 
'Program43.java', 'Program44.java', 'Program45.java', 'Program46.java', 'Program47.java', 'Program48.java', 'Program49.java', 'Program5.java', 'Program50.java', 
'Program51.java', 'Program52.java', 'Program53.java', 'Program54.java', 'Program55.java', 'Program56.java', 'Program57.java', 'Program58.java', 'Program59.java', 
'Program6.java', 'Program60.java', 'Program61.java', 'Program62.java', 'Program63.java', 'Program64.java', 'Program65.java', 'Program66.java', 'Program67.java', 
'Program68.java', 'Program69.java', 'Program7.java', 'Program70.java', 'Program71.java', 'Program72.java', 'Program73.java', 'Program74.java', 'Program76.java', 
'Program77.java', 'Program78.java', 'Program79.java', 'Program8.java', 'Program80.java', 'Program82.java', 'Program83.java', 'Program85.java', 'Program86.java', 
'Program87.java', 'Program89.java', 'Program9.java', 'Program90.java', 'Program92.java', 'Program93.java', 'Program94.java', 'Program95.java', 'Program98.java', 
'Program99.java']

**************************************************************** SKLEARN K-MEANS ALGORITHM WITH K = 2 ****************************************************************
Cluster 1 (with 9 elements):

['Program100.java', 'Program18.java', 'Program75.java', 'Program81.java', 'Program84.java', 'Program88.java', 'Program91.java', 'Program96.java', 'Program97.java']

Cluster 2 (with 91 elements):

['Program1.java', 'Program10.java', 'Program11.java', 'Program12.java', 'Program13.java', 'Program14.java', 'Program15.java', 'Program16.java', 'Program17.java',
 'Program19.java', 'Program2.java', 'Program20.java', 'Program21.java', 'Program22.java', 'Program23.java', 'Program24.java', 'Program25.java', 'Program26.java',
 'Program27.java', 'Program28.java', 'Program29.java', 'Program3.java', 'Program30.java', 'Program31.java', 'Program32.java', 'Program33.java', 'Program34.java',
 'Program35.java', 'Program36.java', 'Program37.java', 'Program38.java', 'Program39.java', 'Program4.java', 'Program40.java', 'Program41.java', 'Program42.java',
 'Program43.java', 'Program44.java', 'Program45.java', 'Program46.java', 'Program47.java', 'Program48.java', 'Program49.java', 'Program5.java', 'Program50.java',
 'Program51.java', 'Program52.java', 'Program53.java', 'Program54.java', 'Program55.java', 'Program56.java', 'Program57.java', 'Program58.java', 'Program59.java',
 'Program6.java', 'Program60.java', 'Program61.java', 'Program62.java', 'Program63.java', 'Program64.java', 'Program65.java', 'Program66.java', 'Program67.java',
 'Program68.java', 'Program69.java', 'Program7.java', 'Program70.java', 'Program71.java', 'Program72.java', 'Program73.java', 'Program74.java', 'Program76.java',
 'Program77.java', 'Program78.java', 'Program79.java', 'Program8.java', 'Program80.java', 'Program82.java', 'Program83.java', 'Program85.java', 'Program86.java',
 'Program87.java', 'Program89.java', 'Program9.java', 'Program90.java', 'Program92.java', 'Program93.java', 'Program94.java', 'Program95.java', 'Program98.java',
 'Program99.java']

**************************************************************** EXPLORING THE BEST K VALUE FOR K-MEANS ****************************************************************
******************************************************************** SELF-IMPLEMENTED VS. SKLEARN ********************************************************************
SCORES FOR CUSTOM K-MEANS ALGORITHM:  [55349.69, 16243.18, 9776.5, 5408.98, 5064.01, 4184.72, 3081.43, 2956.74, 2864.91, 2851.44]
SCORES FOR SKLEARN K-MEANS ALGORITHM: [55349.69, 16338.4, 9674.27, 5376.01, 4816.01, 4743.69, 3067.61, 2924.27, 2842.7, 2716.71]

Process finished with exit code 0
```

Figure 1: Final results for both *k-Means* algorithms (custom and *sklearn*)

k-Means comparison

As it can be seen in Figure 2, the results are very similar for both implementations of the *k-Means* algorithm. All scores are within are reasonable range, save for $k = 5$ & $k = 6$, which differ a little bit between both algorithms. To see the actual values used for plotting, refer to Figure 1.

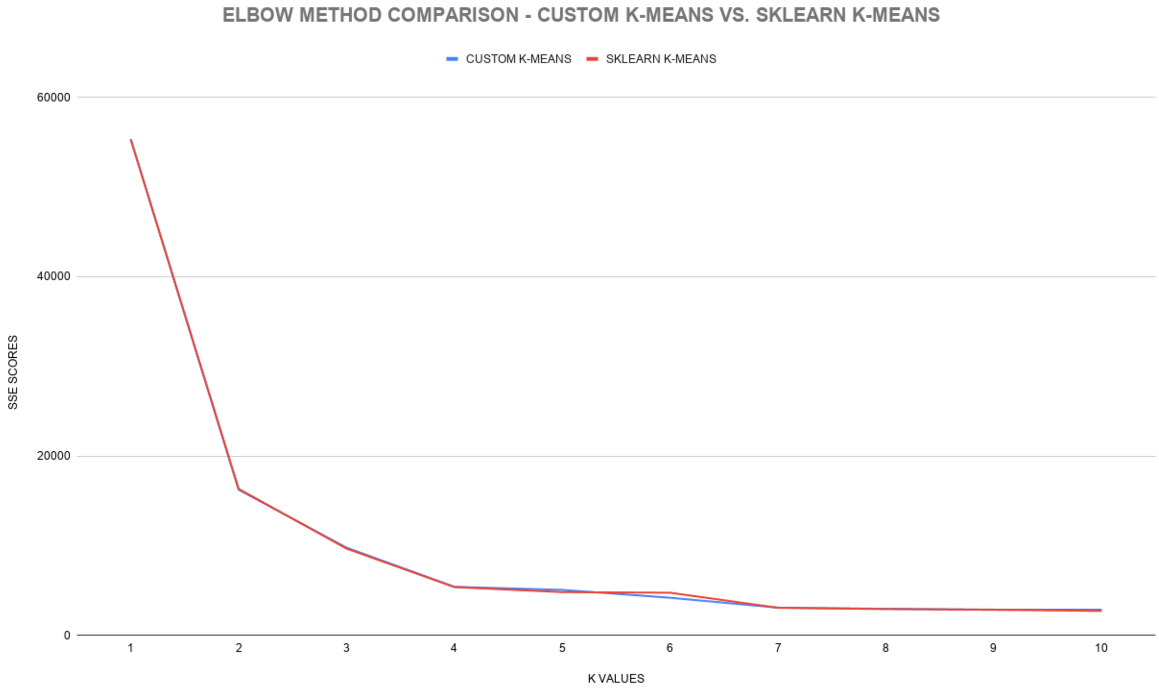


Figure 2: *k-Means* elbow method comparison — self-implemented vs. *sklearn*

Notes

- To modify the behavior of the program, change values in *config.py* accordingly
- Remember to run code using the **version 3 of the *Python* interpreter**

References

[1] *Beyond the k-Means – the Right k*. Retrieved on October 24th from https://www.edupristine.com/blog/beyond-k-means