2022

Assignment-5

INTRODUCTION TO DATA SCIENCE

Course: Introduction to Data Science
Submitted To: Muhammad Sharjeel
Submitted By: Muhammad Faizan
Student Id: CIIT/SP20-BCS-101/LHR
Section: G2 - B

Assignment Report

Natural Language Processing
(IDS)

Question: 1

Compute the BoW model, TF Model, and IDF model for each of the terms in the following three sentences. Then calculate the TF.IDF values

S1 “sunshine state enjoy sunshine”
S2 “brown fox jump high, brown fox run”
S3 “sunshine state fox run fast”

Features:

['brown', 'enjoy', 'fast', 'fox', 'high', 'jump', 'run', 'state', 'sunshine']

Vocabulary:

{'sunshine': 8, 'state': 7, 'enjoy': 1, 'brown': 0, 'fox': 3, 'jump': 5, 'high': 4, 'run': 6, 'fast': 2}

BoW Model:

	Brown	**Enjoy**	**Fast**	**Fox**	**High**	**Jump**	**Run**	**State**	**Sunshine**	**Total Length**
S1	0	1	0	0	0	0	0	1	2	**4**
S2	2	0	0	2	1	1	1	0	0	**7**
S3	0	0	1	1	0	0	1	1	1	**5**

Vectors:

Sentence 1 : [0 1 0 0 0 0 0 1 2]
Sentence 2 : [2 0 0 2 1 1 1 0 0]
Sentence 3 : [0 0 1 1 0 0 1 1 1]

TF Model:

	Brown	Enjoy	Fast	Fox	High	Jump	Run	State	Sunshine
S1	0	1/4	0	0	0	0	0	1/4	2/4
S2	2/7	0	0	2/7	1/7	1/7	1/7	0	0
S3	0	0	1/5	1/5	0	0	1/5	1/5	1/5

IDF Model:

	Idf
Brown	$\log\left(\frac{3}{1}\right) = 0.48$
Enjoy	$\log\left(\frac{3}{1}\right) = 0.48$
Fast	$\log\left(\frac{3}{1}\right) = 0.48$
Fox	$\log\left(\frac{3}{2}\right) = 0.18$
High	$\log\left(\frac{3}{1}\right) = 0.48$
Jump	$\log\left(\frac{3}{1}\right) = 0.48$
Run	$\log\left(\frac{3}{2}\right) = 0.18$
State	$\log\left(\frac{3}{2}\right) = 0.18$
Sunshine	$\log\left(\frac{3}{2}\right) = 0.18$

TF-IDF Model:

	S1	S2	S3
Brown	$0 * 0.48 = \mathbf{0}$	$\frac{2}{7} * 0.48 = \mathbf{0.137}$	$0 * 0.48 = \mathbf{0}$
Enjoy	$\frac{1}{4} * 0.48 = \mathbf{0.12}$	$0 * 0.48 = \mathbf{0}$	$0 * 0.48 = \mathbf{0}$
Fast	$0 * 0.48 = \mathbf{0}$	$0 * 0.48 = \mathbf{0}$	$\frac{1}{5} * 0.48 = \mathbf{0.096}$
Fox	$0 * 0.18 = \mathbf{0}$	$\frac{2}{7} * 0.18 = \mathbf{0.051}$	$\frac{1}{5} * 0.18 = \mathbf{0.036}$
High	$0 * 0.48 = \mathbf{0}$	$\frac{1}{7} * 0.48 = \mathbf{0.068}$	$0 * 0.48 = \mathbf{0}$
Jump	$0 * 0.48 = \mathbf{0}$	$\frac{1}{7} * 0.48 = \mathbf{0.068}$	$0 * 0.48 = \mathbf{0}$
Run	$0 * 0.48 = \mathbf{0}$	$\frac{1}{7} * 0.18 = \mathbf{0.026}$	$\frac{1}{5} * 0.18 = \mathbf{0.036}$
State	$\frac{1}{4} * 0.18 = \mathbf{0.045}$	$0 * 0.18 = \mathbf{0}$	$\frac{1}{5} * 0.18 = \mathbf{0.036}$
Sunshine	$\frac{2}{4} * 0.18 = \mathbf{0.09}$	$0 * 0.18 = \mathbf{0}$	$\frac{1}{5} * 0.18 = \mathbf{0.036}$

Question: 2

Compute the cosine similarity between S1 and S3.

S1 “sunshine state enjoy sunshine”
S2 “brown fox jump high, brown fox run”
S3 “sunshine state fox run fast”

Sentence 1 : [0 1 0 0 0 0 0 1 2]

Sentence 3 : [0 0 1 1 0 0 1 1 1]

$$\cos(\text{S1, S3}) = \frac{(S1\ .\ S3)}{|S1|\ |S3|}$$

$(S1\ .\ S3)$ = (0*0 + 1*0 + 0*1 + 0*1 +0*0 + 0*0 + 0*1 + 1*1 + 2*1) = 3

|S1| = $\sqrt{0*0+1*1+0*0+0*0+0*0+0*0+0*0+1*1+2*2}$ = 2.45

|S3| = $\sqrt{0*0+0*0+1*1+1*1+0*0+0*0+1*1+1*1+1*1}$ = 2.24

$$\cos(\text{S1, S3}) = \frac{3}{2.45*2.24} = \mathbf{0.5477}$$

Hence, the cosine similarity between S1 and S2 is **0.55**