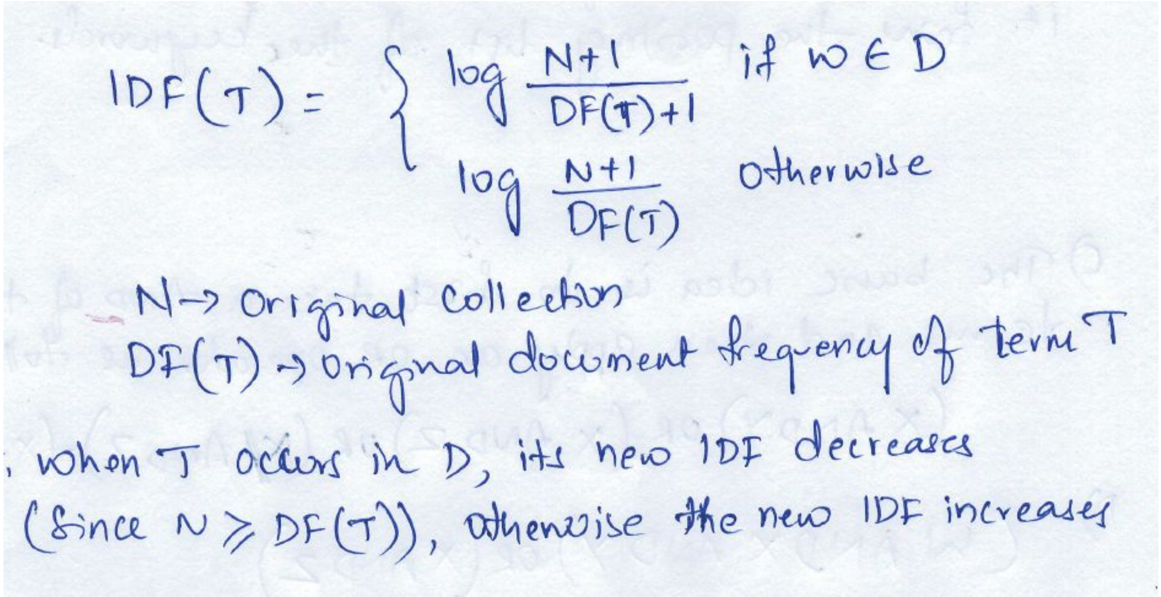
Let D be a document in a text collection. Suppose we add a copy of D to the collection. Would this affect the IDF values of the existing terms in the collection? why or why not? **[2 M]**

It certainly affects the IDF values since IDF(term) = N/DF where N is the number of documents and DF id the document frequency. N increases by 1 when a copy of the document D is added, however the change may be very minimal.



$$IDF(T) = \begin{cases} \log \frac{N+1}{DF(T)+1} & \text{if } w \in D \\ \log \frac{N+1}{DF(T)} & \text{otherwise} \end{cases}$$

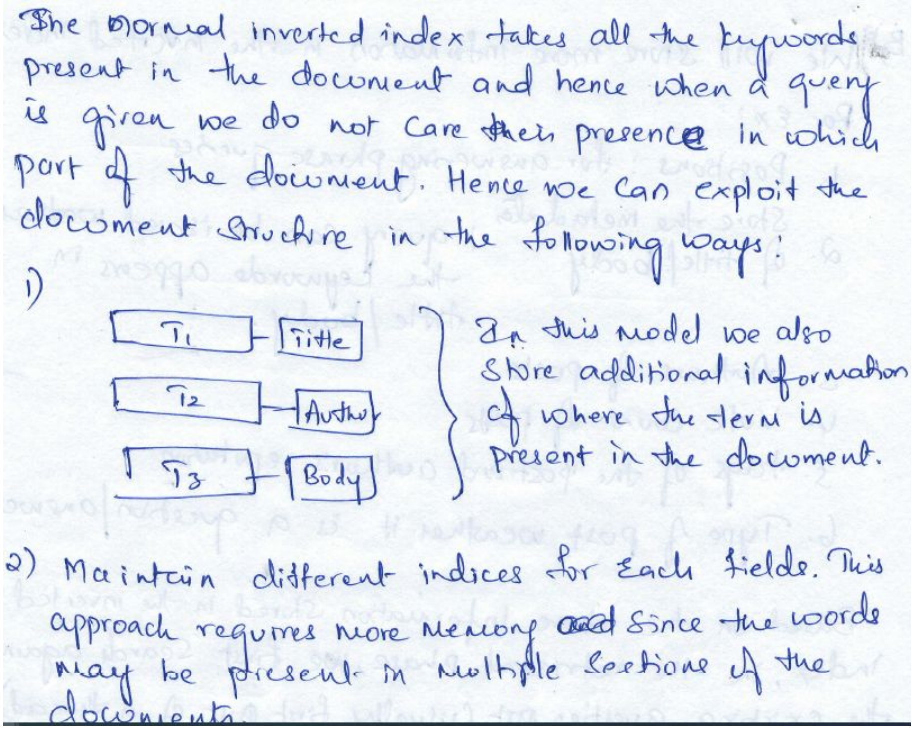
N → Original collection

DF(T) → Original document frequency of term T

, when T occurs in D, its new IDF decreases (since $N \geq DF(T)$), otherwise the new IDF increases

PubMed is a free resource supporting the search and retrieval of peer-reviewed biomedical and life sciences literature. The existing search engine allows you to input keywords and returns all documents having these keywords either in the Title, Author or Body of the document. You are now asked to modify the search engine where the user can also specify where the keyword must be present i.e Title, Author or Body and return documents accordingly.

Suggest two different approaches of constructing inverted index to search in these fields and compare approaches. **[4 M]**



The normal inverted index takes all the keywords present in the document and hence when a query is given we do not care their presence in which part of the document. Hence we can exploit the document structure in the following ways:

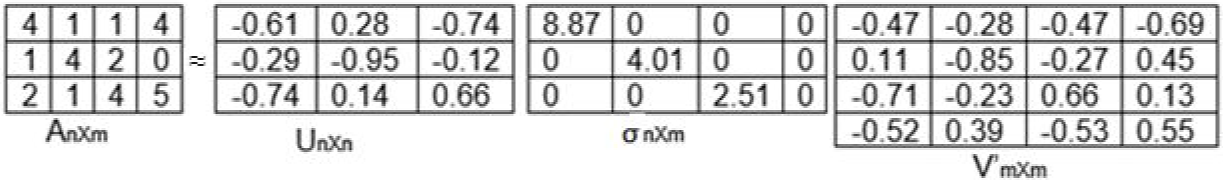
1)

T1
Title
T2
Author
T3
Body

In this model we also store additional information of where the term is present in the document.

2) Maintain different indices for Each fields. This approach requires more Memory ~~and~~ Since the words may be present in multiple Sections of the documents.

Table-2 gives SVD decomposition of the user-movie rating matrix A.



$$\underbrace{\begin{bmatrix} 4 & 1 & 1 & 4 \\ 1 & 4 & 2 & 0 \\ 2 & 1 & 4 & 5 \end{bmatrix}}_{A_{nXm}} \approx \underbrace{\begin{bmatrix} -0.61 & 0.28 & -0.74 \\ -0.29 & -0.95 & -0.12 \\ -0.74 & 0.14 & 0.66 \end{bmatrix}}_{U_{nXn}} \underbrace{\begin{bmatrix} 8.87 & 0 & 0 & 0 \\ 0 & 4.01 & 0 & 0 \\ 0 & 0 & 2.51 & 0 \end{bmatrix}}_{\sigma_{nXm}} \underbrace{\begin{bmatrix} -0.47 & -0.28 & -0.47 & -0.69 \\ 0.11 & -0.85 & -0.27 & 0.45 \\ -0.71 & -0.23 & 0.66 & 0.13 \\ -0.52 & 0.39 & -0.53 & 0.55 \end{bmatrix}}_{V'_{mXm}}$$

Table-2

A. What is the percentage of variance if the first two concepts are retained? What will the resulting sizes of U, σ,V^T after retaining the first two concepts?

$$\frac{8.87+4.01}{15.39} \times 100 = 83.7\ \%$$

If the first two concepts are retained U, σ, V^T will have the following sizes

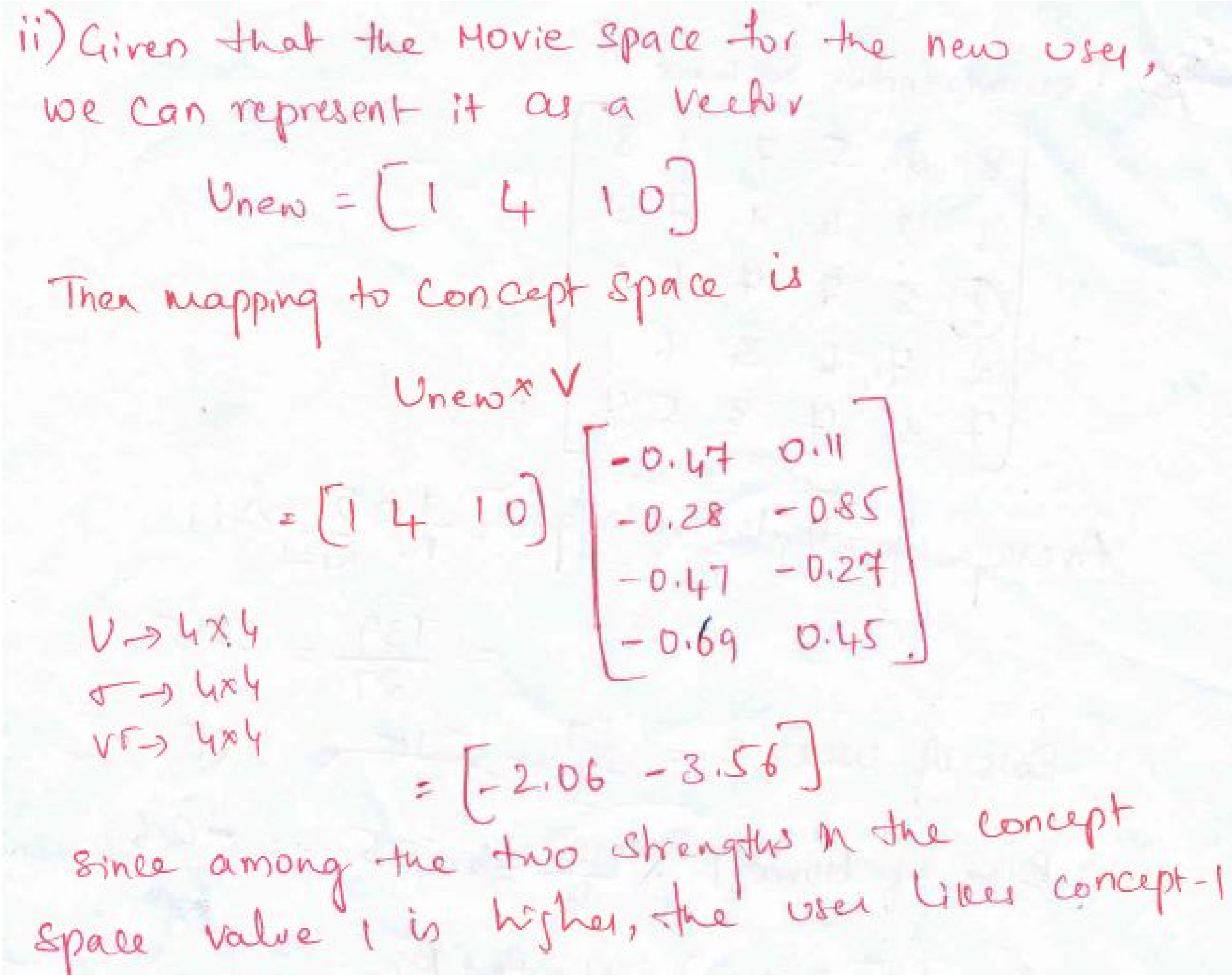
where $n = 3$
$m = 4$

$U \rightarrow n \times 2$

$\sigma = 2 \times 2$

$V^T = 2 \times m$

B. Given the test user U={1,4,1,0} show how to project him in the concept space? What will be resulting sizes of U, σ,V^T matrices? What inference you can make from the values when the user is projected into the concept space?



ii) Given that the Movie space for the new user, we can represent it as a vector

$$U_{new} = [1 \ \ 4 \ \ 1 \ \ 0]$$

Then mapping to concept space is

$$U_{new} * V$$

$$= [1 \ \ 4 \ \ 1 \ \ 0] \begin{bmatrix} -0.47 & 0.11 \\ -0.28 & -0.85 \\ -0.47 & -0.27 \\ -0.69 & 0.45 \end{bmatrix}$$

$V \to 4 \times 4$

$\sigma \to 4 \times 4$

$V\sigma \to 4 \times 4$

$$= [-2.06 \ \ -3.56]$$

Since among the two strengths in the concept space value 1 is higher, the user likes concept-1

[3+3=6M]