Through a step-by-step process, calculate TFIDF for the given corpus

Document 1: ***Johny Johny, Yes Papa,***
Document 2: ***Eating sugar? No Papa***
Document 3: ***Telling lies? No Papa***
Document 4: ***Open your mouth, Ha! Ha! Ha!***

Ans:

1. Create document vectors for the given documents (**Term Frequency Table**)

Johny	Yes	Papa	Eating	Sugar	No	Telling	Lies	Open	your	Mouth	Ha
2	1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	0	0	0
0	0	1	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	3

2. Record the occurrence of word in the document using term frequency table (**Document Frequency Table**)

Johny	Yes	Papa	Eating	Sugar	No	Telling	Lies	Open	your	Mouth	Ha
1	1	3	1	1	2	1	1	1	1	1	1

3. Draw the **inverse document frequency table** wherein, we need to put the document frequency in the denominator while the total number of documents is the numerator. Here, the total number of documents are 4, hence inverse document frequency becomes:

Johny	Yes	Papa	Eating	Sugar	No	Telling	Lies	Open	your	Mouth	Ha
4/1	4/1	4/3	4/1	4/1	4/2	4/1	4/1	4/1	4/1	4/1	4/1

4. The formula of **TFIDF** for any word W becomes: TFIDF(W) = TF(W) * log (IDF(W))

Johny	Yes	Papa	Eating	Sugar	No	Telling	Lies	Open	your	Mouth	Ha
2*log(4/1)	1*log(4/1)	1*log(4/3)	0*log(4/1)	0*log(4/1	0*log(4/2)	0*log(4/1	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)
0*log(4/1)	0*log(4/1)	1*log(4/3)	1*log(4/1)	1*log(4/1)	1*log(4/2)	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)
0*log(4/1)	0*log(4/1)	1*log(4/3)	0*log(4/1)	0*log(4/1)	1*log(4/2)	1*log(4/1)	1*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)	0*log(4/1)
0*log(4/1)	0*log(4/1)	0*log(4/3)	0*log(4/1)	0*log(4/1)	0*log(4/2)	0*log(4/1)	0*log(4/1)	1*log(4/1)	1*log(4/1)	1*log(4/1)	3*log(4/1)

(1 mark for each correct table)