Your name:

B565 HW3 (Fall 2022)

Submission instructions

Submit a PDF file to canvas. The PDF file includes your answers to the calculation problems, results of your exploratory analysis of the movie ratings dataset, and results from your simulation. In addition, submit your code/notebook for Q3 and Q4 to github.iu under HW3 folder in your B565 repository.

Questions

1. (20 points) Working with strings/texts.

   (a) Given two strings $s1$ = ATCGTACGTGTA, and $s2$ = TCGTACGTGTAA, what's their Hamming distance? (3 pts)

   (b) Now represent $s1$ and $s2$ as vectors of 2-shingles (shingles of 2 letters), and compute their Jaccard similarity. (7 pts)

   (c) Based on your calculations above, which of the two metrics (Hamming distance or Jaccard similarity) do you think better captures the similarity/dissimilarity between these two strings? (3 pts)

   (d) Given two strings of average lengths of n letters represented as vectors of 2-shingles, what's the time complexity of computing the Jaccard similarity between the vectors? Use big O notation. You don't need to provide formal proof, but brief explanations are required. (7 pts)

2. (20 points total) Given a shingle(word)-document matrix as below,

$Shingle\_ID$	$d1$	$d2$	$d3$	$d4$	$d5$	$d6$
0	0	0	0	1	0	0
1	0	0	1	0	0	0
2	1	1	1	0	0	0
3	1	1	0	1	1	1
4	1	0	0	1	1	1
5	0	1	0	0	1	1

   (a) Compute the Jaccard similarity for all pairs of the six documents (5 points).

   (b) Compute the minhash signatures for each column (document) if the following hash functions are used: $h_1(x) = (2x+1)\%6$; $h_2(x) = (3x+2)\%6$; $h_3(x) = (5x+2)\%6$; and $h_4(x) = (7x+3)\%6$ (8 points).

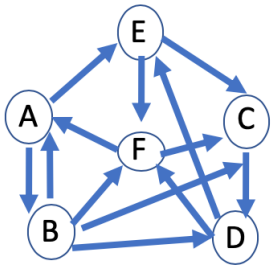
   (c) Compute the similarities for all pairs of the six documents using the signatures (5 points).

   (d) Does minhash provide good signatures for computing the document similarity for this example (2 points)?

3. (10 points) Credit card fraud detection using KNN. Please use this code as the start code.

   - Before calling KNN for classification, were there data processing steps applied in the start code? What distance metric was used in KNN in the start code? (5 points)
   - Using "copy and edit" function on Kaggle to create your only KNN classifier. Try different settings, including different k values and different distance metrics and report how the classifier's performance changes. (5 points)

4. (25 points total) Link analysis



   (a) Given the above toy web (with 6 web pages, A, B, $\cdots$, F), derive its transition probability matrix (5 points).

   (b) Assume a surfer is on web page A, what's the probability that the person will be next visiting B and then D (5 points)?

   (c) Implement the PageRank algorithm that uses power iteration. Your program takes a matrix of web links as the input, and computes the ranks of the web pages. Test your program using the matrix your derived in (a). Try different initial distributions and see if the result changes or not (15 points).

5. Write a summary for this review article about Precision Nutrition (PN): Precision nutrition: Maintaining scientific integrity while realizing market potential. [25 pts]

   - The length of your summary is about one page.
   - Include a paragraph/section summarizing the data mining tasks involved in PN.
   - Write the summary in your own words; don't copy and paste.