IE507, Modeling and Computation Lab
Take Home Test 1-November-2018

**Instructions:**
Solve the given problems and submit the code and report by November 4, 5.00 PM. Prepare different submission files for Part I and Part II of the take home exam.

You are free to consult your previous exercises, books, internet resources and discuss with your friends. However you should write all answers on your own. If you discuss with your friends, you need to clearly state their names in the report. Copied answers will lead to severe penalization for all students involved.

**Take home exam question:** [30 Marks]
Let $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ be an *Alphabet* and let its constituents $a_i$, $\forall i = 1, \ldots, N$ be called *symbols*. A *string* $s = s_1 s_2 \ldots s_\ell$ of length $\ell$ where $s_q \in \mathcal{A}, \forall q = 1, \ldots, \ell$, is constructed by concatenating $\ell$ symbols from the set $\mathcal{A}$. For two strings $s^j$ and $s^k$ of the same length $\ell$, define the $H$-distance as follows:

$$H(s^j, s^k) = \sum_{r=1}^{\ell} \mathbb{I}(s^j_r \neq s^k_r)$$

where $\mathbb{I}(p)$ is zero when proposition $p$ is false and 1 otherwise. In other words, the function $H(s^j, s^k)$ measures the number of positions in which the two strings $s^j$ and $s^k$ differ. Consider two strings $s^1 = ABBA$ and $s^2 = BBBA$. Then $H(s^1, s^2) = 1$. Similarly, if $s^3 = ABAB$ and $s^4 = BABA$, we have $H(s^3, s^4) = 4$.

We consider a set $\mathcal{S}$ of equal-length strings given by $\mathcal{S} = \{s^1, s^2, \ldots, s^M\}$, where the length of each string $s^j \in \mathcal{S}, \forall j = 1, \ldots, M$, is $\ell$.

The aim is to find a string $c = c_1 c_2 \ldots c_\ell$ such that the value $\sum_{m=1}^{M} H(s^m, c)$ is minimum. Note that the symbols $\{c_j\}_{j=1}^{\ell}$ of string $c$ are also from the alphabet $\mathcal{A}$.

1. Formulate a suitable optimization problem for the objective mentioned above. Make your model general so that it can be used for any value of $\ell$, $N$ and $M$.
2. Justify the suitability of your optimization model for the task in question with proper explanation.
3. For the attached files data1.txt and data2.txt we consider $\mathcal{A} = \{O, E, I, R\}$. Both the files contain strings of length 10. There are 4 strings in data1.txt and 15 strings in data2.txt. Use your formulation to find the string $c$ for the data in files data1.txt and data2.txt. Report the string $c$ obtained from your model and the CPU time taken to solve your problem.
4. The attached file data3.txt contains 6 strings of length 55. The alphabet is $\mathcal{A} = \{O, E, I, R\}$. Use your formulation to find the string $c$ for the data in file data3.txt. Report the string $c$ obtained from your model and the CPU time taken to solve your problem.
5. Consider the alphabet $\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, Q, R, S, T, V, W, Y\}$ for the file data4.txt. There are 23 strings of length 11 in data4.txt. Use your formulation to find the string $c$ for the data in file data4.txt. Report the string $c$ obtained from your model and the CPU time taken to solve your problem.
6. For a string $s$ and string set $\mathcal{S}$, define the cohesiveness as $Co(\mathcal{S}, s) = \max_{y \in \mathcal{S}} H(y, s) - \min_{y \in \mathcal{S}} H(y, s)$. Find the value of $Co(\mathcal{S}, c)$ for data1.txt, data2.txt, data3.txt and data4.txt.
7. Suppose if the aim was to minimize $Co(\mathcal{S}, s)$ over all possible $\ell$-length strings $s$, provide a discussion of how the solution to this problem would compare with that of the previous optimization problem. (You need not formulate the model for solving the new problem $\min_s Co(\mathcal{S}, s)$. A simple intuition will be sufficient.)