Quiz #7: Recommendation Systems

Name: _____ ID: _____

1) (4pts) Given four documents A, B, C, and D and their top two TF-IDF words, A: nba, basketball; B: cancer, health; C: vote, democratic; D: basketball, baseball, write the Boolean feature vectors for each document (2pts) and calculate the cosine similarity between A, D (2pts)

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Feature Vector (nba, basketball, cancer, health, vote, democratic, baseball)

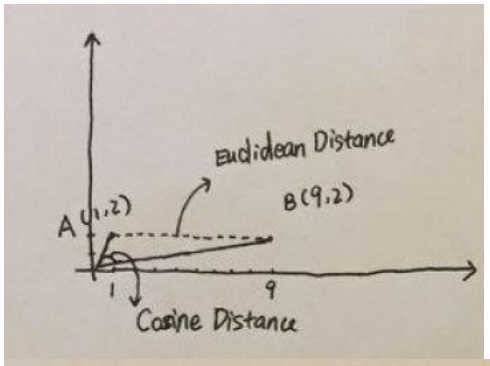
	nba	basketball	cancer	health	vote	democratic	baseball
A	1	1	0	0	0	0	0
B	0	0	1	1	0	0	0
C	0	0	0	0	1	1	0
D	0	1	0	0	0	0	1

(2pts)

Cosine Similarity(A,D) = $1/(\sqrt{2}*\sqrt{2}) = ½$ （2pts）

2) (2pts) On a two-dimensional plane, draw and compare Euclidean Distance and Cosine Distance between A [1, 2] and B [9, 2].

Graph 1pt (no graph will deduct 1pt)



Euclidean Distance $= \sqrt{(1-9)^2+(2-2)^2} = 8$

Cosine Distance $= \arccos\left(\frac{(1,2)\cdot(9,2)}{\sqrt{5}\cdot\sqrt{85}}\right) = \arccos\left(\frac{13}{5\sqrt{17}}\right) \approx 50.9°$

3) (4pts) Given a set of document, briefly explain how to calculate TF and IDF in TF-IDF score. You need to describe any preprocessing you need to apply to the words in

a document (e.g., stemming) (2pts) and how to calculate both the TF and IDF components (2pts).

Preprocessing:

1. Eliminate stop words (1pt)
2. Remove rare words (1pt)
3. Stemming