

## Thursday Quiz

- [1 point] Circle ALL of the statements that are TRUE about Content-based Approach.
  - It is able to recommend to users with unique tastes
  - It is not able to recommend unpopular items
  - It never recommends items outside user's content profile
  - It is widely used because finding the appropriate features is easy
- [1 point] Jaccard Similarity works better than Pearson Correlation for User based Collaborative Filtering System.  
**False**
- [1 points] 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. Calculate the cosine similarity between A and D.

**Answer**

Feature vector	nba	basketball	cancer	health	vote	democratic	baseball
<b>A</b>	1	1	0	0	0	0	0
<b>B</b>	0	0	1	1	0	0	0
<b>C</b>	0	0	0	0	1	1	0
<b>D</b>	0	1	0	0	0	0	1

$$\text{Cosine Similarity} = (1*1) / (\text{sqrt}(1^2 + 1^2) * \text{sqrt}(1^2 + 1^2)) = \frac{1}{2}$$

Rubrics: Numerator (0.5 Marks), Denominator (0.5 Marks)

- [2 points] Give two advantages/pros of Content Based Approach with explanation.
  - No need for data on other users
    - No cold-start (for item) or sparsity problems (i.e., new items can receive recommendations)
  - Able to recommend to users with unique tastes

- c. Able to recommend new & unpopular items
  - i. No first-rater problem (i.e., new products never have been rated, therefore they cannot be recommended)
- d. Able to provide explanations
  - i. Can provide explanations of recommended items by listing content-features that caused an item to be recommended

Rubrics : Both correct pros (2 points), 1 correct (1 point)

5. [5 points] Consider the following table with the rows as Users(A,B,C) and columns as Movies(1,2,3). Values represent movie ratings:

	1	2	3
A	2	3	4
B	1	5	3
C	3	2	5

- Calculate the Cosine Similarity Score between user A, user B and user C for the Features of movie rating [2 points].
- Calculate the normalized ratings [1 point].
- What is the new centralized Cosine Similarity Score between A and B in scaled ratings [1 point]?
- What is the relationship between Pearson correlation and cosine similarity? In other words, under what circumstances are these two measures equivalent? Briefly explain. [1 point]

Answer

a) Cosine Similarity

$$\text{Cos}(A,B) = \frac{2*1 + 3*5 + 4*3}{\sqrt{2^2+3^2+4^2}\sqrt{1^2+5^2+3^2}} = \frac{29}{\sqrt{29}\sqrt{35}}$$

$$\text{Cos}(A,C) = \frac{2*3 + 3*2 + 4*5}{\sqrt{2^2+3^2+4^2}\sqrt{3^2+2^2+5^2}} = \frac{32}{\sqrt{29}\sqrt{38}}$$

$$\text{Cos}(B,C) = \frac{1*3 + 5*2 + 3*5}{\sqrt{1^2+5^2+3^2}\sqrt{3^2+2^2+5^2}} = \frac{28}{\sqrt{35}\sqrt{38}}$$

b) Normalized Ratings

User/Item	Average Rating	1	2	3
A	3	-1	0	1
B	3	-2	2	0
C	10/3	-1/3	-4/3	5/3

c) Centralized Cosine Similarity:

$$\text{Cos}(A,B) = \frac{-1*-2 + 0*2 + 1*0}{\sqrt{-1^2+0^2+1^2}\sqrt{-2^2+2^2+0^2}} = \frac{2}{\sqrt{2}\sqrt{8}}$$

d) Pearson Correlation is given as :

$$w_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}}$$

Cosine Similarity is given as :

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

In cosine similarity, when the cosine vectors are normalized by subtracting the vector means, then they are equivalent to pearson correlation

Rubrics:

Part a) - 1 Correct (1 Mark), 2 Correct (1.5 Marks), 3 Correct (2 Marks)

Part b) - Atleast 50% of table is correct - Give 0.5 Marks

Part c) - Correct answer (1 marks)

Part d) (1 mark) for correct explanation.