Nick Petty
CAP6776 Homework 3

Accuracy = max (sum $\{C_k, L_m\}$ $T(C_k, L_m)$) /n
Where n = number of data points, C_k denotes kth cluster, L_m is the m^{th} class, and $T(C_k, L_m)$ is the number of data points that belong to class m and are assigned to cluster k. Because k and m refer to the number of classes in the data set, they must be equal.

Finding the maximum sum of $T(C_k, L_m)$, when given T in matrix form, is a variation on the Assignment Problem. Solving this problem via brute force will give a run time of m!, which is excessive. Instead, the Munkres algorithm (or the Hungarian Algorithm) is used. This gives a run time of m^3, which is much better.

To use the Munkres algorithm, the Python munkres package has been imported.
Package documentation: https://pypi.python.org/pypi/munkres/
The matrix from $T(C_k, L_m)$ does not immediately work with this algorithm, as it is designed to find the minimum sum. A second matrix, composed of a very large number minus each cell value in the T matrix must also be created. This larger-valued matrix, when given to the algorithm, will then find the maximum sum.

The number of data points, n: 1484

$T(C_k, L_m)$ as a k by m matrix, where each row is a cluster, and each column is a class label:

		Class Labels									
		1	2	3	4	5	6	7	8	9	10
Clusters	1	88	30	47	0	0	4	0	1	2	0
	2	2	4	1	0	0	87	2	4	5	0
	3	13	43	46	0	2	32	1	3	2	0
	4	2	49	12	0	0	5	0	0	0	0
	5	21	89	163	0	2	4	2	6	5	0
	6	12	112	91	0	0	2	3	4	0	0
	7	62	8	12	1	0	1	1	1	1	0
	8	28	31	61	1	1	22	15	8	2	0
	9	1	60	28	0	0	5	1	0	1	0
	10	15	3	2	42	30	1	26	3	2	5

$T(C_k, L_m)$ cells and values that make up the maximum sum:
(1, 1) = 88
(2, 6) = 87
(3, 8) = 3
(4, 5) = 0
(5, 3) = 163
(6, 2) = 112
(7, 10) = 0
(8, 7) = 15
(9, 9) = 1
(10, 4) = 42

511

Accuracy for clustering then equals this maximum sum divided by n.
511/1484 = 0.344339622642