Assignment 3
CIS530 Advanced Data Mining **Spring 2023**

Due: 3/19/2022 23:59PM Submit to myCourses

Submission note:

1. *Include a cover page with group ID and member names.*
2. *Please provide all derivation process. Providing only the result (e.g., a number) will get 0 pts!*
3. *Please submit solutions to Q1-Q3 in a single PDF file.*
4. *Please submit the source code and results of Q4.*

Q1. (14 pts) True or False

a. In binary classification, the higher the classification accuracy is, the lower the classification cost will be.

b. Recall has the same meaning as false positive rate.

c. The precision-recall curve of a random machine learning model will output a straight line at precision = 0.5.

d. Small training data may lead to lower bias and higher variance.

e. Hamming distance and L1 distance have identical meaning on binary vectors.

f. A larger k in k-shingles will always perform better in document representation.

g. In the practice of recommendation system, user-user methods usually perform better than item-item methods.

Q2. (10 pts) Please compute the "Accuracy" and "Cost" for Model M1 and Model M2, respectively.

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	50
	-	1	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	100	50
	-	150	200

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	200	90
	-	10	200

Q3. (36 pts) Given the following Boolean matrix M for S1, S2, S3, S4,

	S_1	S_2	S_3	S_4
A	1	0	1	0
B	1	0	0	1
C	0	1	0	1
D	0	1	0	1
E	0	1	0	1
F	1	0	1	0
G	1	0	1	0

a. (12 pts) please compute the signature matrix for S1 - S4, using following permutations:

   [1]. A,E,B,G,F,C,D

   [2]. E,B,C,F,G,A,D

   [3]. D,B,F,G,A,E,C

b. (12 pts) Please compute the pairwise Jaccard similarity of S1 – S4 using (1) the original representation in M; (2) Minhashing generated by the permutations in question (a)

c. (12 pts) Consider the **S1 and S2 only**. When using two hashing functions: $h(x) = (x + 1) \bmod 7$ and $g(x) = (2x + 3) \bmod 7$, what are the signatures of S1 and S2 after hashing?

Q4. (40 pts) Please finish the following text mining mini-project. Please submit your source code together with your PDF file. All text data are provided in the folder "Data".

a. (10 pts) Preprocess all documents by considering the following typical steps:
   [1]. Convert to lowercase
   [2]. Remove stop words
   [3]. Remove Punctuation
   [4]. Single Characters
   [5]. Stemming and Lemmatization, e.g., change "playing" and "played" to play
   [6]. Converting Numbers, e.g., "1000" to one thousand

b. (10 pts) Extract TF-IDF features for all documents. You may consider each unique word as a token and compute the frequency of each word. This will give you a TF-IDF vector for each document.

c. (10 pts) Compute the pairwise cosine similarities of all documents, save in a matrix (.txt) and submit together with your PDF solution.

d. (10 pts) Among these documents, find the top-10 retrieval results (using file names) for the following two sentences. The retrieval can be done by computing the cosine similarity between the "query" and "reference" documents in the data.
   [1]. Once upon a time . . . there were three little pigs, who left their mummy and daddy to see the world.
   [2]. There once lived a poor tailor, who had a son called Aladdin, a careless, idle boy who would do nothing but play all day long in the streets with little idle boys like himself.