Q1: Jaccard

a. Jaccard-Distance

I have selected five sentences from the internet that introduce Tesla.

```
target1 = "Tesla is an US company involved in car manufacturing and energ
y."
target2 = "Tesla is an American multinational automotive and clean energy
company."
target3 = "Tesla is building a world powered by solar energy, running on b
atteries and transported by electric vehicles."
target4 = "Tesla is an American manufacturer of electric automobiles, sola
r panels, and batteries for cars and home power storage."
target5 = "Tesla is an electric vehicle manufacturer and clean energy comp
any"
```

Then I completed the normalization of them. Lowercase the text, remove punctuations and co nvert numbers to English words. Modify the Jaccard-Index program, and the Jaccard-Distanc e between them is calculated as follows.

AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
0.73	0.88	0.94	0.73	0.88	0.88	0.6	0.79	0.8	0.81

Table 1: Jaccard-Distance

Check property of triangle inequality:

Obviously, AB + BC > AC, AC + CD > AD

Triangle inequality stays true.

b. Dice Coefficient

$$J = \frac{S}{(2 - S)}$$

$$S = \frac{2J}{(1 + j)}$$

By the derivation of the above two formulas, the value of S (Dice Coefficient) can be obtaine d calculated as follows.

AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
0.43	0.21	0.11	0.43	0.21	0.21	0.57	0.35	0.33	0.32

Table 2: Dice Coefficient

Check property of triangle inequality:

BC + CE < BE

Triangle inequality stays false.

These two measures provide different results in terms of analyzing the similarity between the sets of texts. However, the results of Dice Coefficient show a higher similarity rate between t he above-mentioned sets, while Jaccard shows a lower one. However, in measuring the distan ce, Dice Coefficient does not obey the triangle inequality and is not an appropriate distance-measuring tool compared to Jaccard-Distance.

Q2: Cosine Similarity

a: Cosine Similarity

Recently, Ireland is about to have a satellite. So I selected three sentences from the news from three articles about launching a satellite in Ireland. Moreover, I also chose three sentences from three news about launching a satellite in other countries.

```
doc1=('d1', 'IRELAND IS set to launch its first satellite next year after Tánaiste Leo Varadkar today signed an Exchange of Letters with the Europea n Space Agency (ESA).')

doc2=('d2', 'Ireland preparing for first space satellite mission a Leo Var adkar signs off on major milestone with untold opportunities')

doc3=('d3', 'Tanaiste Leo Varadkar confirmed yesterday that Ireland would launch its first- space satellite mission in the new year.')

doc4=('d4', 'China launches Yunhai-1 03 Earth-observing satellite into orb it.')

doc5=('d5', 'A classified satellite for the U.S. National Reconnaissance O ffice was launched into space from California on Sunday.')

doc6=('d6', 'Next batch of OneWeb satellites arrive in India for launch.')
```

The results are calculated by Cosine.py as follows.

d1→d2	d1→d3	d1→d4	d1→d5	d1→d6	d2→d3	d2→d4	d2→d5
0.103	0.111	0.001	0.016	0.036	0.094	0.001	0.058

d2→d6	d3→d4	d3→d5	d3→d6	d4→d5	d4→d6	d5→d6
0.019	0.001	0.02	0.0	0.046	0.0	0.018

Table 3: Cosine similarity in groups

Select d1 as the base document then draw its scatter plot with the values of other documents.

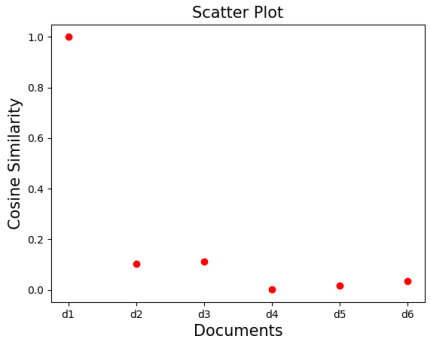

Figure 1: Cosine.py

d1 has a higher similarity with d2 and d3. Largely, they are derived from reports of satellites launchin g in Ireland. Moreover, in the selected clips, there are several sharing parts, the name of the vice prem ier and “launch satellite”, “Ireland”. The similarity between d1 and d4, d5, and d6 is lower probably b ecause they only share the keyword “launch satellite” or “satellite”.

b: Manhattan Distance

Select d1 as the base document then draw its scatter plot with the values of other documents using two methods in sklearn combined with TF-IDF method.

d1→d1	d1→d2	d1→d3	d1→d4	d1→d5	d1→d6
1	0.217	0.325	0.228	0.0745	0.102

Table 4: sklearn.metrics.pairwise.cosine_similarity()

d1→d1	d1→d2	d1→d3	d1→d4	d1→d5	d1→d6
0	6.693	5.812	7.469	7.768	7.025

Table 5: sklearn.metrics.pairwise.manhattan_distances()

It is not the same. Compared to the previous results, the value of Cosine Similarity calculated by calli ng the sklearn method is larger than that calculated by calling Cosine.py. This difference is reflected i n the reason for the different treatment of the two methods when calculating the TF-IDF values.

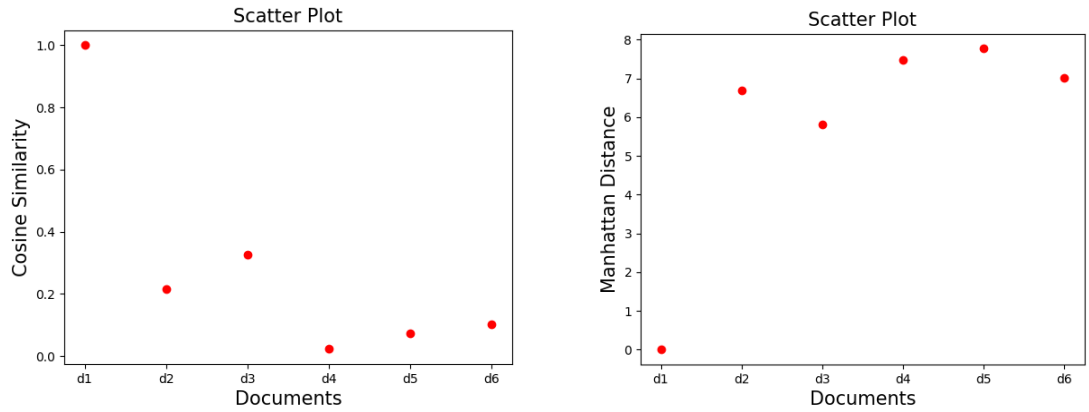

Figure 2: CS and MD

A low Manhattan Distance score means that the two documents have a high similarity, and a high scor e means that the two documents have low similarity. In the documents, the two metrics present the sa me similarity results. Looking through Figure 2, the Manhattan Distance plot looks like a flip of the C osine Similarity plot.