**Cloud Computing for Data Analysis**

**VIDEO CASE 05 : Finding Similar Items**

**Student Name: Balasundaram Avudai Nayagam**

**Student ID: 801104903**

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. Jaccard coefficient (A, B) is given by (A ∩ B)/ (A ∪ B)

D1 ∩ Q = {the, sky, is}

D1 ∪ Q = {the, sky, is, blue, sun, in, bright}

D2 ∩ Q = {the, sun, in, bright}

D2 ∪ Q = {the, sun, in, bright, sky, is}

J (D1, Q) = 3/7

J (D2, Q) = 4/6

1. What is the advantage of using cosine similarity over Jaccard coefficient?
2. - Less complex than Jaccard coefficient

* Cosine Similarity considers term frequency whereas Jaccard coefficient does not.
* Cosine Similarity gives better length normalization since it divides intersection with the square root of the union of the items in the collection.

1. Where do you think, these measures can be used?
2. - They are used in applications such as Information retrieval, biological taxonomy and gene feature mapping.

* The Jaccard coefficient is used to measure dissimilarity between two collections.
* It is also used to calculate similarities in gene features in micro-arrays.