Lab Report 6

Submitted by:

Iris Raj Pokharel

Roll No:510 ‘A’

Submitted to:

Birodh Rijal

(Instructor)

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 ∩ 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. Now the Jaccard similarity is as follows:

JS(d1,d2) = 3/3 =1 JS(d1/d3)= 1/10 =0.1

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

Answer:

If we do not neglect the stop words like (is,am, I, …..) and calculate the similarity the result is as follows:

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


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

Intersection Set:  {'i'}
Union Set:  {'sam', 'ham', 'green', 'eggs', 'am', 'i', 'not', 'like', 'do', 'and'}
Jaccard Similarity between Document1 and Document 3 is  0.1

Intersection Set:  {'i', 'sam', 'am'}
Union Set:  {'sam', 'them', 'am', 'i', 'not', 'like', 'do'}
Jaccard Similarity between Document1 and Document 4 is  0.42857142857142855

Intersection Set:  {'i'}
Union Set:  {'sam', 'ham', 'green', 'eggs', 'am', 'i', 'not', 'like', 'do', 'and'}
Jaccard Similarity between Document2 and Document 3 is  0.1

Intersection Set:  {'i', 'sam', 'am'}
Union Set:  {'sam', 'them', 'am', 'i', 'not', 'like', 'do'}
Jaccard Similarity between Document2 and Document 4 is  0.42857142857142855

Intersection Set:  {'i', 'like', 'do', 'not'}
Union Set:  {'sam', 'ham', 'green', 'eggs', 'them', 'am', 'i', 'not', 'like', 'do', 'and'}
Jaccard Similarity between Document3 and Document 4 is  0.36363636363636365
```

But, if we neglect 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.
fij = 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
idfi = 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:

= tfij idfi = tfij log2 (N/ dfi)

• 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.

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

Calculate the Cosine similarity of the above documents.

Answer:

If we do not neglect the stop words and calculate 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  0.21658124988136848

Cosine Similarity between Document2 and Document3 is  0.0

Cosine Similarity between Document2 and Document4 is  0.21658124988136848

Cosine Similarity between Document3 and Document4 is  0.25395862933166535
```

But, if we neglect the stop words and calculate 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:

Hence, from the above results we can conclude that the output result differ on whether the stops words have been neglected or not. More accurate result will be obtained after neglecting the stop words as they are common in all documents which might not be similar also. Similarly, we can

observe the difference result produced by different methods. More accurate result is given by Cosine similarity which also consider the word frequency and rare terms are given the highest weight.