**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

|  |
| --- |
| Intersection: The.sky.is  Union: The.sun.in.sky.is.bright.blue  J(Q, D1) = 3/7  Intersection: The.sun.in.bright  Union: The.sun.in.the.sky.is.bright  J(Q, D2) = 4/6 |

1. What is the advantage of using cosine similarity over Jaccard coefficient?

|  |
| --- |
| Jaccard coefficient does not consider term frequency. Also, compared with cosine similarity, Jaccard coefficient provide a less sophisticated way of normalizing for length. |

1. Where do you think, these measures can be used?

|  |
| --- |
| Jaccard coefficient and cosine similarity can be used in applications for information retrieval, biologic taxonomy, gene feature mapping, plagiarism check, etc. |