TCSS 588 Hands-on assignment #2, due Wednesday 5/10/2017 @ 4pm

Credit: The data and ideas behind these exercises and homeworks are from the NIH LINCS DCIC Crowdsourcing Portal and Ma'ayan Lab @ Mt Sinai, New York.
http://www.maayanlab.net/crowdsourcing/megatask2.php

The overarching goal is to identify gene clusters and subsequently annotate these clusters using gene sets.

This is a group assignment. You can work in a group consisting of 1, 2, or 3 members. Each group make one submission via canvas. Please state the names of all your team members in your submission. This assignment will be graded out of 10 points.

Upload 2 files for this assignment:

1. A document answering each of the following questions as a pdf or Word document.
2. One single R script named "hw2.R" containing all the code to answer all of the following questions. Please use "#" (comment lines) in your R script to indicate the question number and to document your code.
3. *(Bonus 2 points)* Create a Jupyter notebook (http://jupyter.org/) for Q1 and Q2. Submit the interactive notebook via canvas. If you pursue this option, you should answer all the questions from Q1 and Q2 in this notebook.

In your R script, you can assume the files "**gene_expression_n438x978.txt**" and "**gene_lists.txt**" are in your working directory. I will grade your script by running

```
R CMD BATCH hw2.R hw2.out
```

Please make sure your code will run using the above batch command.
I do not expect your script to take in any input files or ask any questions.

Note that Q2 may take quite some computational time to finish. Do not leave this until the last minute.

(Total 5 points) Clustering

1. Write R code to apply k-means clustering algorithm to identify gene clusters from "**gene_expression_n438x978.txt**". Points for discussion in your report
   a. What did you use as the input number of clusters? Why?
   b. What is your cluster size distribution? In other words, did you get any small or large clusters? Discuss the cluster size distribution in your report.
   c. How do you know if the resulting clusters are any good? Discuss.
   d. How do you visualize the resulting clusters? Discuss and show sample visualization images in the report.

(Total 5 points) Gene set enrichment

2. Using the following data “**gene_expression_n438x978.txt**” and “**gene_lists.txt**” to answer this question.

   In class, we discussed how to compute enrichment comparing a set of differentially expressed genes and predefined gene sets. Here, treat each of your gene clusters from Q1 as your gene set and compare each of your gene clusters to the predefined gene sets in “**gene_lists.txt**”.

   a. Write R code to compute this enrichment.
   b. Show one example gene cluster with a high enrichment (i.e. low p-value from fisher’s test) in your report.
   c. Show one example gene cluster with a low enrichment (i.e. high p-value from fisher’s test) in your report.
   d. Discuss any additional observations or challenges you encountered.