Atharva Prashant Pawar (9427) Comps-A [Batch-D]

[NLP - Ass - 2]

TF - IDF

Q1) Doc - Database.

* Corpus D :

		mon
D1	: Information Retrieval System	3
D2	: Information storage	2
D3	: Digital speech Synthesis System	4
D4	: Speech filtering	2
D5	: Speech retrieval	2

(1) TF value table: (Term Freq. Matrix)

		Document / Term.				
	Words.	D1	D2	D3	D4	D5
w1	Information	1/3	1/2	0	0	0
w2	Retrieval	1/3	0	0	0	1/2
w3	system	1/3	0	1/4	0	0
w4	Storage	0	1/2	0	0	0
w5	Digital	0	0	1/4	0	0
w6	Speech	0	0	1/4	1/2	1/2
w7	Synthesis	0	0	1/4	0	0
w8	filtering	0	0	0	1/2	0

* Formula :

$$TF(w,d) = \frac{\text{Occurance of } w \text{ in } d}{\text{total no. of } w \text{ in } d}$$

Teacher's Signature:

IDF formula: $IDF(w_1 \text{ in } D) = \log\left(\frac{N}{df_i}\right) = \log\left(\frac{5}{2}\right)$

↳ Corpus D

② TF → IDF (Inverse Doc. Freq)

Formula: $w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$

$tf_{i,j}$ = no. of occurrences of i in j
df_i = no. of documents containing i
N = total no. of documents

$TF - IDF(w_1 \text{ in } D1) = 0.33 * \log\left(\frac{5}{2}\right) = 0.1313$

$TF\text{-}IDF(w_2 \text{ in } D1) = 0.33 * \log(5/2) = 0.1313$

$TF\,IDF(w_3 \text{ in } D1) = 0.33 * \log(5/2) = 0.1313$

⋮

② Find IDF for words:

Formula: $IDF(w_i, D) = \log\left(\frac{N}{df_i}\right)$

$IDF(w_1, D) = \log\left(\frac{5}{2}\right) = 0.39$

$IDF(w_2, D) = \log(5/2) = 0.39$
$IDF(w_3, D) = \log(5/2) = 0.39$
$IDF(w_4, D) = \log(5/1) = 0.69$
$IDF(w_5, D) = \log(5/1) = 0.69$
$IDF(w_6, D) = \log(5/3) = 0.22$
$IDF(w_7, D) = \log(5/1) = 0.69$
$IDF(w_8, D) = \log(5/1) = 0.69$

Teacher's Signature:.....

(3) Find TF-IDF for word = "Speech"

~~TF-IDF ("speech" in D)~~ =

Table TF IDF for word = Speech.

Document	TF(speech)	IDF(speech)	TF-IDF
D1	0	0.22	0
D2	0	0.22	0
D3	1/4	0.22	0.055
D4	1/2	0.22	0.11
D5	1/2	0.22	0.11

TF-IDF formula:

$$\text{TF-IDF}(\text{Word in Docu}) = tf * \log\left(\frac{N}{dfi}\right)$$

D1 → 0 * log 0.22 = 0
D2 → 0 * log 0.22 = 0
D3 → 0.25 * log 0.22 = 0.055
D4 → 0.5 * log 0.22 = 0.11
D5 → 0.5 * log 0.22 = 0.11

So the TF-IDF values for the term "speech" in 2 docum are approx 0.11 for D4, D5 & for D3 it's 0.055. This means that speech is most relevant to ~~the~~ these 3 docs. Among 3 two are more relevant to the term 'Speech' because they have the highest TF-IDF score.

Teacher's Signature

Q# 2) Explain.

A. Synonym:
A synonym is a word or phrase that has a similar or identical meaning to ~~other~~ another word or phrase in the same language.
eg: happy & joyful

B. Antonym:
An Antonym is a word that has the opposite meaning of another word.
eg: hot & cold

C. Hyponym:
It is a word whose meaning is a subset of a more general word.
eg: Rose & flower

D. Hypernym:
It is a word that represents a more general category or superclass of words.
eg: Fruit & apple

E. Meronym:
It is a word that represents a part or a component of a larger whole.
eg: Wheel & car.

Teacher's Signature:.....

a) Couch - Sofa
Synonym
Couch & sofa are synonyms because they both refer to a similar piece of furniture designed for seating.

b) Car - Wheel.
Meronym:
Wheel is a part of a car, so Pts a meronymic

c) Meal - Breakfast:
Hyponym
Breakfast is a specific type of meal, so Pts a hyponymic relationship.
Breakfast is a subset of meal.

d) I left my heart - and my suitcase.
No direct lexical relationship.
No direct lexical relat b/w heart & suitcase.
they do not represent typical word relationship

e) Mammal - Dog:
Hyponym
Dog is a specific type of mammal., so Pts a hyponymic relation.

Dogs belongs to category mammals.

Teacher's Signature:.....

Q3)

A: Jupiter is the largest planet

B: Mars is the fourth planet from the earth

* Vector Form

a)

① Traditional word count method.

Words	A	B
Jupiter	1	0
is	1	1
the	1	2
largest	1	0
planet	1	1
Mars	0	1
Fourth	0	1
from	0	1
earth	0	1

∴ Vector A = [1, 1, 1, 1, 1, 0, 0, 0, 0]

Vector B = [0, 1, 2, 0, 1, 1, 1, 1, 1]

② estimate similarity b/w doc A & B using cosine similarity Formula:

$$\text{Similarity}(A, B) = \frac{\sum A_i \times B_i}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$\therefore A \cdot B = (1\times0 + 1\times1 + 1\times2 + 1\times0 + 1\times1 + 0\times1 + 0\times1 + 0\times1 + 0\times1)$$

$$= 0+1+2+0+1+0+0+0+0$$

$$= 4$$

$$|A| = \sqrt{5} \qquad |B| = \sqrt{10}$$

Teacher's Signature:.....

IDF = log(2

∴ Similarity (A, B) = $\frac{4}{\sqrt{5} \times \sqrt{10}} = \frac{4}{5.9160} = 0.56$

b) TF IDF method. cosine similarity.

words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
Jupiters	1/5	0	log(2/1) = 0.30	0.06	0
Is	1/5	1/8	log(2/2) = 0	0	0
the	1/5	2/8	log(2/2) = 0	0	0
largest	1/5	0	log(2/1) = 0.30	0.06	0
planet	1/5	1/8	log(2/2) = 0	0	0
mars	0	1/8	log(2/1) = 0.30	0	0.0375
fourth	0	1/8	log(2/1) = 0.30	0	0.0375
from	0	1/8	log(2/1) = 0.30	0	0.0375
earth	0	1/8	log(2/1) = 0.30	0	0.0375

Vector A = [0.06, 0, 0, 0.06, 0, 0, 0, 0, 0]

Vector B = [0, 0, 0, 0, 0, 0.0375, 0.0375, 0.0375, 0.0375]

Similarity (A, B) = $\frac{A \cdot B}{|A| \times |B|}$

∴ A · B = 0 |A| = 2 |B| = 4

similarity (A, B) = $\frac{0}{2 \times 4} = \frac{0}{8} = 0$

Teacher's Signature:.....

c) Traditional word count method yields a higher cosine similarity of approx. 0.56 indicating that the 2 doc share more common words. However it doesn't consider the importance of words.

TF-IDF method takes into account the importance of words. It gives a lower cosine similarity of approx 0., indicating that the 2 doc. are less similar when considering the importance of words.

d) Other similarity measures.

1. Jaccard similarity.
measures the similary b/w 2 sets by comparing the size of their intersection & size of their union.

$$\text{Jaccard similarity}(A, B) = \left| \frac{A \cap B}{A \cup B} \right|$$

2. Euclidean Distance:
It measures the straigh-line distance b/w 2 points in a multidimensional space.

$$\text{Euclidean Distance}(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

3. Pearson Correlation Coefficient:
measures the linear relationship b/w 2 vectors.

$$\text{Person corr coef}(A, B) = \frac{\sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^{n} (B_i - \bar{B})^2}}$$

Teacher's Signature: