deerwalk
DWIT College

Lab Report 6

Submitted by:

Siddhartha Giri

0534

2019 ‘A’

Submitted to:

Birodh Rijal

(Artificial Intelligence Lecturer)

1.2.1 Part A

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

Answer:

The Jaccard similarity is defined as,

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

Where, A and B are two documents, $A \cap B$ denote the total number of unique words on both of them and $A \cup B$ denotes the number of common words both the documents have.

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

For Document 1 and 2:
Intersection:  {'am', 'sam', 'i'}
Union:  {'am', 'i', 'sam'}
Jaccard similarity:  1.0

For Document 1 and 3:
Intersection:  {'i'}
Union:  {'eggs', 'not', 'and', 'ham', 'do', 'sam', 'i', 'am', 'green', 'like'}
Jaccard similarity:  0.1

For Document 1 and 4:
Intersection:  {'am', 'sam', 'i'}
Union:  {'not', 'do', 'sam', 'i', 'them', 'am', 'like'}
Jaccard similarity:  0.42857142857142855

For Document 2 and 3:
Intersection:  {'i'}
Union:  {'eggs', 'not', 'and', 'ham', 'do', 'sam', 'i', 'am', 'green', 'like'}
Jaccard similarity:  0.1

For Document 2 and 4:
Intersection:  {'am', 'sam', 'i'}
Union:  {'not', 'do', 'sam', 'i', 'them', 'am', 'like'}
Jaccard similarity:  0.42857142857142855

For Document 3 and 4:
Intersection:  {'not', 'do', 'like', 'i'}
Union:  {'not', 'and', 'ham', 'green', 'do', 'sam', 'i', 'them', 'am', 'eggs', 'like'}
Jaccard similarity:  0.36363636363636365
```

1.2.2 Part B

Calculate the Cosine similarity of the above documents.

Answer:

Cosine similarity is calculated as,

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

Where $\overrightarrow{d_j}$ is a document vector which is calculated by the weights of all the words in both the documents with respect to document j. It is computed as

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

where,

$tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

f_{ij} is the frequency of ith word in jth document.

df_i = document frequency of term *i*

```
For Document 1 and 2
Cosine Similarity:  1.0

For Document 1 and 3
Cosine Similarity:  0.0

For Document 1 and 4
Cosine Similarity:  0.21658124988136848

For Document 2 and 3
Cosine Similarity:  0.0

For Document 2 and 4
Cosine Similarity:  0.21658124988136848

For Document 3 and 4
Cosine Similarity:  0.25395862933166535

Process finished with exit code 0
```