Advanced Human Language Technologies
Exercises on Distances and Similarities

Distances and Similarities

Exercise 1.

Given the sentences:

S1: *The man saw a car in the park*
S2: *I saw the man park the car*

Compute *similarity* between them using the following measures (if the measure yields a distance, convert the result to a similarity).

1. Euclidean
2. Vector cosine
3. Jaccard
4. Overlap

Provide the vector or set representation for each sentence used in each case. Develop your computations.

SOLUTION

Vector representations:

	the	man	saw	a	car	in	park	I
S1:	**2**	1	1	1	1	1	1	0
S2:	**2**	1	1	0	1	0	1	1

1. Euclidean: It is a distance, requires conversion.

   $$d = \sqrt{(2-2)^2 + (1-1)^2 + (1-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (1-1)^2 + (0-1)^2} = \sqrt{3} \approx 1.732$$

   Conversion to similarity:

   $$s = \frac{1}{1+d} = \frac{1}{1+1.732} = 0.366$$

2. Vector cosine: Already a similarity

   $$s = \frac{2 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 1}{\sqrt{2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2}\sqrt{2^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2}} = \frac{8}{\sqrt{10}\sqrt{9}} \approx 0.843$$

Set representations:

	the	man	saw	a	car	in	park	I
S1:	1	1	1	1	1	1	1	0
S2:	**1**	1	1	0	1	0	1	1

3. Jaccard: Already a similarity

$$s = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{5}{8} = 0.625$$

4. Overlap: Already a similarity

$$s = \frac{|S1 \cap S2|}{min(|S1|, |S2|)} = \frac{5}{min(7, 6)} = 0.833$$

Exercise 2.

Given the following term×document matrix and the number of words in each document, compute the TF·IDF score for each word/document.

Term/Doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
car	3	0	0	5	12	0	0	2	8	1
auto	8	6	0	12	0	0	9	1	3	10
best	0	1	7	0	1	5	12	0	0	0
Doc. size	40	22	15	38	29	19	47	10	25	26

SOLUTION

First, let's compute $TF(t,d) = \frac{|\{x \in d : x = t\}|}{|d|}$ for each term t and document d:

Term/Doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
car	3/40	0	0	5/38	12/29	0	0	2/10	8/25	1/26
auto	8/40	6/22	0	12/38	0	0	9/47	1/10	3/25	10/26
best	0	1/22	7/15	0	1/29	5/19	12/47	0	0	0

And now inverse document frequencies, $IDF(t,\mathcal{D}) = \log\left(\frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : t \in d\}|}\right)$ for each term:

Word *car* occurs in 6 out of 10 documents: $df(\text{car}) = 6/10 \longrightarrow IDF(\text{car}) = \log_2(10/6) = 0.74$
Word *auto* occurs in 7 out of 10 documents: $df(\text{auto}) = 7/10 \longrightarrow IDF(\text{auto}) = \log_2(10/7) = 0.51$
Word *best* occurs in 5 out of 10 documents: $df(\text{best}) = 5/10 \longrightarrow IDF(\text{best}) = \log_2(10/5) = 1$

Final TF·IDF scores result of multiplying mormalized term frequencies by their corresponding inverse document frequency, that is:

Term/Doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
car	$0.74 \times 3/40$	0	0	$0.74 \times 5/38$	$0.74 \times 12/29$	0	0	$0.74 \times 2/10$	$0.74 \times 8/25$	$0.74 \times 1/26$
auto	$0.51 \times 8/40$	$0.51 \times 6/22$	0	$0.51 \times 12/38$	0	0	$0.51 \times 9/47$	$0.51 \times 1/10$	$0.51 \times 3/25$	$0.51 \times 10/26$
best	0	$1.0 \times 1/22$	$1.0 \times 7/15$	0	$1.0 \times 1/29$	$1.0 \times 5/19$	$1.0 \times 12/47$	0	0	0

Exercise 3.

Papazom.com also needs to match offers from different suppliers that correspond to the same product, as well as to match user queries with product descriptions.

For this, they asked us to propose a similarity model able to establish how similar two product description are.

For instance, given the product descriptions.

s_1	*smartphone Hoewai x23-A with latest super AMOLED display and 64Gb*
s_2	*smartphone x23-A with 64Gb and AMOLED charge indicator*
s_3	*Hoewai smartphone z21-B with super AMOLED display and 32Gb*

1. Represent each description as a word set, and compute $sim_{jac}(s_1, s_2)$, $sim_{jac}(s_1, s_3)$, and $sim_{jac}(s_2, s_3)$ using Jaccard similarity

2. Represent each description as a word-bigram set (i.e set elements are not single words, but word-bigrams in the sentece), and compute $sim_{cos}(s_1, s_2)$, $sim_{cos}(s_1, s_3)$, and $sim_{cos}(s_2, s_3)$ using Cosine similarity.

3. A Papazon.com user wrote the search *Hoewai smartphone AMOLED display*. Compute the similarities of this query with s_1, s_2, and s_3 with each of the above metrics (unigram Jaccard and bigram Cosine).

SOLUTION

1.

	s_1	s_2	s_3
smartphone	1	1	1
Hoewai	1	0	1
x23-A	1	1	0
with	1	1	1
latest	1	0	0
super	1	0	1
AMOLED	1	1	1
display	1	0	1
and	1	1	1
64Gb	1	1	0
charge	0	1	0
indicator	0	1	0
z21-B	0	0	1
32Gb	0	0	1

$$sim_{jac}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} = \frac{6}{12} = 0.50$$

$$sim_{jac}(s_1, s_3) = \frac{|s_1 \cap s_3|}{|s_1 \cup s_3|} = \frac{7}{12} = 0.58$$

$$sim_{jac}(s_2, s_3) = \frac{|s_2 \cap s_3|}{|s_2 \cup s_3|} = \frac{4}{13} = 0.31$$

2.

	s_1	s_2	s_3
smartphone Hoewai	1	0	0
Hoewai x23-A	1	0	0
x23-A with	1	1	0
with latest	1	0	0
latest super	1	0	0
super AMOLED	1	0	1
AMOLED display	1	0	1
display and	1	0	1
and 64Gb	1	0	0
smartphone x23-A	0	1	0
with 64Gb	0	1	0
64Gb and	0	1	0
and AMOLED	0	1	0
AMOLED charge	0	1	0
charge indicator	0	1	0
Hoewai smartphone	0	0	1
smartphone z21-B	0	0	1
z21-B with	0	0	1
with super	0	0	1
and 32Gb	0	0	1

$$sim_{cos}(s_1, s_2) = \frac{|s_1 \cap s_2|}{\sqrt{|s_1|}\sqrt{|s_2|}} = \frac{1}{\sqrt{9}\sqrt{7}} = 0.13$$

$$sim_{cos}(s_1, s_3) = \frac{|s_1 \cap s_3|}{\sqrt{|s_1|}\sqrt{|s_3|}} = \frac{3}{\sqrt{9}\sqrt{8}} = 0.35$$

$$sim_{cos}(s_2, s_3) = \frac{|s_2 \cap s_3|}{\sqrt{|s_2|}\sqrt{|s_3|}} = \frac{0}{\sqrt{7}\sqrt{8}} = 0.00$$

3.

	query (q)
smartphone	1
Hoewai	1
x23-A	0
with	0
latest	0
super	0
AMOLED	1
display	1
and	0
64Gb	0
charge	0
indicator	0
z21-B	0
32Gb	0

$$sim_{jac}(s_1, q) = \frac{|s_1 \cap q|}{|s_1 \cup q|} = \frac{4}{10} = 0.40$$

$$sim_{jac}(s_2, q) = \frac{|s_2 \cap q|}{|s_2 \cup q|} = \frac{2}{10} = 0.20$$

$$sim_{jac}(s_3, q) = \frac{|s_3 \cap q|}{|s_3 \cup q|} = \frac{4}{9} = 0.44$$

	query (q)
smartphone Hoewai	0
Hoewai x23-A	0
x23-A with	0
with latest	0
latest super	0
super AMOLED	0
AMOLED display	1
display and	0
and 64Gb	0
smartphone x23-A	0
with 64Gb	0
64Gb and	0
and AMOLED	0
AMOLED charge	0
charge indicator	0
Hoewai smartphone	1
smartphone z21-B	0
z21-B with	0
with super	0
and 32Gb	0
smartphone AMOLED	**1**

$$sim_{cos}(s_1, q) = \frac{|s_1 \cap q|}{\sqrt{|s_1|}\sqrt{|q|}} = \frac{1}{\sqrt{9}\sqrt{3}} = 0.19$$

$$sim_{cos}(s_2, q) = \frac{|s_2 \cap q|}{\sqrt{|s_2|}\sqrt{|q|}} = \frac{0}{\sqrt{7}\sqrt{3}} = 0.00$$

$$sim_{cos}(s_3, q) = \frac{|s_3 \cap q|}{\sqrt{|s_3|}\sqrt{|q|}} = \frac{2}{\sqrt{8}\sqrt{3}} = 0.41$$