WEB MINING BIA 660

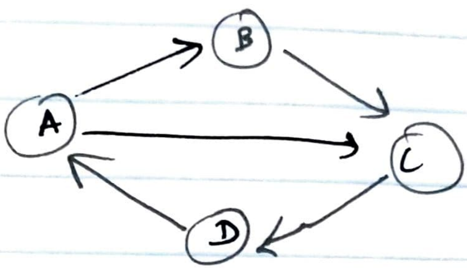
MID TERM

Name :- Hariom Mehta

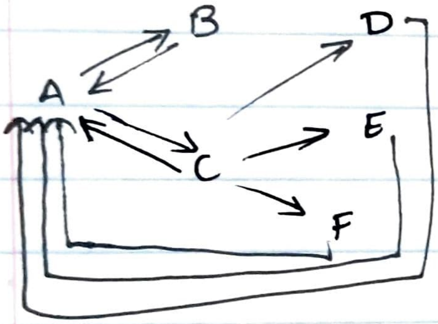
CWID :- 10453223

Q.1

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 1 & 1 & 0 \\ B & 0 & 0 & 1 & 0 \\ C & 0 & 0 & 0 & 1 \\ D & 1 & 0 & 0 & 0 \end{array}$$



Q.2



Formula :-

$$PR(A) = (1-d) + d\left(\frac{PR(T_i)}{C(T_i)} + \dots + \frac{PR(T_n)}{C(T_n)}\right)$$

PR(A) = Page Rank of A

PR(T_i) = is page Rank of T_i page which is linked to A. C_{T_i} = No. of Links

d = demping factor = 0.15 (given)

Initially we assume that pageRang for all page is equal = 1.

$$PR(A) = (1-0.15) + 0.15\left(\frac{1}{1} + \frac{1}{4} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1}\right)$$

$$= 0.85 + 0.15(4.25)$$

$$= 1.4875$$

→ $PR(B) = (1 - 0.15) + 0.15\left(\frac{1.4875}{2}\right)$

$= 0.85 + 0.15(0.74375)$

$= 0.85 + 0.115$

$\boxed{PR(B) = 0.9615}$

→ $PR(C) = 0.85 + 0.15\left(\frac{1.4875}{2}\right)$

$\boxed{PR(C) = 0.9615}$

→ $PR(D) = (1 - 0.15) + 0.15(0.9615/4)$

$= 0.85 + 0.15(0.24037)$

$\boxed{PR(D) = 0.886}$

→ $PR(E) = (1 - 0.15) + 0.15(0.9615/4)$

$= 0.85 + 0.15(0.2403)$

$\boxed{PR(E) = 0.886}$

→ $PR(F) = 0.85 + 0.15(0.9615/4)$

$\boxed{PR(F) = 0.886}$

PR(A) is maximum.

Q.3.

The main differenie between Page Rank Algorithm and Certrality Measures (degree centrality) is pagerank Algo. works on number of pages and quality of pages links of a perticular Page where as degree centrality focuses on nodes (page) with highest degree it has.

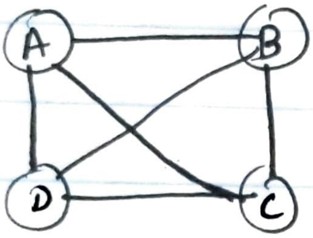
Q.4. Network's characteristic path length :- It's average number of edges in the shortest path between all vertex pairs.

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)}$$

a = path length

d(s,t) = distance between s & t

n = total number of nodes in graph G



(G1)

Let's calculate a for given graph G1

node - 4

edge - 6

a = for node (A) + for node (B) + for node (C) + for node (D)

$$\therefore a = \frac{1}{4(4-1)}[d(A,B) + d(A,C) + d(A,D)] + \frac{1}{4(4-1)}[d(D,A) + d(D,B) + d(D,C)]$$

$$\frac{1}{4(4-1)}[d(B,A) + d(B,C) + d(B,d)]$$

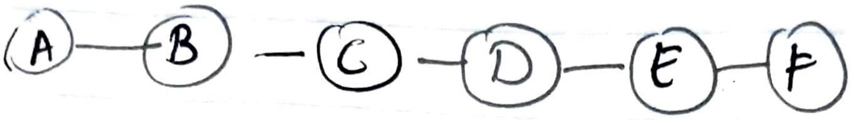
$$+ \frac{1}{4(4-1)}[d(C,A) + d(C,B) + d(C,D)]$$

$$\therefore a = \frac{1}{4\cancel{(3)}}[\cancel{3}] + \frac{1}{4\cancel{(3)}}[\cancel{3}] + \frac{1}{4\cancel{(3)}}[\cancel{3}] + \frac{1}{4\cancel{(3)}}[\cancel{3}]$$

$$\therefore \boxed{a = 1}$$

node 6
edge = 5

Network Q2 :-



$$a = 2 \times \frac{1}{6(5)} \underset{\text{for}(A,F)}{[1+2+3+4+5]} + \frac{1}{6(5)} [1+1+2+3+4] \times 2 \text{ (for B,E}$$

$$+ \frac{1}{6(5)} [2+1+1+2+3] \times 2 \text{ (for C,D)}$$

$$= \frac{1}{3(\cancel{5})} [\cancel{15}] + \frac{1}{3(5)} [11] + \frac{1}{3(5)} [9]$$

$$= 1 + \frac{11}{15} + \frac{9}{15}$$

$$= 1 + \frac{20}{15}$$

$$= 1 + \frac{4}{3}$$

$$a = 3.33 > 2$$

Q.5 Tim Berners-Lee was british scientist who developed World wide web (www) while working at CERN.

The main idea of www was to merge the new technologies of Computer into one global information system.

Q.6

HTTP code 200 - shows [OK] success. That HTTP request is accepted by client.

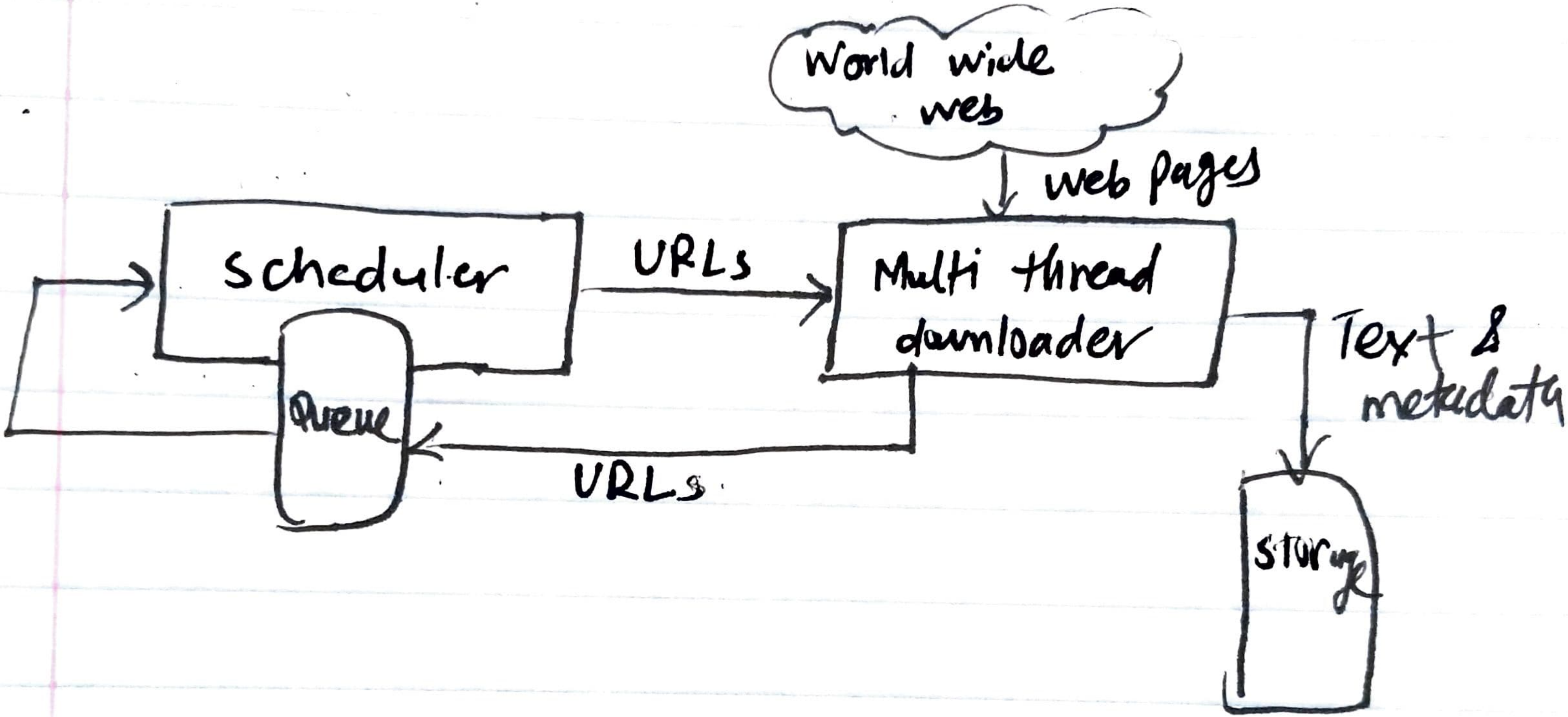
HTTP code - 404 :- Not Found. Server could not find the request.

HTTP code 401 - unauthorized request made

Q. 7. Regular Expression is sequence of characters that forms a search pattern.

```
import re
Pattern = 'r"\bS\w+"
test = "This is Super car"
result = re.match(pattern, test)
if result:
    print("successful")
Else:
Print("Unsuccessful")
```

Q.8



Architecture of a web crawler.

Q.9.

while crawling web, we come accross some websites which has little bit different name but it'll get us to same website. With the Help of dedullication it can compaire the response of websites and will keep origtnal or latest .version of it.

Example: (i) www.google.co.in

(ii) www.google.com.

Q.10 Revising policy :- It tells crawler's How often to check for changes to a Page.

Age tells about how old page is and freshness is when was the last time information was updated to that page.

Q.11 POS Tagging is performed when we want to Extract the words (with detecting the contex) from the Speech (Sentenses) in linguistic analysis. It identify words as nouns, verbes, adjectives etc.

Q.12 TF - IDF :- TF means term frequency. Why IDF (Inverse document frequency) is important ?

For Example if in a text document 'the' is very common and it's frequency will be high according to TF. But IDF removes the weight of frequent words and gives the weight to other words like 'Black', 'Bird' which is actually important.

TF - IDF = TF(word) * IDF(word)

so, if there is stop word 'the' it's IDF value will be '0'. so TF - IDF = 0.

Q.13 word2vec is 2-layer neural net. It requires a format of 'List of list' for training. Each document contained list and every list has tokens of document.

Q14

1) wait 2 sec between visits

User-agent :- bot

Crawl-delay :- 2

2) Avoid visiting /secrets directory

User-agent :- bot

Disallow :- /secrets/

Q.15 By using SVM classifiation (support vector machine classification) to Prepare the data. we can use tokenization and streamming svm classifiers and categories the data as positive, neutral or negative and corresponding polarity can be calculated.

Q.16

Classification	Positive	Negative	Recall(+)	Accuracy
Positive	60	80	$\frac{60}{60+20} = 75\%$	$\frac{60+20}{200} = 40\%$
Negative	40	20	$\frac{40}{40+20} = 66.67\%$	$\frac{40+20}{200} = 30\%$

Q.17 In LDA, from lots of document it builds topic per doucument and words per topic from the given set of document

→ Topic found by the n number of docs by Running LDA. (Supported by relevent words

→ Words are Extreacted to define a topic from given document sets.

Q.18 Selenium is very useful while performing test for webApplication, iOS Apps or Android Apps & Software testing.

Suppose, we need perticular type of data from given websit selenium can be useful to Extract data.

Q.19. By saying network 'A small world' means adding very small number of edges randomly. the diamoter tends to drop. This is known as the small world.

For Example :- In social media if any person turnout to connect other 6 person (connection) is small world network.

Q.20. I would like to learn How to mine crypto currency ? How web cryptominer works ?

Q.21 $\boxed{5}$