deerwalk

DWIT College

Lab Report 6: Document Similarity

Submitted by:	**Submitted to:**
Abhishek Kadariya	Birodh Rijal
0501	(Artificial Intelligence Lecturer)
2019 ‘A’	

Four different text documents:

D1: I am Sam.

D2: Sam I am.

D3: I do not like green eggs and ham.

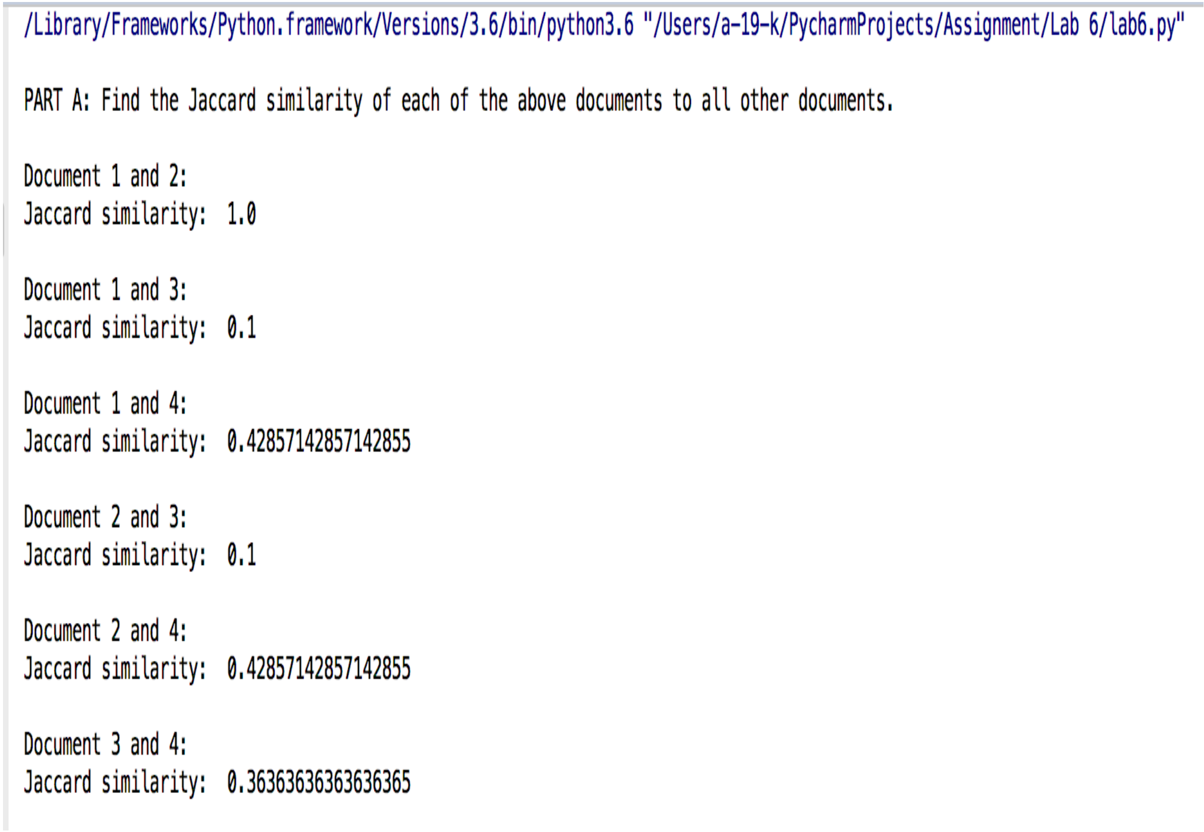
D4: I do not like them, Sam I am.

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

The Jaccard similarity is a common index for binary variables. It is defined as the quotient between the intersection and the union of the pairwise compared variables among two objects. It is defined as,

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

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 ∩ 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.



```
/Library/Frameworks/Python.framework/Versions/3.6/bin/python3.6 "/Users/a-19-k/PycharmProjects/Assignment/Lab 6/lab6.py"

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

Document 1 and 2:
Jaccard similarity:  1.0

Document 1 and 3:
Jaccard similarity:  0.1

Document 1 and 4:
Jaccard similarity:  0.42857142857142855

Document 2 and 3:
Jaccard similarity:  0.1

Document 2 and 4:
Jaccard similarity:  0.42857142857142855

Document 3 and 4:
Jaccard similarity:  0.36363636363636365
```

Part B: Calculate the Cosine similarity of the above documents.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle in the interval [0,2π]. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Cosine similarity is calculated as,

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

Where $\vec{d_j}$ 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

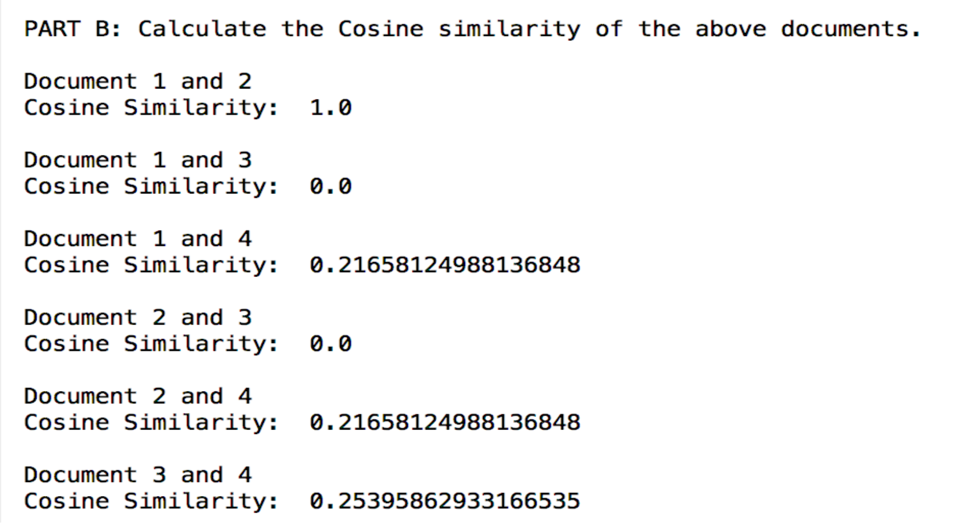
$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$

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



```
PART B: Calculate the Cosine similarity of the above documents.

Document 1 and 2
Cosine Similarity:  1.0

Document 1 and 3
Cosine Similarity:  0.0

Document 1 and 4
Cosine Similarity:  0.21658124988136848

Document 2 and 3
Cosine Similarity:  0.0

Document 2 and 4
Cosine Similarity:  0.21658124988136848

Document 3 and 4
Cosine Similarity:  0.25395862933166535
```