Quiz5

1) (2pts) 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 (1pt) and calculate the cosine similarity between A, D (1pt)

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

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

	nba	basketball	cencer	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

Cosine Similarity(A,D) = 1/(√2*√2) = ½

2) (2pts)Briefly explain one advantage and one disadvantage of 1) using the content-based approach for finding recommendations and 2) using the dimensionality reduction techniques in your CF recommendation systems.

1.Advantage: No need for data on other users, Able to recommend to users with unique tastes, Able to recommend new & unpopular items, Able to provide explanations

Disadvantage: Finding the appropriate features is hard, Recommendations for new users, Overspecialization

2.Advantage: Reduce the computation

Disadvantage: May lose useful information

3)

1. $w_{2,3} = \frac{(4-2.5)(3-2.5)+(1-2.5)(2-2.5)}{\sqrt{(4-2.5)^2+(1-2.5)^2} \times \sqrt{(3-2.5)^2+(2-2.5)^2}} = 1$

$w_{3,4} = 0$

$$w_{3,5} = \frac{(3-3)\left(2-\frac{10}{3}\right)+(2-3)\left(3-\frac{10}{3}\right)+(4-3)(5-\frac{10}{3})}{\sqrt{(3-3)^2+(2-3)^2+(4-3)^2} \times \sqrt{(2-\frac{10}{3})^2+(3-\frac{10}{3})^2+(5-\frac{10}{3})^2}} = 0.65$$

$P_{2,3} = \frac{\left(2-\frac{5}{2}\right)\times 1+(4-4)\times 0+(1-\frac{10}{3})\times 0.65}{1+0+0.65} + 3 = 1.78$ (1pt, no need to show the result)

4)

(1) average ratings based on all ratings

$$w_{1,3} = \frac{\left(3-\frac{10}{3}\right)\left(5-\frac{8}{3}\right)+\left(5-\frac{10}{3}\right)\left(1-\frac{8}{3}\right)}{\sqrt{\left(3-\frac{10}{3}\right)^2+\left(5-\frac{10}{3}\right)^2}\times\sqrt{\left(5-\frac{8}{3}\right)^2+\left(1-\frac{8}{3}\right)^2}} = -0.73$$

$$w_{2,3} = \frac{\left(4-\frac{8}{3}\right)\left(2-\frac{8}{3}\right)+\left(3-\frac{8}{3}\right)\left(1-\frac{8}{3}\right)}{\sqrt{\left(4-\frac{8}{3}\right)^2+\left(3-\frac{8}{3}\right)^2}\times\sqrt{\left(2-\frac{8}{3}\right)^2+\left(1-\frac{8}{3}\right)^2}} = -0.58$$

$$w_{3,4} = \frac{\left(5-\frac{8}{3}\right)\left(2-\frac{8}{3}\right)+\left(2-\frac{8}{3}\right)\left(3-\frac{8}{3}\right)}{\sqrt{\left(5-\frac{8}{3}\right)^2+\left(2-\frac{8}{3}\right)^2}\times\sqrt{\left(2-\frac{8}{3}\right)^2+\left(3-\frac{8}{3}\right)^2}} = -0.98$$

(0.5pt, no need to show the result to get full credit)

$P_{1,3} = \frac{2\times w_{1,3}+1\times w_{2,3}}{|w_{1,3}|+|w_{2,3}|}$ (0.5pt, no need to show the result.)

(2) average ratings for co-rated items

$$w_{3,1} = \frac{(3-4)(5-3)+(5-4)(1-3)}{\sqrt{(3-4)^2+(5-4)^2}\times\sqrt{(5-3)^2+(1-3)^2}} = -1$$

$$w_{3,2} = \frac{(4-7/2)(2-3/2)+(3-7/2)(1-3/2)}{\sqrt{(4-7/2)^2+(3-7/2)^2}\times\sqrt{(2-3/2)^2+(1-3/2)^2}} = 1$$

$$w_{3,4} = \frac{(5-7/2)(2-5/2)+(2-7/2)(3-5/2)}{\sqrt{(5-7/2)^2+(2-7/2)^2}\times\sqrt{(2-5/2)^2+(3-5/2)^2}} = -1$$

$P_{1,3} = \frac{1\times w_{3,2}+2\times w_{3,1}}{|w_{3,2}|+|w_{3,1}|}$ or $P_{1,3} = \frac{1\times w_{3,2}+3\times w_{3,4}}{|w_{3,2}|+|w_{3,4}|}$

5) Inverse frequency. fi = log(N/ni) (1pt)

e.g) clustering --- get rid of the outsiders. Case Amplification (1pt)

6) Advantage: Simple representation

Disadvantage: Complicated for large dataset