**Name: Nikhil Jadhav**

**Student ID: 801075504**

**Cloud Computing for Data Analysis**

**VIDEO CASE 05 : Finding Similar Items**

Watch following videos:

**Video 1:** <https://youtu.be/wrkVnwaKTjo>

**Video 2:** <https://youtu.be/ubqGFxHeg7Q>

Video 1 gives a sample example for Jaccard coefficient and its limitations

Video 2 gives a sample example for Cosine similarity

**Video Case Questions:**

**D1:** The sky is blue

**D2:** The sun in bright

**Query:** The sun in the sky is bright

1. Find Jaccard coefficient for the above documents (D1 and D2) for the query Q
2. What is the advantage of using cosine similarity over Jaccard coefficient?
3. Where do you think, these measures can be used?

Answers:

1. The given document and query are as follows:
   1. D1: The sky is blue
   2. D2: The sun in bright
   3. Q: The sun in the sky is bright

The Jaccard coefficient for D1 is:

(Q, D1) = 3/7

(Q, D2) = 4/6 = 2/3

1. a. Jacard coefficient is used to measure overlap it does not consider frequency.

b. Jacard coefficient requires sophisticated way of normalizing length.

c. In cosine similarity a document can be represented by a bag of terms or a long vector, with each attribute recording the frequency of a particular term.

d. Consine similarity is used for applications involving information retrieval, biological taxonomy, gene feature mapping etc.

1. The Cosine index could be used to identify plagiarism but will not be a good index to identify mirror sites on the internet. Whereas the Jaccard index, will be a good index to identify mirror sites, but not so great at catching copy pasta plagiarism (within a larger document).