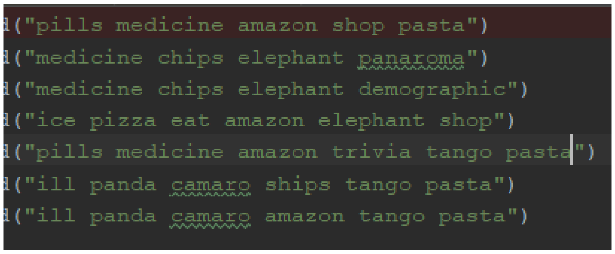
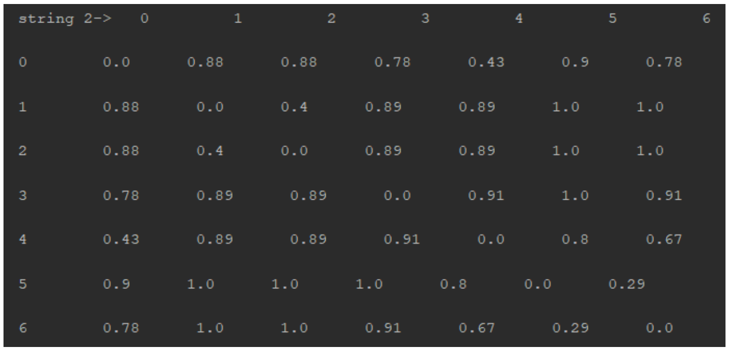
Q1) entities chosen are



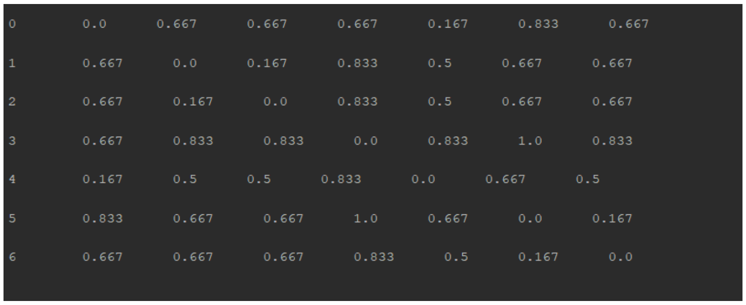
```
d("pills medicine amazon shop pasta")
d("medicine chips elephant panaroma")
d("medicine chips elephant demographic")
d("ice pizza eat amazon elephant shop")
d("pills medicine amazon trivia tango pasta")
d("ill panda camaro ships tango pasta")
d("ill panda camaro amazon tango pasta")
```

Applying Jaccard distance on the entities we get



```
string 2->   0          1          2          3          4          5          6

0         0.0       0.88       0.88       0.78       0.43       0.9        0.78

1         0.88       0.0       0.4        0.89       0.89       1.0        1.0

2         0.88       0.4       0.0        0.89       0.89       1.0        1.0

3         0.78       0.89       0.89       0.0        0.91       1.0        0.91

4         0.43       0.89       0.89       0.91       0.0        0.8        0.67

5         0.9       1.0       1.0       1.0       0.8       0.0       0.29

6         0.78       1.0       1.0       0.91       0.67       0.29       0.0
```

Triangle inequality states that sum of distances of two sides of triangle are greater than the third sides.We can see from the data that the max length for third side can be only one.The distance between two strings is max when the string in the middle is almost in the middle of the other two strings. This can be empirically seen in the data above
For example points
0,1,2 [0,2]=0.88 [0,1]=0.88 [1,2]=0.4
0.88<1.24
4,5,6 [4,6]=0.67 [4,5]=0.8 [5,6]=0.29
0.8<0.96
Hence jaccard follows triangle inequality
b)Dice coefficient



```
0        0.0       0.667      0.667      0.667      0.167      0.833      0.667

1        0.667      0.0       0.167      0.833      0.5       0.667      0.667

2        0.667      0.167      0.0       0.833      0.5       0.667      0.667

3        0.667      0.833      0.833      0.0       0.833      1.0       0.833

4        0.167      0.5       0.5       0.833      0.0       0.667      0.5

5        0.833      0.667      0.667      1.0       0.667      0.0       0.167

6        0.667      0.667      0.667      0.833      0.5       0.167      0.0
```

Triangle inequality doesn’t hold for the dice coefficient. See the points 0,1,4
[0,4]=0.16 [4,1]=0.5 [0,1]=0.67
0.67>0.66
Hence inequality is not satisfied in Dice method.

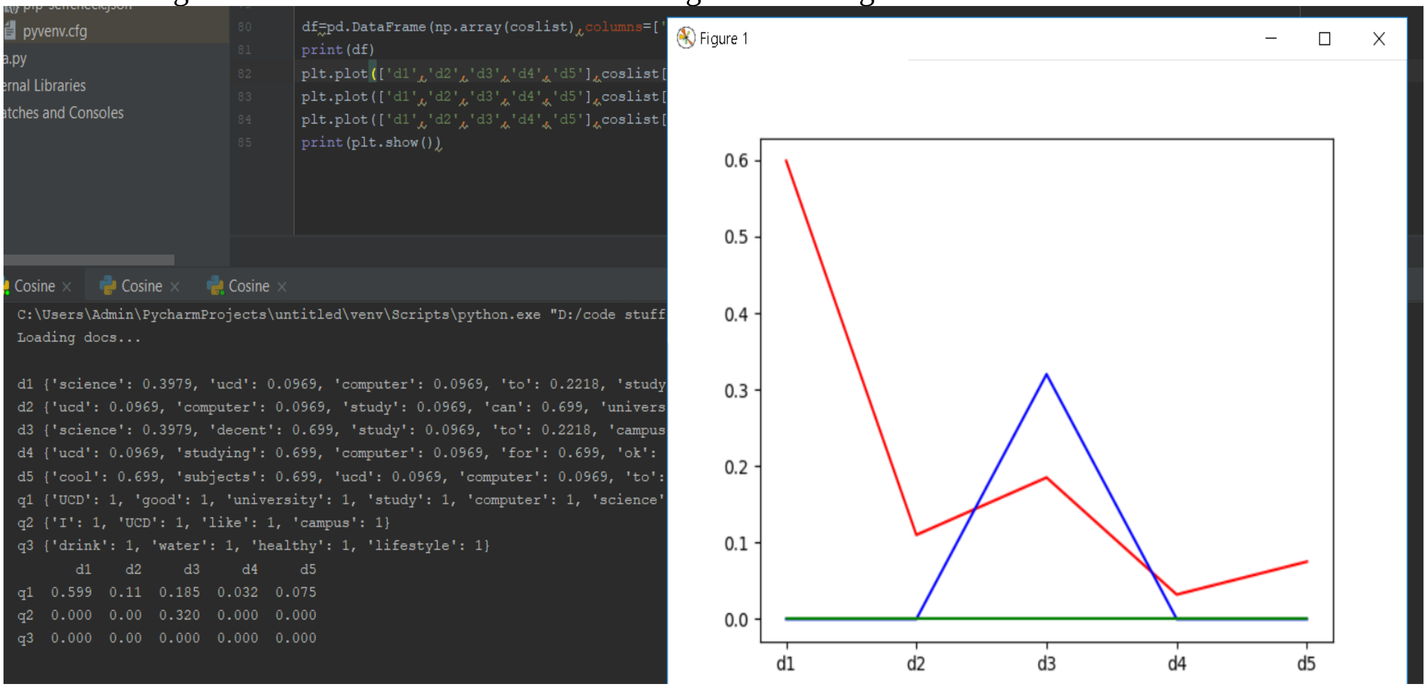
Q2)
Text chosen:

```
vector_dict['q1'] = {'UCD' : 1, 'good' : 1, 'university' : 1, 'study' : 1, 'computer' : 1, 'science' : 1}
vector_dict['q2'] ={'I':1,'UCD':1,'like':1,'campus':1}
vector_dict['q3']={'drink':1,'water':1,'healthy':1,'lifestyle':1}
```

variations used:

```
doc1=('d1', 'UCD is a good university to study computer science')
doc2=('d2', 'UCD university can study computer ')
doc3=('d3', ' decent university  to study campus science')
doc4=('d4', 'UCD is a ok university for studying computer ')
doc5=('d5', 'UCD is a cool university to study computer subjects')
```

After scanning differences between variations and original text we get



```
df=pd.DataFrame(np.array(coslist),columns=['
print(df)
plt.plot(['d1','d2','d3','d4','d5'],coslist[
plt.plot(['d1','d2','d3','d4','d5'],coslist[
plt.plot(['d1','d2','d3','d4','d5'],coslist[
print(plt.show())
```

```
C:\Users\Admin\PycharmProjects\untitled\venv\Scripts\python.exe "D:/code stuff
Loading docs...

d1 {'science': 0.3979, 'ucd': 0.0969, 'computer': 0.0969, 'to': 0.2218, 'study
d2 {'ucd': 0.0969, 'computer': 0.0969, 'study': 0.0969, 'can': 0.699, 'univers
d3 {'science': 0.3979, 'decent': 0.699, 'study': 0.0969, 'to': 0.2218, 'campus
d4 {'ucd': 0.0969, 'studying': 0.699, 'computer': 0.0969, 'for': 0.699, 'ok':
d5 {'cool': 0.699, 'subjects': 0.699, 'ucd': 0.0969, 'computer': 0.0969, 'to':
q1 {'UCD': 1, 'good': 1, 'university': 1, 'study': 1, 'computer': 1, 'science'
q2 {'I': 1, 'UCD': 1, 'like': 1, 'campus': 1}
q3 {'drink': 1, 'water': 1, 'healthy': 1, 'lifestyle': 1}
       d1     d2     d3     d4     d5
q1  0.599  0.11  0.185  0.032  0.075
q2  0.000  0.00  0.320  0.000  0.000
q3  0.000  0.00  0.000  0.000  0.000
```

Pictured to bottom left is the matrix.
Red denotes difference between first text and all the variation documents. As expected its highest since all are variations of first text.
Blue is difference of second text and variation documents with one common word.
Green is a completely different text and as expected the cosine is zero.

c)Testing (previous documents)similarity with Euclidian distance results in

	Euclidian	Cosine
doc1	0.22860036	0.365
doc 2	0.1308152	0.224
doc 3	0.	0.

```
New doc chosen is  'UCD university are important for nations
economy'
```

Euclidian is understandably lower than cosine. This is because it measures straight line distance to the documents instead of measuring angles.

Q3

Tweets chosen

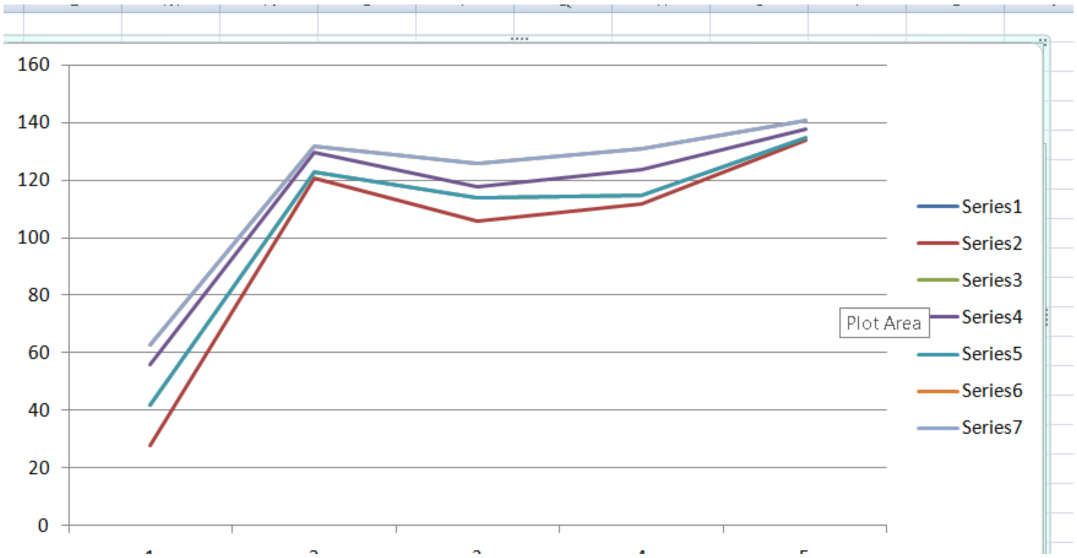
```
s1 = 'Former Rangers defender David Bates eyed by Arsenal and Everton as well as other Premier League clubs'
s2 = 'We will always be grateful to them for their courage. Today, I had the honour of meeting Lalti Ram Ji, an
INA veteran. It was wonderful spending time with him.'

s3 = 'Their success will inspire many other youngsters to shine on the playing field. Wishing these young stars
the very best for their future endeavours.'
s4 = 'Its amazing what a simple thermal paste change can do to an 11 month old laptop that was pretty clean.
Sorry for it being a photo and not a screenshot.'

s5 = 'Mobile Phones is one of the dominant sub-sector in the electronics industry. It witnessed a jump of 60 %
as manufacturing of mobile phones reached 175 Mn units during 2016-17'
```

Comparisons in matrix form of tweets and spam generated from first tweet

```
T1   t2   t3   t4  t5
  28 121 106 112 134  s1
  42 123 114 115 135  s2
  56 130 118 124 138  s3
  42 123 114 115 135  s4
  63 132 126 131 141
  63 132 126 131 141
  56 130 118 122 142
  56 130 123 121 141
  49 124 117 123 137
  21 118 108 114 133
  21 118 108 114 133
  35 122 110 117 135
  56 130 118 124 138
  56 130 123 121 141
  42 126 116 119 138
  42 126 116 119 138
  63 132 126 131 141
  42 123 114 115 135
  42 123 114 115 135
  56 130 118 122 142  s20
```

As expected for the first tweet the distance is lowest because spam is generated from it

Plotting for the first six values of spam and text comparisons we get.



We can see that the overall trend is same for all spam. For first comparison(of variations and first tweet) we get a small value as they are variations of same tweet. And otherwise all values are high. This implies that spam is usually very different to actual unique tweets .

If we plot graphs between different text we will get completely random trends since the tweets are quite unique and all the distance values will tend to be random