Document Clustering

In this question, you solve a document clustering problem using unsupervised learning algorithms (i.e., soft and hard Expectation Maximization for document clustering**.**)

EM for Document Clustering

Task 1
Derive **Expectation** and **Maximisation** steps of the hard-EM algorithm for Document Clustering show your work in your submitted report. In particular, include all model parameters that should be learnt and the exact expression (using the same math convention that we saw in the lecture) that should be used to update these parameters during the learning process (ie., E step, M step and assignments).

Task 2
Implement the hard-EM (you derived above) and soft-EM (derived in notes **5**). Please provide enough comments in your submitted code.
Hint. If it helps, feel free to base your code on the provided code for EM algorithm for GMM in reference 2 or the reference 1 provided).

Task 3
Load **TaskA.text** file and necessary libraries (if needed, perform text preprocessing similar to what we did in reference 3), set the number of clusters K=4, and run both the soft-EM and hard-EM algorithms on the provided data.

Task 4
Perform a PCA on the clusterings that you get based on the hard-EM and soft-EM in the same way we did in reference 3. Then, visualise the obtained clusters with different colors where x and y axes are the first two principal components (similar to reference 3). Save your visualizations as plots in your Notebook file.

Submission:
The files that you need to submit are:
1. Jupyter Notebook file containing the code for question with the extension ".ipynb".
2. You must add enough comments to your code to make it readable and understandable by the tutor.
3. A PDF file that contains your answer to task 1 (it is the only pen & paper question in this question, all the other questions are answered in Notebook file.