**Cloud Computing for Data Analysis**

**VIDEO CASE 05: Finding Similar Items**

**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  
   **Answer:**J(q, d1) = 3/8

J(q,d2) = 4/7

1. What is the advantage of using cosine similarity over Jaccard coefficient?  
   **Answer:**Jaccard similarity does not take term frequency into consideration as compared to cosine frequency which does. To elaborate, the term frequency must be considered as, the words which do not appear frequently but are important- Jaccard similarity does not consider this. Hence, Cosine similarity is comparatively much more accurate. Furthermore, One could make use of cosine similarity to identify and copy paste plagiarism, however, Jaccard coefficient cannot do this.
2. Where do you think, these measures can be used?   
   **Answer:**The measures of similarity could be used to train a model for document classification into classes. These could also be used for information retrieval. Thus, saving a lot of time and manual labor. In terms of plagiarism, one could use these measures on internet web-sites on the to identify and test credibility.