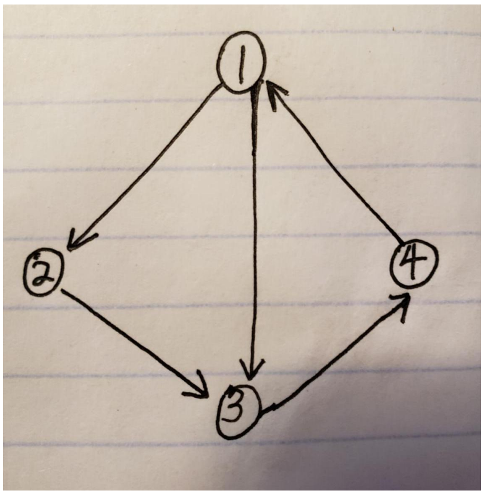
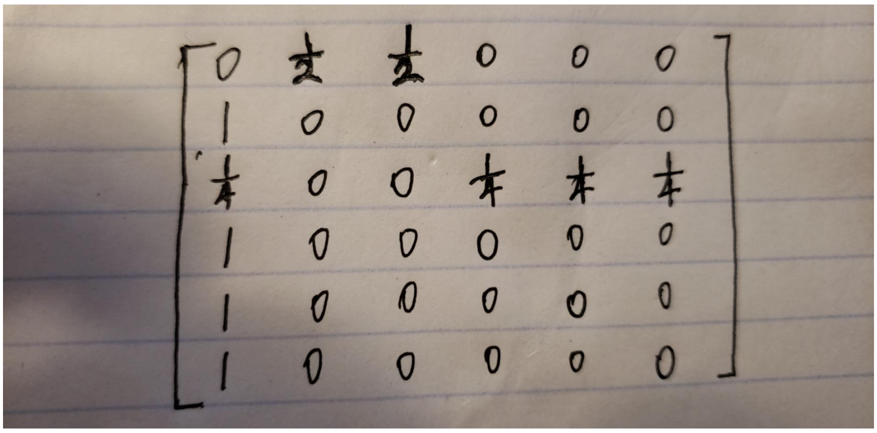
1. Draw a network consistent with the following adjacency matrix.



2. Walk through the steps of the PageRank algorithm, explaining how to compute the PageRank of all pages in the following directed network. Set *d* to 0.15. You do not need to run the algorithm until convergence. You do not need to compute the precise numerical values for each intermediate step. Instead, you should show and explain what computations are performed at each step, in enough detail that someone without knowledge of the algorithm could reimplement it using your walkthrough.

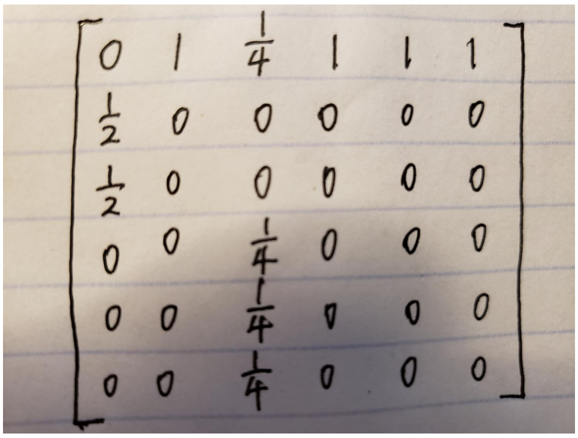
Answer:

Step 1, convert this graph into a adjacency matrix (A matrix that a value[i,j] = 1 means node i is directed to j).



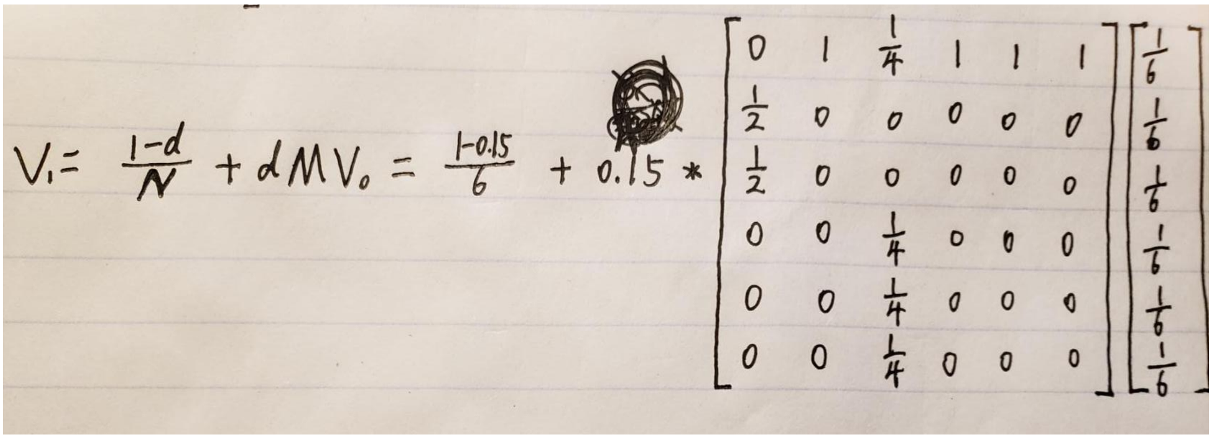
$$\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Step 2, get the stochastic matrix from the matrix above.



$$\begin{bmatrix} 0 & 1 & \frac{1}{4} & 1 & 1 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \end{bmatrix}$$

Step 3, assuming that the probability of each page is equal = 1/6, calculate the probability distribution of the initial test by multiply it with a probability vector.



$$V_1 = \frac{1-d}{N} + dMV_0 = \frac{1-0.15}{6} + 0.15 * \begin{bmatrix} 0 & 1 & \frac{1}{4} & 1 & 1 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \end{bmatrix}$$

Step 4, Repeat step3 iteratively.

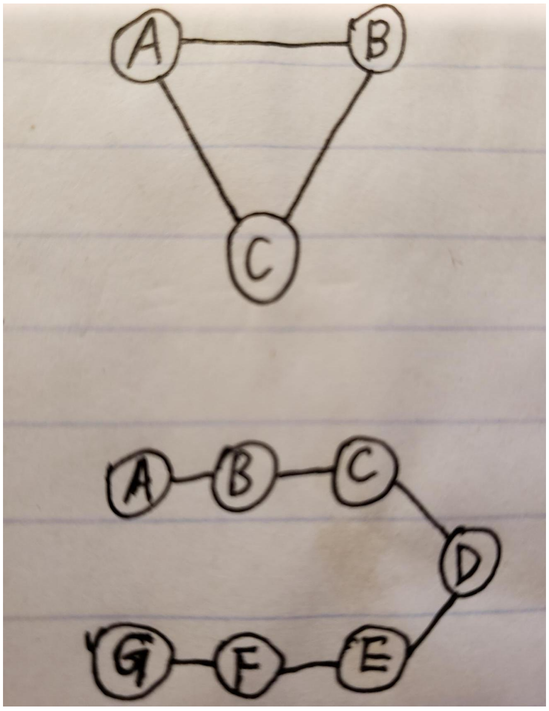
3. In at most two sentences, describe the essential differences between the PageRank algorithm and another measure of node centrality of your choosing.

Answer： I compared PageRank and Degree Centrality.

The difference is that PageRank considers weight while Degree Centrality dosen't consider weight.

4. In one sentence, define a network's characteristic path length. Draw a network with a characteristic path length of 1. Draw a second network with a characteristic path length that is greater than 2.

Answer: Average Path Length is the average number of steps along the shortest paths for all possible pairs of network nodes.



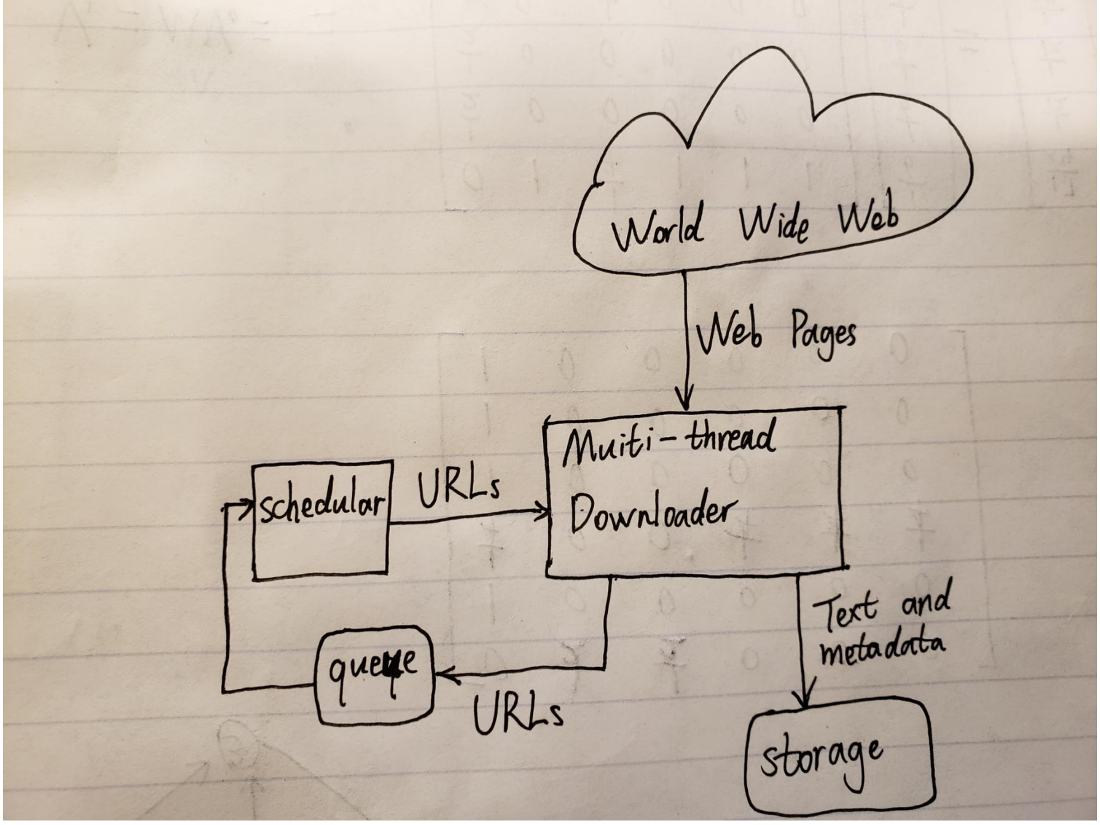
5. What is the connection between Tim Berners-Lee, CERN, and the web?
Answer: Tim Berners-Lee is a British Scientist who invented the world wide web(WWW) in 1989, while working at CERN

6. Give 3 HTTP status codes and what it means when a server returns that response.
Answer:
200: Standard response for successful HTTP requests.
202: The request has been accepted for processing, but the processing has not been completed.
404: The requested resource could not be found but may be available in the future.

7. What is a regular expression? Give an example of a regular expression in Python and a string that is matched and unmatched by it. Your regular expression should make use of at least 3 of the following symbols: . ^ $ * + ? {m} \ | (...) \w \W \S \d \D
Answer: Regular expression is a sequence of characters that define a search pattern.

> **Regular Expression:** "^\d{1,3} dogs$"
> **Matched:** "100 dogs"
> **Unmatched:** "dogs"

8. Draw a diagram OR write a paragraph describing the basic architecture of a web crawler.
Answer



9. Why is URL deduplication useful when crawling the web? Give an example of two URLs that are difffferent but probably should be considered the same.
Answer: URL deduplication is useful because sometime two different URL may may have the same content.

For example: https://en.wikipedia.org/wiki/Water

VS https://en.wikipedia.org/wiki/Water?foo=bar

10. In one sentence, define a web crawler's "revisit policy". In a second sentence, describe the distinction between "age" and "freshness" of a page.
Answer: revisit policy is used to check whether page has changed.

Age records how long the page has been updated while freshness only judge if the local data has been synchronized or modified.

11. At what level of linguistic analysis is part-of-speech (POS) tagging? Explain your answer in 1 sentence.
Answer: POS analysis at **word** level.

12. TF-IDF is a metric of how important a word is to a document in a corpus of many documents. Why is IDF necessary (i.e., why isn't TF enough?)?
Answer: Because there are some very common words such as "the" that will incorrectly emphasize documents which use "the" more frequently. The IDF will diminish the weight of those common terms and increase the weight of terms that occur rarely. This will increase the accuracy of our algorithm.

13. What sort of data is used to train vector space models of words such as word2vec?
Answer: Word2vec is a group of related models that are used to produce word embeddings. It used document as training data.

14. Write a robots.txt fle requesting that crawlers (1) wait 2 seconds between visits and (2) avoid visiting the /secrets directory.
Answer:

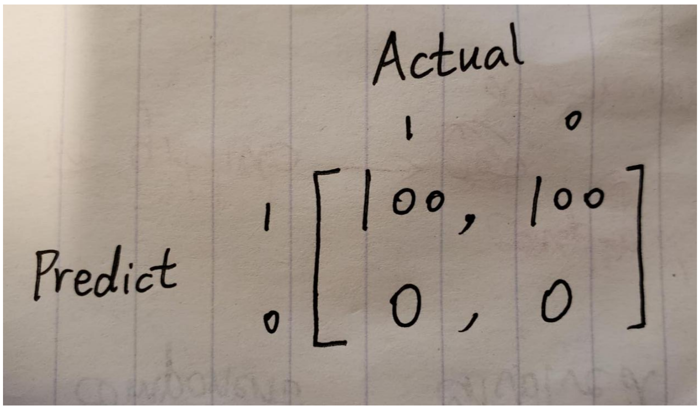
```
crawl-delay: 2
Disable: /secrets
```

15. In 2 or 3 sentences, how might you approach creating a sentiment analysis algorithm that, rather than using an exhaustive hand-tagged word list, makes use of vector representations of a few words known to be highly positive and negative in polarity.
Answer： For a given document, I first calculate TF(term frequency) of each highly positive and negative words. Then, the sentiment score of this document is

$$Sentiment\_Score = \sum_{i \in positive} tf_i - \sum_{i \in negative} tf_i$$

If the sentiment score is a positive number, then classify this document as positive. Otherwise, negative.

16. Show a 2 × 2 matrix of classification results for a sentiment analysis algorithm that sees 100 positive sentences, 100 negative sentences, and achieves high *recall* for positive-sentiment sentences despite low *accuracy*.



Actual

$$\text{Predict}\quad\begin{array}{c|cc} & 1 & 0 \\ \hline 1 & 100, & 100 \\ 0 & 0, & 0 \end{array}$$

17. What is the precise relationship between topics and words in Latent Dirichlet Allocation?
Answer： In Latent Dirichlet Allocation, each topic represents a set of words.

18. When are browser automation tools such as Selenium most useful?
Answer: It is most useful when we wish to spend less time testing the front end of their web applications but still want to be confident that every feature works fine. Selenium will save you time by automating repetitive online tasks with Selenium WebDriver. You will find a step-by-step example for automating and testing the login function of WordPress, but you can also adapt the example for any other login form.

19. What does it mean for a network to be a “small world”? Explain it using web pages as an example.

Answer: A small-world network is a type of mathematical graph in which most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a small number of hops or steps. Specifically, a small-world network is defined to be a network where the typical distance L between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes N in the network. While the clustering coefficient is not small. In the context of a social network, this results in the small world phenomenon of strangers being linked by a short chain of acquaintances. Many empirical graphs show the small-world effect, including social networks, wikis such as Wikipedia, gene networks, and even the underlying architecture of the Internet. It is the inspiration for many network-on-chip architectures in contemporary computer hardware.

Example: many pages inside the stevens website.

20. (automatic +1 for answering) Is there anything that you are particularly interested in learning about web mining that you are hoping we'll cover in the coming weeks?

Answer: I am pretty interested in whether there is a proper way to apply NLP to java code or Python code to analyze relationships between different code file.

21. On a scale from 1 (way too easy) to 4 (just right) to 7 (way too hard), how was this exam for you?

Answer: 4