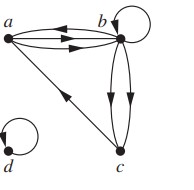
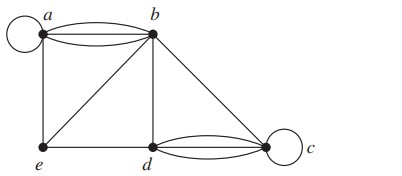
**Assignment 02**

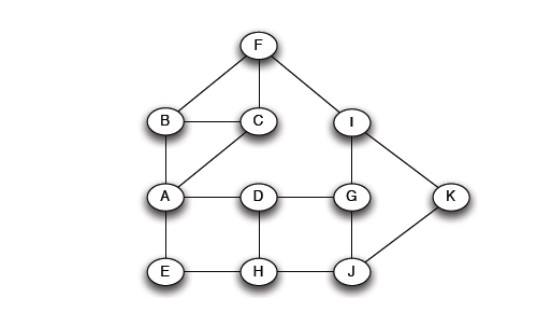
1. 

Consider the above two graphs and answer the following questions”

i) Calculate in-degree

ii) Calculate out-degree

iii) Represent them using adjacency matrix, adjacency list.

1. 

Consider the above graph where the start node is A and the destination node is K having uniform weights. Now answer the following questions:

i) Measure Degree Centrality

ii) Measure Closeness Centrality

iii) Find the shortest path using BFS algorithm

1. For this task, you will work with a dataset of network graphs and perform various analyses using Python. You can use any network dataset of your choice for this resource (Link), such as social networks, communication networks, or transportation networks. Your tasks are as follows:

i) Load the network data into Python and visualize it using appropriate graph layouts and visualization techniques. Explain the structure of the network and any interesting patterns or features that you observe. (Not more than 150 words)

ii) Analyze the degree distribution of the network and determine whether it follows a power law distribution. If so, estimate the parameters of the distribution and plot the fitted line on a log-log scale. Explain the implications of a power law distribution for the network's structure and dynamics.(not more than 150 words)

iii) Compute the clustering coefficient of the network and compare it to random and regular networks with the same number of nodes and edges. Explain the significance of the clustering coefficient for the network's resilience, efficiency, and vulnerability to attacks or failures.(Not more than 200 words )

Your submission should include a Jupyter notebook or Python script that demonstrates your implementation and results, as well as a written report that explains your methodology and findings. With all the visualization images of the results . You should also provide clear and concise code comments and documentation to facilitate understanding and reproducibility.

1. Term-frequency matrix for the six documents (D1 to D6) is shown below. Answer the following questions:

i) What is the TF-IDF value for (D3, Lay off)

ii) Find the cosine similarity between documents? Identify the two documents that are most similar.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Documents | Corona | Inflation | Trump | Lay off | Soccer |
| D1 | 19 | 12 | 4 | 6 | 0 |
| D2 | 0 | 2 | 0 | 0 | 14 |
| D3 | 8 | 0 | 3 | 12 | 0 |
| D4 | 0 | 5 | 10 | 0 | 1 |
| D5 | 0 | 0 | 0 | 0 | 4 |
| D6 | 4 | 19 | 2 | 6 | 0 |

iii) Often TF-IDF is used as multidimensional data in classifying document collections.Can you think of a weakness in the approach and how you may overcome it?

1. In this assignment question, you will work with a dataset of text documents and perform various text-mining tasks using the nltk library in Python. You can use any text dataset of your choice, such as a collection of news articles, social media posts, or scientific papers(refer to online sources). Your tasks are as follows:

i) Load the text data into Python and preprocess it by tokenizing the text into individual words, removing stop words, and converting all the words to lowercase.

ii) Implement stemming and lemmatization techniques to further process the text data. Compare and contrast the differences between the two approaches and explain which approach you would choose in different situations. (explanation not more than 250 words )

iii) Use the nltk library to calculate various statistics on the text data, such as the frequency distribution of words, the most common words, and the average word length. Visualize the results using appropriate charts and graphs.

iv) Apply topic modeling techniques to the text data, such as Latent Dirichlet Allocation (LDA), to identify the underlying themes or topics in the dataset. Explain how you would interpret the results and what insights you can gain from them.(Bonus Question)

Discuss the limitations and potential biases of your approach, and suggest possible improvements or future directions(not more than 300 words ).

Your submission should include a Jupyter notebook or Python script that demonstrates your implementation and results, as well as a written report that explains your methodology and findings, With all the visualization images of the results. You should also provide clear and concise code comments and documentation to facilitate understanding and reproducibility.

1. Using the friendship network in this class to study the following questions:

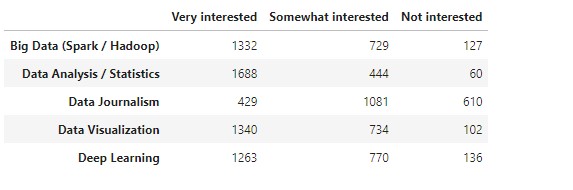
<https://docs.google.com/spreadsheets/d/16zRiKesRjmdvtKp7xoqCjPQW97WpS8zroGvsnIyOtRQ/edit?usp=sharing>

1. Which students have the maximum in-degree and out-degree?
2. Measure Degree Centrality
3. Measure Closeness Centrality

7. The participants had three options for each topic: Very Interested, Somewhat interested, and Not interested. 2,233 respondents completed the survey. The survey results have been saved in a csv file and can be accessed through this link: <https://cocl.us/datascience_survey_data>.

If you examine the csv file, you will find that the first column represents the data science topics and the first row represents the choices for each topic.

1. Write a python script that will generate the data frame as below:



2. Use the artist layer of Matplotlib to replicate the bar chart below to visualize the percentage of the respondents' interest in the different data science topics surveyed.

Hints: Sort the dataframe in descending order of Very interested.  
 Convert the numbers into percentages of the total number of respondents. Recall that 2,233 respondents completed the survey. Round percentages to 2 decimal places.