<table>
<tr><th></th><th>Doc1</th><th>Doc2</th><th>Doc3</th></tr>
<tr><td>Step 0:

Raw data</td><td>Apple is a fruit which is red in colour.

Apple a day keeps doctor away.</td><td>Orange is a fruit which is orange in colour.

It is rich in vitamins.</td><td>Kiwi is a fruit which is green in colour.

It is a native fruit of new Zealand.</td></tr>
<tr><td>Step 1:

Remove all stop words (like is, a, which, in, it, of, etc)</td><td>Apple fruit red colour apple day keeps doctor away</td><td>Orange fruit orange colour rich vitamins</td><td>Kiwi fruit green colour native fruit new zealand</td></tr>
<tr><td>Step 2:

Lemmatize (i.e find the root word. Eg: root word of “seeing” is “see)</td><td>Keeps => keep

(no change in other words as they are root words already)

Apple fruit red colour apple day keep doctor away</td><td>Vitamins => vitamin

Orange fruit orange colour rich vitamin</td><td>(no change here)

Kiwi fruit green colour native fruit new zealand</td></tr>
<tr><td>Step 3:
Identify the frequency of each word within the document</td><td>
<table>
<tr><td>Apple</td><td>2</td></tr>
<tr><td>Fruit</td><td>1</td></tr>
<tr><td>Red</td><td>1</td></tr>
<tr><td>Colour</td><td>1</td></tr>
<tr><td>day</td><td>1</td></tr>
<tr><td>keep</td><td>1</td></tr>
<tr><td>doctor</td><td>1</td></tr>
<tr><td>away</td><td>1</td></tr>
<tr><td>total</td><td>9</td></tr>
</table>
</td><td>
<table>
<tr><td>Orange</td><td>2</td></tr>
<tr><td>fruit</td><td>1</td></tr>
<tr><td>colour</td><td>1</td></tr>
<tr><td>rich</td><td>1</td></tr>
<tr><td>Vitamin</td><td>1</td></tr>
<tr><td>total</td><td>6</td></tr>
</table>
</td><td>
<table>
<tr><td>kiwi</td><td>1</td></tr>
<tr><td>fruit</td><td>2</td></tr>
<tr><td>green</td><td>1</td></tr>
<tr><td>color</td><td>1</td></tr>
<tr><td>native</td><td>1</td></tr>
<tr><td>new</td><td>1</td></tr>
<tr><td>zealand</td><td>1</td></tr>
<tr><td>total</td><td>8</td></tr>
</table>
</td></tr>
<tr><td>Step 4:
Identify Term Frequency (TF) in each document

TF = (freq of word)/(no. of words in document)

Or

TF = 0.5 + 0.5*(freq of word)/(freq of most repeating word across documents)</td><td colspan="3">
<table>
<tr><td>term</td><td>TF of d1</td><td>TF of d2</td><td>TF of d3</td></tr>
<tr><td>Apple</td><td>0.222222
freq of word=2,
No. of words in document=9,
So TF=2/9=0.222222</td><td>0</td><td>0</td></tr>
<tr><td>Fruit</td><td>0.111111</td><td>0.166667</td><td>0.25</td></tr>
<tr><td>Red</td><td>0.111111</td><td>0</td><td>0</td></tr>
<tr><td>Colour</td><td>0.111111</td><td>0.166667</td><td>0.125</td></tr>
<tr><td>day</td><td>0.111111</td><td>0</td><td>0</td></tr>
<tr><td>keep</td><td>0.111111</td><td>0</td><td>0</td></tr>
<tr><td>doctor</td><td>0.111111</td><td>0</td><td>0</td></tr>
<tr><td>away</td><td>0.111111</td><td>0</td><td>0</td></tr>
<tr><td>orange</td><td>0</td><td>0.333333</td><td>0</td></tr>
<tr><td>rich</td><td>0</td><td>0.166667</td><td>0</td></tr>
<tr><td>vitamin</td><td>0</td><td>0.166667</td><td>0</td></tr>
<tr><td>kiwi</td><td>0</td><td>0</td><td>0.125</td></tr>
<tr><td>green</td><td>0</td><td>0</td><td>0.125</td></tr>
<tr><td>native</td><td>0</td><td>0</td><td>0.125</td></tr>
<tr><td>new</td><td>0</td><td>0</td><td>0.125</td></tr>
<tr><td>zealand</td><td>0</td><td>0</td><td>0.125</td></tr>
</table>
</td></tr>
</table>

Step 5:
Find IDF (inverse document frequency for each word)

IDF = 1+log(n/k)
n = total no. of documents
k = no. of documents containing this word

term	**n/k**	**log(n/k)**	**IDF=1+log(n/k)**
Apple	3	0.477121	1.477121
Fruit	1	0	1
Red	3	0.477121	1.477121
Colour	1	0	1
day	3	0.477121	1.477121
keep	3	0.477121	1.477121
doctor	3	0.477121	1.477121
away	3	0.477121	1.477121
orange	3	0.477121	1.477121
rich	3	0.477121	1.477121
vitamin	3	0.477121	1.477121
kiwi	3	0.477121	1.477121
green	3	0.477121	1.477121
native	3	0.477121	1.477121
new	3	0.477121	1.477121
zealand	3	0.477121	1.477121

Step 6 :
Find TF IDF for each document
TF IDF = TF * IDF

TF values for each word in a document are available in Step 4. IDF value is common for all documents and is available in Step 5.

TF-IDF(d1)	TF-IDF(d2)	TF-IDF(d3)
0.3282492	0	0
0.1111111	0.018519	0.027778
0.1641246	0	0
0.1111111	0.018519	0.125
0.1641246	0	0
0.1641246	0	0
0.1641246	0	0
0.1641246	0	0
0	0.492374	0
0	0.246187	0
0	0.246187	0
0	0	0.18464
0	0	0.18464
0	0	0.18464
0	0	0.18464
0	0	0.18464
0	0	0.18464

Step 7:
Find Cosine similarity between two documents

Cosine Similarity = (D1.D2)/(sqrt of ||D1|| * sqrt of ||D2||)

Where

||D1|| = sum of squares of the TF IDF values of D1
||D2|| = sum of squares of the TF IDF values of D2

d1.d2	squares of d1	sqrt(squares sum of d1)	squares of d2	sqrt(squares sum of d1)	|d1|.|d2|	Cosine Sim
0	0.107747516	0.516839694	0	1.429534568	0.73884	0.0055698
0.002057613	0.012345679		0.037037037			
0	0.026936879		0			
0.002057613	0.012345679		0.037037037			
0	0.026936879		0			
0	0.026936879		0			
0	0.026936879		0			
0	0.026936879		0			
0	0		0.984747503			
0	0		0.492373752			
0	0		0.492373752			
0	0		0			
0	0		0			
0	0		0			
0	0		0			
0	0		0			

Similarly find Cosine Similarity for D1 and D3

d1.d3	squares of d1	sqrt(squares sum of d1)	squares of d3	sqrt(squares sum of d3)	|d1|.|d3|	Cosine Sim
0	0.107747516	0.516839694	0	1.466955052	0.758181	0.0223895
0.00308642	0.012345679		0.055555556			
0	0.026936879		0			
0.013888889	0.012345679		0.25			
0	0.026936879		0			
0	0.026936879		0			
0	0.026936879		0			
0	0.026936879		0			
0	0		0			
0	0		0			
0	0		0			
0	0		0.369280314			
0	0		0.369280314			
0	0		0.369280314			
0	0		0.369280314			
0	0		0.369280314			
0.016975309			2.151957124			

And find Cosine Similarity for D2 and D3

d2.d3	squares of d2	sqrt(squares sum of d2)	squares of d3	sqrt(squares sum of d3)	|d2|.|d3|	Cosine Sim

Step 8: **Identify most similar documents based on their cosine similarity values.**

Finally compare CosSim(d1, d2), CosSim(d1, d3) and CosSim(d2, d3).
The 2 documents which have the highest CosSim value are the most similar documents.