**Subject: Information Retrieval**





ốộ



ốộ

Corpus1:

Tôi yêu ML

Tôi ghét DL

DL là phương\_pháp ML số\_một

Tôi yêu Sơn\_Tùng

Sơn\_Tùng là số\_một

1.5

Idf: 0.73 1.32. 1.32. 2.32 1.32 1.32. 2.32. 1.32. 1.32

Toi yeu ML ghet DL la pp so1 ST

C1: 1 1 1 0 0 0 0 0 0

C2: 1 0 0 1 1 0 0 0 0

C3: 0 0 1 0 1 1 1 1 0

C4: 1 1 0 0 0 0 0 0 1.

C5: 0 0 0 0 0 1 0 1 1

Toi yeu ML ghet DL la pp so1 ST

C1: 0.73 1.32 1.32 0 0 0 0 0 0

C2: 0.73 0 0 2.32 1.32 0 0 0 0

C3: 0 0 1.32 0 1.32 1.32 2.32 1.32 0

C4: 0.73 1.32 0 0 0 0 0 0 1.32

C5: 0 0 0 0 0 1.32 0 1.32 1.32

Corpus

Normalization

Term frequency: tf

inversed document frequency: idf = log\_2(N/df) = log\_2(5/3)

df(tôi) = 3

idf(tôi) = log(1/3) = log\_2(N/df)

Document frequence: df: df càng lớn --> độ quan trọng càng nhỏ

df: df càng nhỏ--> độ quan trọng càng lớn

inversed document frequency:

idf: df càng nhỏ--> idf càng lớn --> độ quan trọng càng lớn

{music, sport}

D1: Son\_Tung

D2: Messi

D3: và

===

Corpus2:

Tôi yêu ML Tôi ghét DL DL là phương\_pháp ML số\_một

Tôi yêu Sơn\_Tùng Sơn\_Tùng là số\_một

Toi yeu ML ghet DL la pp so1 ST

C1: 1 0.5 1 0.5 1 0.5 0.5 0.5 0

C2: 0.5 0.5 0 0 0 0.5 0 0.5 1

Toi yeu ML ghet DL la pp so1 ST

C1: 0 0 1 1 1 0 1 0 0

C2: 0 0 0 0 0 0 0 0 1