STA 141C - Big Data & High Performance Statistical Computing **Spring 2023**

Homework 3

Lecturer: Bo Y.-C. Ning *Due May 26, 2023*

Due **May 26, 2023** by 11:59pm.

A few notes:

1. Submit your homework using the file name "**LastName_FirstName_hw3**"
2. Answer all questions with complete sentences.
3. Your code should be readable; writing a piece of code should be compared to writing a page of a book. Adopt the **one-statement-per-line** rule. Consider splitting a lengthy statement into multiple lines to improve readability. (You will lose one point for each line that does not follow the one-statement-per-line rule)
4. To help understand and maintain code, you should always add comments to explain your code. (homework with no comments will receive 0 points). For a very long comment, break it into multiple lines.
5. Submit your final work with one **.pdf** (or **.html**) file to Canvas. I encourage you to use LaTeX for writing equations. Handwriting is acceptable, you have to scan it and then combine it with the coding part into a single .pdf (or .html) file. Handwriting should be clean and readable.
6. For Jupyter Notebook users, put your answers in new cells after each exercise. You can make as many new cells as you like. Use code cells for code and Markdown cells for text.
7. This assignment will be graded based on how you implement your code .

In this homework, you will work with the Google PageRank problem. You will need to compare the computational speed between direct methods to iterative methods.

Questions:

1. Open the ucd-web folder from Piazza webpage. The folder contains two files U.txt and A.txt. U.txt lists the 500 URL names. A.txt is the 500×500 connectivity matrix. Read data into R or python. Once you read in the data, **take the transpose** of the dataset of A.txt to obtain the A matrix.

   Compute summary statistics:

   (a) number of pages
   (b) number of edges (page links)
   (c) number of dangling nodes
   (d) max in-degree
   (e) max out-degree
   (f) visualize sparsity pattern of A

2. Set the teleportation parameter at $p = 0.85$. Try the following methods to solve for x using the ucd-web data and report the speed of each method. You may want to remove some strange URLs (it depends on you how to remove them as long as it makes sense)

(a) Dense linear system solver: LU decomposition

(b) Dense linear system solver: QR factorization

(c) A simple iterative linear system solver such as Jacobi or Gauss-Seidel

(d) Choose either a dense eigen-solver such as SVD or iterative method such as the power method

(e) Comparing the computational speed for all the methods

(f) List the top 20 ranked URLs you found for each method and comment on your findings.

3. As of Monday, 13 Feb 2023, there are at least 4.61 billion indexed webpages on internet according to `http://www.worldwidewebsize.com/`. Comment on whether each of these methods may or may not work for the PageRank problem at this scale.