IDS-GII
Assignment no 5 (NLP)
FA19-BCS-131

Question: no 1
Compute the BoW, TF, IDF and TFIDF.
S_1 "Sunshine state enjoy sunshine"
S_2 " Brown fox jump high, brown fox run"
S_3 " Sunshine State fox fun fast".

For Bow document term matrix

	Sunshine	state	enjoy	brown	fox	jump	high	run	fast
S_1:	2	1	1	0	0	0	0	0	0
S_2	0	0	0	2	2	1	1	1	0
S_3	1	1	0	0	1	0	0	1	1

Vector S_1: [2, 1, 1, 0, 0, 0, 0, 0, 0]
Vector S_2: [0, 0, 0, 2, 2, 1, 1, 1, 0]
Vector S_3: [1, 1, 0, 0, 1, 0, 0, 1, 1]

Term - Frequency.

	Sunshine	state	enjoy	brown	fox	jump	high	run	fast
tf-S_1	2/4	1/4	1/4	0	0	0	0	0	0
tf-S_2	0	0	0	2/7	2/7	1/7	1/7	1/7	0
Tf-S_3	1/5	1/5	0	0	1/5	0	0	1/5	1/5

: Inverse Document Frequency (IDF)

Idf (Sunshine) = log (3/2) = 0.18
Idf (State) = log (3/2) = 0.18
Idf (enjoy) = log (3/1) = 0.48
Idf (brown) = log (3/1) = 0.48
Idf (fox) = log (3/2) = 0.18
Idf (jump) = log (3/1) = 0.48
Idf (high) = log (3/1) = 0.48
Idf (run) = log (3/2) = 0.18
Idf (fast) = log (3/1) = 0.48

TF · IDF
Term Frequency · Inverse Document Frequency.

S1 →
tf · idf (Sunshine) = 2 × 0.18 = 0.09
tf · idf (state) = 1/4 × 0.18 = 0.045
tf · idf (enjoy) = 1/4 × 0.48 = 0.12

S2 →
tf · idf (brown) = 2/7 × 0.48 = 0.14
tf · idf (fox) = 2/7 × 0.18 = 0.05
tf · idf (jump) = 1/7 × 0.48 = 0.07
tf · idf (high) = 1/7 × 0.48 = 0.07
tf · idf (run) = 1/7 × 0.18 = 0.03

$S_3 \rightarrow$

tf.idf (sunshine) = 1/5 x 0.18 = 0.04
tf.idf (state) = 1/5 x 0.18 = 0.04
tf.idf (fox) = 1/5 x 0.18 = 0.04
tf.idf (run) = 1/5 x 0.18 = 0.04
tf.idf (fast) = 1/5 x 0.48 = 0.096

Q2

Cosine Similarity.
S_1 & S_3.
TF. IDF vectors.

$S1 = [0.09, 0.045, 0.12, 0, 0, 0, 0, 0, 0]$
$S_3 = [0.04, 0.04, 0, 0, 0.04, 0, 0, 0, 0.04, 0.096]$

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

$S_1 \cdot S_3 = (0.09 \times 0.04) + (0.045 \times 0.04) + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0.0054$

$|S_1| = \sqrt{(0.09 \times 0.09) + (0.045 \times 0.045) + (0.12)^2 + 0 + 0 + 0 + 0 + 0 + 0} = 0.15660$

~~$|S_3| = \sqrt{0.0016 + 0.0016 +}$~~

$|S_3| = \sqrt{(0.04)(0.04) + (0.04 \times 0.04) + (0.04 \times 0.04) + (0.04 \times 0.04) + (0.096 \times 0.096)}$

$= 0.125$

$$\cos(S_1, S_3) = \frac{0.0054}{0.1566 \times 0.125}$$

$$\cos(S_1, S_3) = 0.2759$$