# Introduction To Data Science


Name: Muhammad Zubair
Registration no: SP-20-BCS-025
Group: 4 (IV)
Submitted to: Sir Muhammad Sharjeel


## Assignment no 5
## Question no 1

$S_1$: "sunshine state enjoy sunshine"
$S_2$: "brown for jump high, brown for sun"
$S_3$: "sunshine state for sun fast"

### Bow Model

| | Sunshine | state | enjoy | brown | for | jump | high | sun | fast | Total length |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| $S_2$ | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 1 | 0 | 7 |
| $S_3$ | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 5 |

### If Model

| | Sunshine | state | enjoy | brown | for | jump | high | run | fast | Total sun |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 2/4 | 1/4 | 1/4 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| $S_2$ | 0 | 0 | 0 | 2/7 | 2/7 | 1/7 | 1/7 | 1/7 | 0 | 7 |
| $S_3$ | 1/5 | 1/5 | 0 | 0 | 1/5 | 0 | 0 | 1/5 | 1/5 | 5 |

# IDF Model

Formula

$$IDF('word') = \log\left(\frac{\text{Total no. of Documents}}{\text{no of Documents containing}}\right)$$

IDF (sunshine) = $\log(3/2) = 0.1761$

IDF (state) = $\log(3/2) = 0.1761$

IDF (enjoy) = $\log(3/1) = 0.4771$

IDF (brown) = $\log(3/1) = 0.4771$

IDF (fox) = $\log(3/2) = 0.1761$

IDF (Jump) = $\log(3/1) = 0.4771$

IDF (high) = $\log(3/1) = 0.4771$

IDF (run) = $\log(3/2) = 0.1761$

IDF (fast) = $\log(3/1) = 0.4771$

| | sunshine | state | enjoy | brown | fox | jump | high | run | fast |
|---|---|---|---|---|---|---|---|---|---|
| IDF | 0.1761 | 0.1761 | 0.4771 | 0.4771 | 0.1761 | 0.4771 | 0.4771 | 0.1761 | 0.4771 |

# TF-IDF values

For $S_1$:

tf · idf (sunshine) = 2/4 * 0.1761 = 0.0880

tf · idf (state) = 1/4 * 0.1761 = 0.440

tf · idf (enjoy) = 1/4 * 0.4471 = 0.1192

for $S_2$:

tf · idf (brown) = (2/7)(0.4771) = 0.1363

tf · idf (fox) = (2/7)(0.1761) = 0.0503

tf · idf (jump) = (1/7)(0.4771) = 0.0681

tf · idf (high) = (1/7)(0.4771) = 0.0681

tf · idf (sun) = (1/7)(0.1761) = 0.0251

for $S_3$

tf · idf (sunshine) = (1/5)(0.1761) = 0.0352

tf · idf (state) = (1/5)(0.1761) = 0.0352

tf · idf (fox) = (1/5)(0.1761) = 0.0352

tf idf (sun) = (1/5)(0.1761) = 0.0352

tf idf (fast) = (1/5)(0.4771) = 0.0954

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| sunshine | 0.0880 | 0 | 0.0352 |
| state | 0.0440 | 0 | 0.0362 |
| enjoy | 0.1192 | 0 | 0 |
| brown | 0 | 0.1363 | 0 |
| fox | 0 | 0.0503 | 0.0352 |
| jump | 0 | 0.0681 | 0 |
| high | 0 | 0.0681 | 0 |
| sun | 0 | 0.0251 | 0.0352 |
| fast | 0 | 0 | 0.0954 |

## Question no 02:

Cosine Similarity b/w $S_1$ & $S_3$ using bow model to generate vectors

$S_1 = < 2\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 >$

$S_3 = < 1\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1 >$

formula

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

$S_1 \cdot S_3 = (2)(1) + (1)(1) + (1)(0) + (0)(0) + (0)(1) + (0)(0) + (0)(0) + (0)(1) + (0)(1)$

$= 2 + 1$

$= 3$

$|S_1| = \sqrt{2^2+1^2+1^2+0^2+0^2+0^2+0^2+0^2+0^2}$

$= \sqrt{4+1+1} = \sqrt{6}$

$|S_1| = 2.4494$

$|S_3| = \sqrt{1^2+1^2+0^2+0^2+1^2+0^2+0^2+1^2+1^2}$

$= \sqrt{5}$

$|S_3| = 2.2360$

$\cos(S_1, S_3) = \frac{3}{(2.4494)(2.2360)}$

$\boxed{\cos(S_1, S_3) = 0.5477}$

$S_1, S_3 = \cos^{-1}(0.5477)$

$\boxed{S_1, S_3 = 56.78}$