WEB MINING MIDTERM

BIA 660, Spring 2020

Stevens Institute of Technology

Instructions: Concisely answer each of the following questions. You may use notes, website, and textbooks when completing this exam. However, you may NOT consult with other students. The work you submit must be your own. Submit your completed exam as a PDF. You may type out your answers, write them by hand and scan them, or whatever else works for you. However, you must upload a PDF.

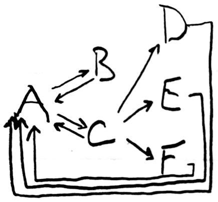
1.

Draw a network consistent with the following adjacency matrix.

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

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.



3.

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

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.

5.

What is the connection between Tim Berners-Lee, CERN, and the web?

6.

Give 3 HTTP status codes and what it means when a server returns that response.

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
```

8.

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

9.

Why is URL deduplication useful when crawling the web? Give an example of two URLs that are different but probably should be considered the same.

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.

11.

At what level of linguistic analysis is part-of-speech (POS) tagging? Explain your answer in 1 sentence.

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?)?

13.

What sort of data is used to train vector space models of words such as word2vec?

14.

Write a robots.txt file requesting that crawlers (1) wait 2 seconds between visits and (2) avoid visiting the `/secrets` directory.

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.

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*.

17.

What is the precise relationship between topics and words in Latent Dirichlet Allocation?

18.

When are browser automation tools such as Selenium most useful?

19.

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

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?

21. (NOT GRADED)

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

Last updated on March 25, 2020