S1: "data science is one of the most important courses in computer sciences"

S2: "this is one of the best data science courses"

S3: "the data scientists perform data analysis"

BOW:

	data	science	is	one	of	the	most	important	courses	in	computer	this	best	scientists
S1	1	2	1	1	1	1	1	1	1	1	1	0	0	0
S2	1	1	1	1	1	1	0	0	1	0	0	1	1	0
S3	2	0	0	0	0	1	0	0	0	0	0	0	0	1

	perform	analysis	Total length
S1	0	0	12
S2	0	0	9
S3	1	1	6

S1 vector: [1 2 1 1 1 1 1 1 1 1 0 0 0 0 0]

S2 vector: [1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 0]

S3 vector: [2 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1]

Term Frequency:

	data	science	is	one	of	the	most	important	courses	in	computer	this	best	scientists	perform	analysis
S1	0.08	0.17	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0	0	0	0	0
S2	0.11	0.11	0.11	0.11	0.11	0.11	0	0	0.11	0	0	0.11	0.11	0	0	0
S3	0.33	0	0	0	0	0.16	0	0	0	0	0	0	0	0.16	0.16	0.16

Inverse Document Frequency:

IDF

data	0
science	0.18
is	0.18
one	0.18
of	0.18
the	0
most	0.48
important	0.48
courses	0.18
in	0.48
computer	0.48
this	0.48
best	0.48
scientists	0.48
perform	0.48
analysis	0.48

TFIDF

	S1	S2	S3
data	0	0	0
science	0.031	0.0198	0
is	0.014	0.0198	0
one	0.014	0.0198	0
of	0.014	0.0198	0
the	0	0	0
most	0.038	0	0
important	0.038	0	0
courses	0.014	0.0198	0
in	0.038	0	0
computer	0.038	0	0
this	0	0.53	0
best	0	0.53	0
scientists	0	0	0.77
perform	0	0	0.77
analysis	0	0	0.77

Cosine Similarity

→ between S1 and S2

$$\cos = \frac{S1 \cdot S2}{|S1||S2|}$$

$S1 \cdot S2 = 0 + 0 + 0 + 0.08 + 0.048 + 0 + 0 + 0.08 + 0 + 0.08 + 0 + 0.16 + 0 + 0.48$

$= 0.496$

$|S1| = \sqrt{(.33)^2 + (.25)^2 + (.19)^2 + (.33)^2 + (.33)^2 + (.25)^2 + (.33)^2 + (.25)^2 + (.25)^2 + (.50)^2 + (.19)^2}$

$|S1| = \sqrt{1.316} = 1.147$

$|S2| = \sqrt{(.42)^2 + (.32)^2 + (.25)^2 + (.32)^2 + (.32)^2 + (.32)^2 + (.32)^2 + (.25)^2 + (.42)^2}$

$|S2| = 0.995$

$\cos = \frac{0.496}{1.147 \times 0.995}$