CS494 - IR EXAM 2

FALL 2021

Name:

University NetID:

This test consists of 4 questions. The number of points for each question is shown below.

- Read all questions carefully before starting to answer them.
- Write all your answers in the space provided in the exam paper.
- The order of the questions is arbitrary, so the difficulty may vary from question to question. Do not get stuck by insisting on doing them in order.
- Show your work. Correct answers without justification will not receive full credit. However, also be concise. Excessively verbose answers may be penalized.
- Clearly state any simplifying assumptions you may make when answering a question.
- **Be sure to write your name on the test paper.**

Question	1	2	3	4	total
Points	20	20	20	20	80
Your Points					

Exercise 1 - 20 points. (Naïve Bayes)

You are given a collection of 800 documents, which are classified into one of the two classes: *entertainment* and *science*. The vocabulary of words in the collection is as follows:

$$V = \{actor, monkey, scientist, arm, mind, computer, meteor\}$$

Assume there are 300 documents from *entertainment* and 500 documents from *science*. Assume further that the frequency counts of words in each class are as follows:

$$entertainment : actor(200), monkey(150), scientist(50), arm(300), mind(42), computer(40), meteor(72)$$

$$science : actor(20), monkey(483), scientist(400), arm(40), mind(230), computer(421), meteor(53)$$

Train a Multinomial Naïve Bayes model on the above dataset and assign probabilistic labels to the unlabeled documents below (round probabilities to two decimals):

Scientists Train Monkeys to Move Two Virtual Arms With Their Minds
The main character who played in "A Beautiful Mind" was a scientist, a real scientist.

Ignore any words that are not in the vocabulary (assume stemming). Do add-1 smoothing.

Exercise 2 - 20 points. (Web Crawling)

Consider the following web graph:

Page A points to pages F, B, and G.
Page B points to pages C, E, D, and G.
Page E points to page F.
Page F points to pages B and E.
Page G points to page E.

Show the order in which the pages are indexed when starting at page A and using a breadth-first spider (with duplicate page detection). Assume links on a page are examined in the orders given above. Assume also that the robots.txt file at the domain of webpage B includes the following lines:

```
User-agent: *
Disallow: /
```

Exercise 3 - 20 points. (HITS)
Consider the web graph from Exercise 2, shown below for convenience.

Page A points to pages F, B, and G.
Page B points to pages C, E, D, and G.
Page E points to page F.
Page F points to pages B and E.
Page G points to page E.

Run the HITS (Hubs and Authorities) algorithm on this graph of web pages. Simulate the algorithm for two iterations.

Exercise 4 - 20 points. (Page Rank)
Consider the following pages and the set of web pages that they link to:

```
Page A points to pages B, C, D.
Page B points to page D.
Page C points to pages B.
Page D points to page C.
```

Consider running the PageRank algorithm on this graph of pages. Assume $\epsilon = 0.15$. Simulate the algorithm for two iterations. Show the page rank scores for each page for each iteration. Order the elements in the vectors in the sequence: A, B, C, D.

Remember:

$$R(A) = \frac{\epsilon}{n} + (1 - \epsilon) \sum_{(B,A) \in G} \frac{R(B)}{out(B)}$$