Name: Muhammmad Umair Tariq **Reg#FA19-BCS-055**

<u>Introduction to Data Science</u>

Assignment#5

Sentences:

- sunshine state enjoy sunshine
- brown fox jump high, brown fox run
- sunshine state fox 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	0	4
S2	0	0	0	2	2	1	1	1	1	0	8
S3	1	1	0	0	1	0	0	0	1	1	5

Term Frequencies:

	Sunshine	State	Enjoy	Brown	Fox	Jump	High	,	Run	fast
S1	1/2	1/4	1/4	0	0	0	0	0	0	0
S2	0	0	0	1/4	1/4	1/8	1/8	1/8	1/8	0
S3	1/5	1/5	0	0	1/5	0	0	0	1/5	1/5

Inverse Document Frequencies:

Idf('sunshine')=log(3/2)=0.1760

Idf('state)=log(3/2)= 0.1760

Idf('enjoy')=log(3/1)=0.4771

Idf('brown')=log(3/1)=0.4771

Idf('fox')=log(3/2)= 0.1760

Idf('jump')=log(3/1)=0.4771

Idf('high')=log(3/1)=0.4771

Idf(',')=log(3/1)=0.4771

Idf('run')=log(3/2)=0.1760

Idf('fast')=log(3/1)=0.4771

	Sunshine	State	Enjoy	Brown	Fox	Jump	High	,	Run	fast
IDF	0.1760	0.1760	0.4771	0.4771	0.1760	0.4771	0.4771	0.4771	0.1760	0.4771

TF-IDF:

S1:

Idf(‘sunshine’)=0.1760x1/2=0.088

Idf(‘state)= 0.1760x1/4=0.044

Idf(‘enjoy’)= 0.4771x1/4=0.044

Idf(‘brown’)=0.4771x0=0

Idf(‘fox’)= 0.1760x0=0

Idf(‘jump’)=0.4771x0=0

Idf(‘high’)=0.4771x0=0

Idf(‘,’)=0.4771x0=0

Idf(‘run’)=0.1760x0=0

Idf(‘fast’)=0.4771x0=0

S2:

Idf(‘sunshine’)=0.1760x0=0

Idf(‘state)= 0.1760x0=0

Idf(‘enjoy’)= 0.4771x0=0

Idf(‘brown’)=0.4771x1/4=0.1192

Idf(‘fox’)= 0.1760x1/4=0.044

Idf(‘jump’)=0.4771x1/8=0.0596

Idf(‘high’)=0.4771x1/8=0.0596

Idf(‘,’)=0.4771x1/8=0.0596

Idf(‘run’)=0.1760x1/8=0.022

Idf(‘fast’)=0.4771x0=0

S3:

Idf(‘sunshine’)=0.1760x1/5=0.0352

Idf(‘state)= 0.1760x1/5=0.0352

Idf(‘enjoy’)= 0.4771x0=0

Idf(‘brown’)=0.4771x0=0

Idf(‘fox’)= 0.1760x1/5=0.0352

Idf(‘jump’)=0.4771x0=0

Idf(‘high’)=0.4771x0=0

Idf(‘,’)=0.4771x0=0

Idf(‘run’)=0.1760x1/5=0.0352

Idf(‘fast’)=0.4771x1/5=0.0954

	Sunshine	State	Enjoy	Brown	Fox	Jump	High	,	Run	fast
Idf(S1)	0.088	0.044	0.044	0	0	0	0	0	0	0
Idf(S2)	0	0	0	0.1192	0.044	0.0596	0.0596	0.0596	0.022	0
Idf(S3)	0.0352	0.0352	0	0	0.0352	0	0	0	0.0352	0.0954

Cosine Similarity Between S1 and S3:

Cos(S1,S3)=S1.S3/|S1||S3|

Taking Bag of Words Vector:

S1=[2,1,1,0,0,0,0,0,0,0]

S3=[1,1,0,0,1,0,0,0,1,1]

S1.S3=2*1 + 1*1 + 1*0 + 0*0 + 0*1 + 0*0 + 0*0 + 0*0 + 0*1 + 0*1=3

|S1|= (2*2 + 1*1 + 1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0)^0.5=2.4494

|S2|= (1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0 + 0*0 + 1*1 + 1*1)^0.5=2.2360

Cos(S1,S3)=3/2.4494*2.2360

Cos(S1,S3)=3/5.4768

Cos(S1,S3)=0.5477