NLP Assignment-2

9206
Oswin
Comps-A

Q.1)

→ a) $TF(w,d) = \frac{\text{Occurences of } w \text{ in } d}{\text{Total no. of } w \text{ in } d}$

Words	TF(D_1)	TF(D_2)	TF(D_3)	TF(D_4)	TF(D_5)
Information	1/3	1/2	0	0	0
Retrieval	1/3	0	0	0	1/2
System	1/3	0	1/4	0	0
Storage	0	1/2	0	0	0
Digital	0	0	1/4	0	0
Speech	0	0	1/4	1/2	0
Synthesis	0	0	1/4	0	0
Filtering	0	0	0	1/2	0

b) $IDF_{(w,C)} = \log\left(\frac{\text{Total no. of doc in } C}{\text{No. of doc containing } w}\right)$

Words	IDF
Information	$\ln(5/2) = 0.91$
Retrieval	$\ln(5/2) = 0.91$
System	$\ln(5/2) = 0.91$
Storage	$\ln(5/1) = 1.60$
Digital	$\ln(5/1) = 1.60$
Speech	$\ln(5/3) = 0.51$
Synthesis	$\ln(5/1) = 1.60$
Filtering	$\ln(5/1) = 1.60$

c) TF - IDF (Speech) = TF (Speech) x IDF (Speech)

Document	TF(Speech)	IDF(Speech)	TF-IDF
D_1	0	0.51	0
D_2	0	0.51	0
D_3	1/4	0.51	0.1275
D_4	1/2	0.51	0.255
D_5	1/2	0.51	0.255

The TF-IDF value indicates the relevance of a word in each document. Here, the word 'speech' is most relevant in document D_4, D_5 & D_3.

Q 2)

→ ① Synonym :- Two or more words that have the same sense or meaning ~~is kn~~ are known as synonyms.

Example :- Small, little, tiny
large, Huge, Massive.

② Antonyms :- Two or more words that have exactly opposite meaning are known as Antonyms.

Example :- Happy x Sad
Small x large

③ Hyponym :- A term which is an instance of a more generic term is known as Hyponym.

Example :- Car is a hyponym of vehicle
Dog is a hyponym of animal

④ Hypernym :- A generic term from which other terms can be derived are known as Hypernym.

Example :-

⑤ Meronym :- Meronym is referred to a small component of something whole.
For eg :- Engine is a component of vehicle.

a) Couch - Sofa :- Synonym

b) Car - Wheel :- Meronym

c) Meal - Breakfast :- Hypernym
Meal is a generic term for Breakfast, lunch & dinner

d) I left my heart - and my suitcase

e) Mammal - Dog :- Hypernym
Mammal is a generic term for Dog, cats, etc.

Q.3)

→ a)

Words	A	B
Jupiter	1	0
is	1	1
the	1	2
largest	1	0
planet	1	1
Mars	0	1
fourth	0	1
from	0	1
~~the~~		
Earth	0	1

vector-A = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

vector-B = [0, 1, 2, 0, 1, 1, 1, 1, 1]

$$\text{Similarity}(A,B) = \frac{\sum A_i \times B_i}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$$

$$A.B = (1\times0 + 1\times1 + 1\times2 + 1\times0 + 1\times1 + 0\times1 + 0\times1 + 0\times1 + 0\times1)$$

$$A.B = 4$$

$$|A| = \sqrt{5}$$

$$|B| = \sqrt{10}$$

$$\text{Similarity} = \frac{4}{\sqrt{5}\times\sqrt{10}} = 0.56$$

b)

Words	TF(A)	TF(B)	IDF
Jupiter	1/5	0	$\ln(2/1) = 0.69$
is	1/5	1/8	$\ln(2/2) = 0$
the	1/5	2/8	$\ln(2/2) = 0$
largest	1/5	0	$\ln(2/1) = 0.69$
planet	1/5	1/8	$\ln(2/2) = 0$
Mars	0	1/8	$\ln(2/1) = 0.69$
fourth	0	1/8	$\ln(2/1) = 0.69$
from	0	1/8	$\ln(2/1) = 0.69$
~~the~~ earth	0	1/8	$\ln(2/1) = 0.69$

TF - IDF(A)	TF - IDF(B)
0.138	0
0	0
0	0
0.138	0
0	0
0	0.086
0	0.086
0	0.086
0	0.086

Vector - A = [0.138, 0, 0, 0.138, 0, 0, 0, 0, 0]

Vector - B = [0, 0, 0, 0, 0, 0.086, 0.086, 0.086, 0.086, 0.086]

c) TF-IDF gives more meaningful ~~accurate~~ and contextually accurate representations than simple word count method.

d)

① Jaccard Similarity :- Measures the similarity between two sets by comparing the size of their intersection and size of their union

$$\text{Jaccard Similarity}(A,B) = \left| \frac{A \cap B}{A \cup B} \right|$$

Where,

A and B are sets being compared

$|A \cap B|$ size of intersection of sets A and B

$|A \cup B|$ size of union of sets A and B

② Euclidean distance :- Euclidean distance measures the straight-line distance between two points in a multidimensional space.

$$\text{Euclidean dist}(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

(3) Pearson - Correlation Coefficient:
Measures the linear correlation between two variables. It ranges from -1 to +1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, +1 indicates perfect positive linearee relationship.

$$\text{Pearson - Corr. Coef.}(A,B) = \frac{\sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^{n} (B_i - \bar{B})^2}}$$