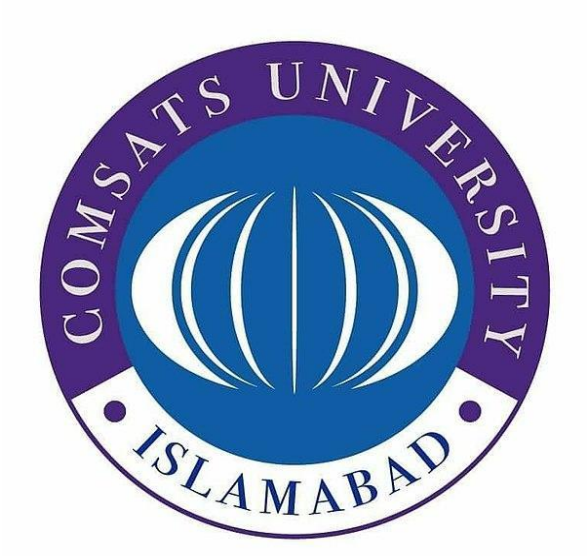

Introductionto Data Sciences

(CSC461)

Submitted by: Ahsan Aijaz

Submitted to: Dr M. Sharjeel

Reg No: SP20-BCS-044

Section: B

Group: G-II

Assignment-5

Q1. 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”

Vocabulary: sunshine, state, enjoy, brown, fox, jump, high, run, fast

Bag of Words

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

Vectors:
S1 : [2 1 1 0 0 0 0 0 0]
S2 : [0 0 0 2 2 1 1 1 0]
S3 : [1 1 0 0 1 0 0 1 1]

Term Frequency:

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

Inverse Document Frequency:

	Idf
Sunshine	0.18
State	0.18
Enjoy	0.48
Brown	0.48
Fox	0.18
Jump	0.48
High	0.48
Run	0.48
Fast	0.48

Term Frequency inverse document frequency:

	S1	S2	S3
Sunshine	0.09	0	0.036
State	0.045	0	0.036
Enjoy	0.12	0	0
Brown	0	0.137	0
Fox	0	0.051	0.036
Jump	0	0.068	0
High	0	0.068	0
Run	0	0.068	0.096
Fast	0	0	0.096

Q2. Compute the cosine similarity between S1 and S3.

Vector S1: [2 1 1 0 0 0 0 0 0]

Vector S3: [1 1 0 0 1 0 0 1 1]

cos(S1, S3) = ($S1$. $S3$) $|S1|$ $|S3|$

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

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

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

cos(S1, S3) = 3 $2.45 * 2.24 = 0.5466$