CS 5002: Discrete Structures

Final project

Assigned: Thursday, December 9th

Due: 12 noon on Friday, December 17th

Instructions: Please choose **only one** of the following problems. Solve the chosen problem, and then write up your solution. Your write-up should contain the following elements:

- Problem statement
- Mathematical description of the solution
- Observations and insights drawn from your approach
- Pseudocode, or executable, commented, and clear code

Note 1: You may choose to work on more than one problem, but you are expected to submit your report for one problem.

Note 2: You may choose to prepare your project as a stand-alone Jupyter notebook, but that is not required. Any well-written and meaningfully formatted project report is acceptable.

In preparing your project report, please consider:

- Writing mathematical descriptions using complete sentences, paragraphs, and formulas.
- Commenting your pseudocode/code as necessary, with a description of what each function does, and all major steps.
- Labeling plots axes, using legends, and appropriate plot titles.

Although you may discuss the project with others, you should turn in your own, original work.

The DBLP Publication Network

The following datasets, provided on Canvas, contain bipartite graphs where one set of nodes (vertices) are authors and the other set are academic papers. Each edge (a, p) connects an author a to a paper p.

- out.dblp_author-short.txt
- out.dblp_author-medium.txt
- out.dblp_author-long.txt
- out.dblp_author-all.txt

The short dataset contains 1,000 edges, the medium 10,000 edges, the long version 100,000 edges, and the "all" version has all the edges.

Note: The network is encoded with two numbers per line separated by spaces. You will need to open and read the file, and then turn the read data into a meaningful format. If you use Python, you should be able to use built-in method `split` to get a list of numbers. The numbers at even indices (starting at 0) are authors, and the numbers at odd indices are publications. You will want to be mindful not to do things that take a lot of operations or memory.

Your tasks:

a) Find the minimum, maximum, and average number authors per paper.

b) Find the minimum, maximum, and average number of papers per author.

c) Find the (not necessarily unique) author who has written the most papers. Call this author X. An author other than X has an X-index of 1 if she has co-authored at least one paper with X. An author has an X-index of 2 if she does not have an X-index of 1, but has co-authored a paper with someone who has an X-index of 1. Similarly, you can define having an X-index of 3, 4, etc.

Write a function that produces the set of authors for some given index, n.

Graphs and Linear Algebra

Recall that the adjacency matrix of an undirected graph has entry $A_{i,j} = 1$ if and only if nodes i and j are adjacent. Also, recall that a graph is **regular** with degree k if every node has k neighbors. We call such a graph k-regular. Finally, for shorthand, we say that the eigenvalues of a graph are the eigenvalues of its adjacency matrix.

Your tasks:

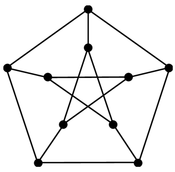
a) Find three graphs with more than five nodes that are 2-regular, 3-regular, and 4-regular. Draw the chosen graphs, and find their adjacency matrices. These will be running examples for this problem.

b) Find the eigenvalues of the three examples, along with the multiplicities of the eigenvalues.

c) Show that if G is k-regular, then k is an eigenvalue of G.

d) Show that G is k-regular and connected, then the eigenvalue k of G has multiplicity one.

e) Show that the following graph, called the Petersen Graph, is 3-regular by finding its eigenvalues.



Simple Text Processing

When performing data analyses involving raw text data, we usually start by applying some simple pre-processing steps to clean our data. In this problem, we are going to examine some of those simple processing steps.

Here is a sample text from Cory Doctorow's recent novel, **Unauthorized Bread**:

"Long before she got to that point, she'd grown certain that it was a lost cause. But these were the steps that you took when the electronics stopped working, so you could call the 800 number and say, "I've turned it off and on, I've unplugged it, I've reset it to factory defaults and..."

There was a touchscreen option on the toaster to call support, but that wasn't working, so she used the fridge to look up the number and call it. It rang seventeen times and disconnected. She heaved a sigh. Another one bites the dust.

The toaster wasn't the first appliance to go (that honor went to the dishwasher, which stopped being able to validate third-party dishes the week before when Disher went under), but it was the last straw. She could wash dishes in the sink but how the hell was she supposed to make toast—over a candle?

Just to be sure, she asked the fridge for headlines about Boulangism, and there it was, their cloud had burst in the night. Socials crawling with people furious about their daily bread. She prodded a headline and learned that Boulangism had been a ghost ship for at least six months because that's how long security researchers had been contacting the company to tell it that all its user data—passwords, log-ins, ordering and billing details—had been hanging out there on the public internet with no password or encryption. There were ransom notes in the database, records inserted by hackers demanding cryptocurrency payouts in exchange for keeping the dirty secret of Boulangism's shitty data handling. No one had even seen them.

Boulangism's share price had declined by 98 percent over the past year. There might not even be a Boulangism anymore. When Salima had pictured Boulangism, she'd imagined the French bakery that was on the toaster's idle-screen, dusted with flour, woodblock tables with serried ranks of crusty loaves. She'd pictured a rickety staircase leading up from the bakery to a suite of cramped offices overlooking a cobbled road. She'd pictured gas lamps."

Your tasks:

Using the provided sample of the **Unauthorized Bread** novel as the original input text:

a) Find the total number of alphanumeric characters in the original input text.

b) Find the total number of unique numeric characters in the original input text.

c) Convert all text to lowercase.

d) Find the total number of words in the provided text.

e) Find the total number of unique words in the provided text.

f) Find the most frequent noun and the most frequent verb in the provided text.

g) A **digram** is defined as a group of two successive letters or other symbols in some text. Find the most frequent digram in the provided text.

Analyzing Credit Card Clients Data Set

Acknowledgement: In this problem, we are using a data set available from Kaggle: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset (The dataset also available on Canvas.)

More specifically, we will use the 'Default of Credit Card Clients Dataset'. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

The dataset contains the following 25 variables:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

Note: You will need to open and read the provided file, and then turn the read data into a meaningful format. If you use Python, the following line of code might be helpful:

credit = pd.read_csv("UCI_Credit_Card.csv")

Your Tasks:

Analyze the provided dataset, and answer the following questions:

a) How many data entries (rows) are present in the provided dataset?

b) How many credit card users are at least 70 years old?

c) How many credit card users are younger than 25 years?

d) What percentage of users do not have a default payment? (1=yes, 0=no)

e) Among all female users, what percentage of do not have a default payment?

f) Who is more likely do default - users younger than 25 or older than 70 years?

The Problems

1. The DBLP Publication Network
2. Graphs and Linear Algebra
3. Simple Text Processing
4. Analyzing Credit Card Clients Data Set