4 Programming: Text Clustering

EM for Mixture of Multinomials

1. Expectation

$$\gamma_{ic} = \frac{\pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}{\sum_{c=1}^{n_c} \pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}$$

2. Maximization

$$\mu_{jc} = \frac{\sum_{i=1}^{n_d} \gamma_{ic} T_{ij}}{\sum_{i=1}^{n_d} \sum_{l=1}^{m_w} \gamma_{ic} T_{il}}$$

$$\pi_c = \frac{1}{n_d} \sum_{i=1}^{n_d} \gamma_{ic}$$

Task [20 pts]

- I have attached my matlab code to the homework submission on Canvas.
- I tried to reduce the number of for loops as much as I could and perform full matrix operations instead whenever possible. I also picked 1e-30 as the difference threshold between the parameters after a loop to determine that the algorithm has convereged. When using isequal instead, it would take a very long time to fully converge for some initial parameter values.
- These choices proved very beneficial, with total running time of the algorithm ranging from 0.1 to 0.5 seconds.
- The algorithm performed fairly well with the test data set, with an average accuracy of 78.3483% across 150 test runs. Here are the complete statistics for the 150 test runs (with randomally initialized parameters):

Mean Accuracy	Median Accuracy	Min. Accuracy	Max. Accuracy	Accuracy Mode
78.3483%	79.125%	61.75%	91.25%	78%