SP20-BSE-009

Date _____

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 S1: [1 2 1 1 1 1 1 | 1 1 1 | 0 0 0 0 0]

Vector S2: [1 1 | 1 1 1 | 0 0 | 1 0 0 | 1 1 | 0 0 0]

Vector S3: [2 0 0 0 0 1 | 0 0 0 0 0 0 | 1 1 1]

Date _____

TF (Term Frequency)

Term	S_1	S_2	S_3
data	1/12	1/9	2/6
scienE	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
coluses	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

Date

Inverse Document Frequency (IDF)

idf (data) = log (3/3) = 0

idf (science) = log (3/2) = 0.18

idf (is) = log (3/2) = 0.18

idf (one) = log (3/2) = 0.18

idf (of) = log (3/2) = 0.18

idf (the) = log (3/3) = 0

idf (most) = log (3/1) = 0.48

idf (important) = log (3/1) = 0.48

idf (courses) = log (3/2) = 0.48

idf (in) = log (3/1) = 0.48

idf (computer) = log (3/1) = 0.48

idf (this) = log (3/1) = 0.48

idf (best) = log (3/1) = 0.48

idf (scientists) = log (3/1) = 0.48

idf (perform) = log (3/1) = 0.48

idf (analysis) = log (3/1) = 0.48

Date _____

TF - IDF :

Term	tf x idf (S1)	tf x idf (S2)	tf x idf (S3)
data	1/12 x 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