Due date: **Mar. 4, 2020, 11:59 PM** (Arlington time). You have **two** late days throughout the semester — use it at as you wish. Once you run out of this quota, the penalty for late submission will be applied. You can either use your late days quota (or let the penalty be applied). Clearly indicate in your submission if you seek to use the quota.

What to turn in:

1. Your submission should include your complete code base in an archive file (`zip, tar.gz`) and `q1/, q2/,` and so on), and a very very clear README describing how to run it.
2. **A report (typed up, submit as a PDF file, NO handwritten scanned copies) describing what you solved, what you implemented and your own interpretation of the results. Your report is the primary material that we will look at to assign scores.**
3. Submit your entire code and report to Canvas.

Notes from instructor:

- Start early!
- You may ask the TA or instructor for suggestions, and discuss the problem with others (minimally). But **all parts of the submitted code must be your own**.
- Use Python for your implementation (you may use built-in libraries for matrix operations, e.g., numpy, and visualization of the result).
- Make sure that the TA can easily run the code by plugging in our test data.

Problem 1

(k-means, **50pts**) Generate 2 sets of 2-D Gaussian random data, each set containing 500 samples using parameters below.

$$\mu_1 = [1, 0],\ \mu_2 = [0, 1.5],\ \Sigma_1 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix} \tag{1}$$

1. **(20pts)** Write a function `cluster = mykmeans(X, k, c)` that clusters data $X \in \mathbb{R}^{n \times p}$ (n number of objects and p number of attributes) into k clusters. The c here is the initial centers, although this is usually not necessary, we will need it to test your function. Terminate the iteration when the ℓ_2-norm between a previous center and an updated center is ≤ 0.001 or the number of iteration reaches 10000.
2. **(15pts)** Apply your code to the data generated above with $k = 2$ and initial centers $c_1 = (10, 10)$ and $c_2 = (-10, -10)$. In your report, report the centers found for each cluster. How many iterations did it take? Show a scatter plot of the data and the centers of clusters found.
3. **(15pts)** Apply your code to the data generated above with $k = 4$ and initial centers $c_1 = (10, 10)$ and $c_2 = (-10, -10)$, $c_3 = (10, -10)$ and $c_4 = (-10, 10)$. In your report, report the centers found for each cluster. How many iterations did it take? Show a scatter plot of the data and the centers of clusters found.

Problem 2

(text clustering, **50pts**)
Dataset: https://www.kaggle.com/noushad24/amazon-reviews/download

1. **(15pts)** Let us focus on the reviews from the dataset without their labels. Build a weight matrix about all the words in the review part. Represent each review as a real-valued vector of tf-idf. Report each preprocessing step you applied and attach the corresponding part of your code. Visualize the matrix with color code (i.e., show the matrix as an 2D image where pixel intensity represents the weight).

2. **(15pts)** Pick your own 5 "positive" words and 5 "negative" words, which indicate if a product is good or bad, respectively. List the words you selected. Represent each review in a vector space of these ten words (i.e., count matrix) as well as tf-idf weight matrix.

3. **(20pts)** For each review, sum up the frequency of "positive" words and "negative" words. Represent each review as a vector of length 2. Now the reviews can be shown in 2D space, while one dimension is about "positive" and the other one is "negative". Apply your code from Problem 1 to this 2D data with $k = 2, 3, 4$ with randomly initialized centers. In your report, report the centers found for each cluster. How many iterations did it take? Show a scatter plot of the data and the centers of clusters found.