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.

Exercise 2.

Given the following term×document matrix:

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	2	0

1. Compute the one-hot vector for each of the three words in the vocabulary
2. Compute the TF·IDF score for each word/document.

   NOTE: Remember to normalize the matrix by the maximum value of each row:
   $\max(\text{car}) = \max(\text{auto}) = \max(\text{best}) = 12$

Exercise 3.

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

Exercise 4.

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).