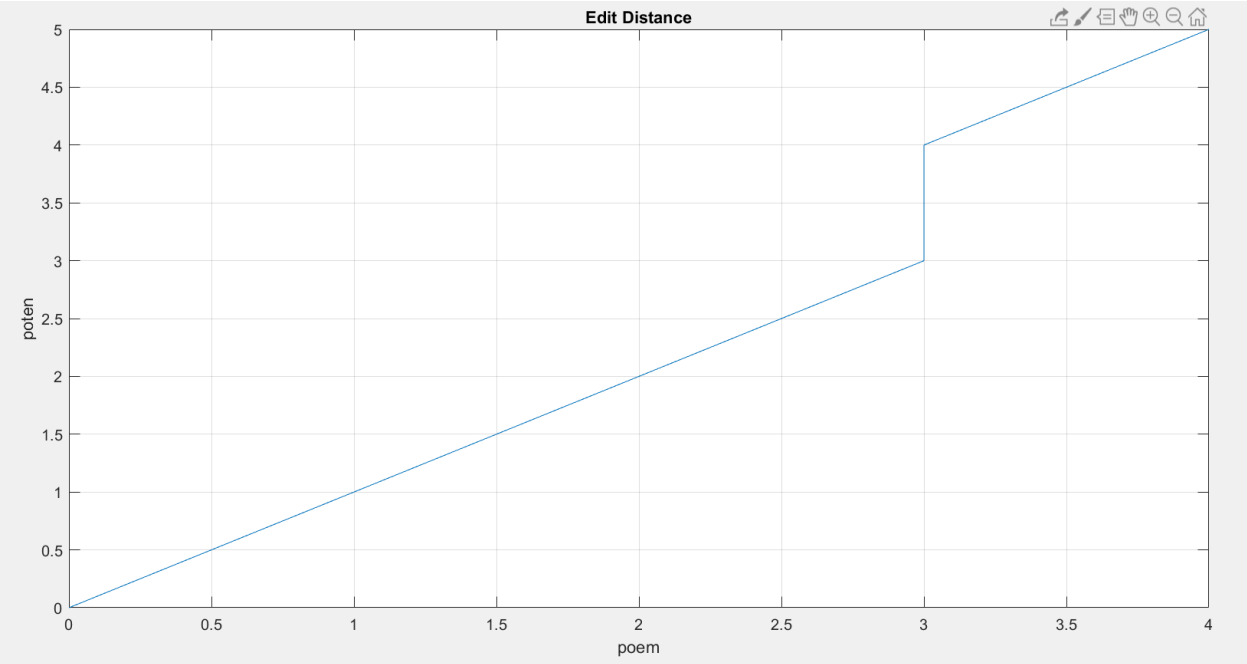
Problem 8.1

The edit distance is shown in the below figure. The edit distance cost is 2 which is shown in the 5th row and 6th column in the table after the figure. The table gives the "distance matrix" which represents the optimal path one would take. "2" representing the deletion of "t" and change of "n" to "m".



(Edit Distance)

0	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	1	1	2
4	3	2	2	2	2

(Distance Matrix)

"Estimating the number of clusters in a dataset via the gap statistic"

1. The first choice is to generate each feature uniformly based on the range of observed values. The second choice is to generate features based on a uniform distribution with a box aligned as well as principal components of the data.
2. The two distributions can be as close as 0.0002 apart from each mean before the gap statistic fails to differentiate them.