DEERWALK INSTITUTE OF TECHNOLOGY

LAB 6: Jaccard and Cosine Similarity

(ARTIFICIAL INTELLIGENCE)

SUBMITTED BY:

NAME: SUSHIL AWALE

PROGRAM: B.SC.CSIT (FIFTH SEM)

ROLL NO.: 0540

SECTION: A

SUBMITTED TO:

BIRODH RIJAL

DATE: 12 MAY 2018

KATHMANDU, NEPAL

2018

INTRODUCTION

Four documents were taken and their similarity was calculated.

JACCARD SIMILARITY

The Jaccard similarity is defined

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

The intersection between two sets A and B is denoted A ∩ B and reveals all items which are in both sets. The union between two sets A and B is denoted A ∪ B and reveals all items which are in either set.

COSINE SIMILARITY

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,

= log2 (N/ df i) (N: total number of documents)

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

$= tf_{ij}\ idf_i = tf_{ij}\ log2\ (N/\ df_i)$

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.

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

OUTPUT

```
"C:\Program Files\JetBrains\IntelliJ IDEA 2017.2.4\bin\runnerw.exe" "
Document 1: i,am,sam
Document 2: sam,i,am
Document 3: i,do,not,like,green,eggs,and,ham
Document 4: i,do,not,like,them,sam,i,am

 Jaccard similarity of Document 1 and 2

Union Set: i,am,sam
Intersection Set: sam,i,am
Jaccard Similarity between document 1 and 2 is 1

 Jaccard similarity of Document 1 and 3

Union Set: i,am,sam,do,not,like,green,eggs,and,ham
Intersection Set: i
Jaccard Similarity between document 1 and 3 is 0.1

 Jaccard similarity of Document 1 and 4

Union Set: i,am,sam,do,not,like,them
Intersection Set: i,sam,am
Jaccard Similarity between document 1 and 4 is 0.42857142857142855
```

```
 Jaccard similarity of Document 2 and 3

Union Set: sam,i,am,do,not,like,green,eggs,and,ham
Intersection Set: i
Jaccard Similarity between document 2 and 3 is 0.1

 Jaccard similarity of Document 2 and 4

Union Set: sam,i,am,do,not,like,them
Intersection Set: i,sam,am
Jaccard Similarity between document 2 and 4 is 0.42857142857142855

 Jaccard similarity of Document 3 and 4

Union Set: i,do,not,like,green,eggs,and,ham,them,sam,am
Intersection Set: i,do,not,like
Jaccard Similarity between document 3 and 4 is 0.36363636363636365
```

```
 Cosine Simalrity of Document 1 and 2

1

 Cosine Simalrity of Document 1 and 3

NaN

 Cosine Simalrity of Document 1 and 4

1

 Cosine Simalrity of Document 2 and 3

NaN

 Cosine Simalrity of Document 2 and 4

1

 Cosine Simalrity of Document 3 and 4

0.39735970711951313
```

CONCLUSION

Using Jaccard Similiarity, the result obtained was not optimal because the frequency of words in a document were ignored. Cosine similarity derived desired output.