

Isha Ikhlaq.

Assignment - 4

Q: 1 Compute BOW, TF, IDF & TF-IDF

Vocabulary (unique terms):

data, science, is, one, of, the, most, important, courses, in, computer, this, best, scientists, perform, analysis.

Bag of Words (BOW):

Term	S_1	S_2	S_3
data	1	1	2
science	2	1	0
is	1	1	0
one	1	1	0
of	1	1	0
the	1	1	1
most	1	0	0
important	1	0	0
courses	1	1	0
in	1	0	0
computer	1	0	0
this	0	1	0
best	0	1	0

Date _____

Term	S_1	S_2	S_3
scientists	0	0	1
perform	0	0	1
analysis	0	0	1

Vector S_1 : [1 2 1 1 1 1 1 1 1 1 0 0 0 0 0]

Vector S_2 : [1 1 1 1 1 0 0 1 0 0 1 1 0 0 0]

Vector S_3 : [2 0 0 0 0 1 0 0 0 0 0 0 1 1 1]

TF (Term Frequency)			
Term	S_1	S_2	S_3
data	$1/12$	$1/9$	$2/6$
science	$2/12$	$1/9$	0
is	$1/12$	$1/9$	0
one	$1/12$	$1/9$	0
of	$1/12$	$1/9$	0
the	$1/12$	$1/9$	$1/6$
most	$1/12$	0	0
important	$1/12$	0	0
courses	$1/12$	$1/9$	0
in	$1/12$	0	0
computer	$1/12$	0	0
this	0	$1/9$	0
best	0	$1/9$	0
scientists	0	0	$1/6$
perform	0	0	$1/6$
analysis	0	0	$1/6$

Inverse Document Frequency (IDF)

$$\text{idf}(\text{data}) = \log(3/3) = 0$$

$$\text{idf}(\text{science}) = \log(3/2) = 0.18$$

$$\text{idf}(\text{is}) = \log(3/2) = 0.18$$

$$\text{idf}(\text{one}) = \log(3/2) = 0.18$$

$$\text{idf}(\text{of}) = \log(3/2) = 0.18$$

$$\text{idf}(\text{the}) = \log(3/3) = 0$$

$$\text{idf}(\text{most}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{important}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{courses}) = \log(3/2) = 0.48$$

$$\text{idf}(\text{in}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{computer}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{this}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{best}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{scientists}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{perform}) = \log(3/1) = 0.48$$

$$\text{idf}(\text{analysts}) = \log(3/1) = 0.48$$

TF - IDF :

Term	$tf \times idf (S1)$	$tf \times idf (S2)$	$tf \times idf (S3)$
data	$\frac{1}{12} \times 0 = 0$	0	0
science	0.03	0.02	0
is	0.015	0.02	0
one	0.015	0.02	0
of	0.015	0.02	0
the	0	0	0
most	0.04	0	0
important	0.04	0	0
courses	0.04	0.02	0
in	0.04	0	0
computer	0.04	0	0
this	0	0.053	0
best	0	0.053	0
scientists	0	0	0
perform	0	0	0
analysis	0	0	0

Q: 2.

Bag of words

$$\cos \theta = \frac{S_1 + S_2 + S_3}{|S_1| |S_2| |S_3|}$$

$$|S_1| |S_2| |S_3|$$

$$\cos(S_1, S_2) = \frac{S_1 \cdot S_2}{|S_1| |S_2|}$$

$$S_1 \cdot S_2 = 1 \cdot 1 + 2 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 1 +$$

$$1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 1$$

$$S_1 \cdot S_2 = 9$$

$$|S_1| = (1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 2)$$

$$= 14^{0.5} = 3.7417$$

$$|S_2| = \sqrt{9} = 3$$

$$|S_1| |S_2| = 11.2251$$

$$\cos(S_1, S_2) = 0.8017$$

$$\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{|S_2| |S_3|}$$

$$S_2 \cdot S_3 = 2 + 1 + 0 = 3$$

$$|S_2| = \sqrt{9} = 3$$

$$|S_3| = \sqrt{6} = 2.4495$$

$$\cos(S_2, S_3) = 0.4082$$

$$\cos(S_1, S_3) = \frac{S_1 \cdot S_3}{|S_1| |S_3|}$$

$$S_1 \cdot S_3 = 1 + 2 + 1 + 1 + 1 + 1 + 0 = 8$$

$$S_1 \cdot S_3 = 2 + 1 + 0 = 3$$

$$|S_1| = 3.7417$$

$$|S_3| = \sqrt{6} = 2.4495$$

$$\cos(S_1, S_3) = 0.327$$