STAT 534 Homework 4
Due May 13, 2019
©Marina Meilă
mmp@stat.washington.edu

Reading:

Problem 1 – Data structures for disjoint sets
this problem is under development
a Write a python module that implements the functions UNION (for union-by-rank) and FIND-SET (for finding the representative with path compression), MAKE-SET and LINK.

In addition, write a FIND-PARENT function that returns the parent of an element *without modifying the data structure*. This function will be used in your homework to display the state of the disjoint set forest at any given time.

b Describe in your homework how you implemented this data structure; e.g. did you use an array? a dictionary?, etc. What did the entries represent? (1-2 paragraphs)

Describe in enough detail that we can evaluate if the functions operating on your data structure achieve the asymptotic running time of their pseudocode versions. *Small constant differences can be ignored.*

b Write a `__main__` function that ...

Next one.
c Apply your algorithm to `statisticiansA-M.txt`.

1. Let the elements be the truncated names as in Homework 1. First, assign each name a number from 0 to $n-1$, representing its rank in the data file,; $n = 415$ is the number of statisticians in the file. We will call this number the *Id* i of the statistician, not to be confused with rank of the node.

   You will use the numbers $i = 0 : n-1$ with the disjoint sets forest.

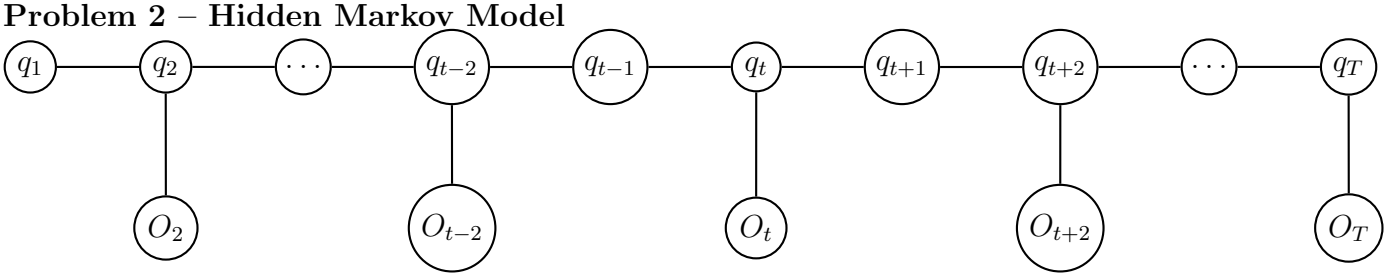
2. Write a function FINDSETSTATISTICIAN that takes an input the last name of a statistician and returns the truncated last name of the representative of the statistician in your data

structure.

3. Read the list of edges $x_{1:m}, y_{1:m}$ from file and perform the following operations.

```
union( x1, y1 )
union( x2, y2 )
...union( xm, ym )
findSet( 'Blackwell' )
findSet( 'Bottou' )
findSet( 'Brad' )
findSet( 'Breslow' )
findSet( 'Wellner' )
findSet( 'Laird' )
findSet( 'Fisher' )
findSet( 'Holmes' )
```

Problem 2 – Hidden Markov Model



In the above HMM, you only observe the outputs on even steps. Hence, the sequence of observations is $O_2, O_4, \ldots O_T$; T is always even. Denote $t : 2 : t'$ with $t' > t$ the sequence of even values in the set $t, t+1, \ldots t'$ (note that this does not agree with python conventions!).

a. Define $\alpha_t(i) = P[O_{1:2:t}, q_t = i]$ for *even* t. Derive the expression of $\alpha_2(i)$, and the expression of $\alpha_t(i)$ as a function of the values of $\alpha_{t-2}(j)$, $j = 1 : N$.

b. Define $\beta_t(i) = P[O_{t+1:2:T}|q_t = i]$ for *even* t. Derive the expression of $\beta_T(i)$, and the expression of $\beta_t(i)$ as a function of the values of $\beta_{t+2}(j)$, $j = 1 : N$.

c. Prove or disprove $P[O_{1:2:T}] = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ for any even t.

d. Define $\gamma_t(i) = P[q_t = i|O_{1:2:T}]$ for any $t = 1 : T$. Derive the expression of $\gamma_t(i)$ as a function

of the model parameters and α and β values.

e. Define $\xi_t(i,j) = P[q_t = i, q_{t+1} = j | O_{1:2:T}]$ for any $t = 1 : T-1$. Derive the expression of $\xi_t(i,j)$ as a function of the model parameters and α and β values.