NAME: ALEENA KHAN ROLL NO: SP20-BCS-112

Group #4:

Q: S1 : sunshine state enjoy sunshine

S2 : brown fox jump high, brown fox run

S3 : sunshine state fox run fast .

| | Bow 1 | Bow 2 | Bow 3 | TF-1 | TF-2 | TF-3 | IDF |
|---|---|---|---|---|---|---|---|
| sunshine | 2 | 0 | 1 | 0.5 | 0 | 0.2 | 0.176 |
| state | 1 | 0 | 1 | 0.25 | 0 | 0.2 | 0.176 |
| enjoy | 1 | 0 | 0 | 0.25 | 0 | 0 | 0.4771 |
| brown | 0 | 2 | 0 | 0 | 0.285 | 0 | 0.4771 |
| fox | 0 | 2 | 1 | 0 | 0.285 | 0.2 | 0.176 |
| jump | 0 | 1 | 0 | 0 | 0.142 | 0 | 0.4771 |
| high | 0 | 1 | 0 | 0 | 0.142 | 0 | 0.4771 |
| run | 0 | 1 | 1 | 0 | 0.142 | 0.2 | 0.176 |
| fast. | 0 | 0 | 1 | 0 | 0 | 0.2 | 0.4771 |
| length: | 4 | 7 | 5 | | | | |

## Formulas:

- BoW = count in single document

- TF = $\frac{\text{count in single doc}}{\text{length of document.}}$

تاریخ: _____

**IDF** = $\log\left(\frac{\text{no of docu}}{\text{no of docu in which it occured}}\right)$

- sunshine = $\log\left(\frac{3}{2}\right) = 0.176$

state = log (3/2) = 0.176

enjoy = log (3/1) = 0.4771

brown = log (3/1) = 0.4771

fox = log (3/2) = 0.176

jump = log (3/1) = 0.4771

high = log (3/1) = 0.4771

run = log (3/2) = 0.176

fast = log (3/1) = 0.4771

تاریخ: _____

For S1.S3:

$= (2\times1) + (1\times1) + (1\times0) + (0\times0) + (0\times1) +$
$(0\times0) + (0\times0) + (0\times1) + (0\times1)$

$= 2 + 1 = 3$

$|S1| = \sqrt{2^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2}$
$= \sqrt{4+1+1} = \sqrt{6} = 2.45$

$|S3| = \sqrt{1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2}$
$= \sqrt{1+1+1+1+1} = \sqrt{5} = 2.24$

So $\cos(S1, S3) = \dfrac{S1 . S3}{|S1||S3|}$

$= \dfrac{3}{(2.45)(2.24)} = 0.55$

When I was trying to compute the TF-IDF matrix by hand using the standard formula, after computing TF and IDF first and then multiplying the two I did realize, there was something different compared to the one I obtained with Scikit-learn on my sample corpus, there I realized the difference between the Scikit-learn version and most standard and traditional version.

**Formulas:**

Scikit-Learn

- $\text{IDF}(t) = \log\frac{1+n}{1+\text{df}(t)} + 1$

Standard notation

- $\text{IDF}(t) = \log\frac{n}{\text{df}(t)}$