

DEERWALK INSTITUTE OF TECHNOLOGY



LAB6: The Jaccard Similarity and Cosine Similarity:
(Artificial Intelligence)

SUBMITTED BY:

NAME: Kriston Pal

ROLL NO.: 0511

SECTION: A

DATE: 12th May, 2018

SUBMITTED TO:

Birodh Rijal

KATHMANDU, NEPAL

1.2.1 Part A

The Jaccard similarity is defined

$$JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

More notation, given a set A, the cardinality of A denoted $|A|$ counts how many elements are in A. The intersection between two sets A and B is denoted $A \cap B$ and reveals all items which are in both sets. The union between two sets A and B is denoted $A \cup B$ and reveals all items which are in either set. Now the Jaccard similarity is as follows:

$$JS(d1, d2) = \frac{3}{3} = 1 \quad JS(d1, d3) = \frac{1}{10} = 0.1$$

Find the Jaccard similarity of each of the above documents to all other documents.

Answer:

Neglecting the stop words and calculate the Jaccard similarity, we get the following results:

PART A: Finding the Jaccard similarity of each of the above documents to all other documents.

```
Intersection Set: {'sam'}
Union Set: {'sam'}
Jaccard Similarity between Document1 and Document 2 is 1.0

Intersection Set: set()
Union Set: {'green', 'sam', 'like', 'eggs', 'ham'}
Jaccard Similarity between Document1 and Document 3 is 0.0

Intersection Set: {'sam'}
Union Set: {'sam', 'like'}
Jaccard Similarity between Document1 and Document 4 is 0.5

Intersection Set: set()
Union Set: {'green', 'sam', 'like', 'eggs', 'ham'}
Jaccard Similarity between Document2 and Document 3 is 0.0

Intersection Set: {'sam'}
Union Set: {'sam', 'like'}
Jaccard Similarity between Document2 and Document 4 is 0.5

Intersection Set: {'like'}
Union Set: {'green', 'sam', 'eggs', 'like', 'ham'}
Jaccard Similarity between Document3 and Document 4 is 0.2
```

1.2.2. Part B

More frequent terms in a document are more important, i.e. more indicative of the topic.

f_{ij} = frequency of term i in document j

Terms that appear in many different documents are less indicative of overall topic df_i = document frequency of term i that is, number of documents containing term

idf_i = inverse document frequency of term i ,

$$= \log_2 (N / df_i) \quad (N: \text{total number of documents})$$

- A typical combined term importance indicator is tf-idf weighting:

$$= tf_{ij} idf_i = tf_{ij} \log_2 (N / df_i)$$

- A similarity measure is a function that computes the degree of similarity between two vectors.
- Using a similarity measure between the query and each document:
 - It is possible to rank the retrieved documents in the order of presumed relevance.
 - It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.

$$\text{CosSim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \cdot |\vec{q}|} = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^t w_{ij}^2 \cdot \sum_{i=1}^t w_{iq}^2}}$$

Calculate the Cosine similarity of the above documents.

Answer:

Neglecting the stop words and calculating the Cosine similarity, we get the following result:

PART B: Finding the Cosine similarity of each of the above documents to all other documents.

Cosine Similarity between Document1 and Document2 is 1.0

Cosine Similarity between Document1 and Document3 is 0.0

Cosine Similarity between Document1 and Document4 is 0.38333288898839096

Cosine Similarity between Document2 and Document3 is 0.0

Cosine Similarity between Document2 and Document4 is 0.38333288898839096

Cosine Similarity between Document3 and Document4 is 0.25616339380286374

Conclusion:

It can be concluded that, Jaccard similarity and Cosine similarity of documents can be calculated and compared and more accurate result is given by Cosine similarity.