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:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>auto</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>best</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

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.

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