CEN 807 Roll No………………………….

B.Tech (Computer Engineering) 8th Semester Examinations 2022

Natural Language Processing and Information Extraction

Paper Code: CEN 807

Maximum Marks: 60 Maximum Time: 3 hr

(Write your Roll No. on the top immediately on receipt of this question paper)

Note: Attempt any two parts from each question Assume suitable data, if necessary.

<table>
<tr><th>S.No.</th><th>Questions</th><th>Marks</th><th>CO</th></tr>
<tr><td>1(a)</td><td>What is tokenization? What is the advantage of using Byte Pair Encoding scheme for tokenization? Train the BPE algorithm using the following table.
<table>
<tr><td>low</td><td>5</td></tr>
<tr><td>lowest</td><td>2</td></tr>
<tr><td>newer</td><td>6</td></tr>
<tr><td>wider</td><td>3</td></tr>
<tr><td>new</td><td>2</td></tr>
</table>
Test the BPE algorithm using the word “lower” and show its tokenization. Show all the steps for training and testing clearly.</td><td>6</td><td>1</td></tr>
<tr><td>1(b)</td><td>Given the following bigram counts.
<table>
<tr><td></td><td>i</td><td>want</td><td>to</td><td>eat</td><td>Chinese</td><td>food</td><td>lunch</td><td>spend</td></tr>
<tr><td>i</td><td>5</td><td>827</td><td>0</td><td>9</td><td>0</td><td>0</td><td>0</td><td>2</td></tr>
<tr><td>want</td><td>2</td><td>0</td><td>608</td><td>1</td><td>6</td><td>6</td><td>5</td><td>1</td></tr>
<tr><td>to</td><td>2</td><td>0</td><td>4</td><td>686</td><td>2</td><td>0</td><td>6</td><td>211</td></tr>
<tr><td>eat</td><td>0</td><td>0</td><td>2</td><td>0</td><td>16</td><td>2</td><td>42</td><td>0</td></tr>
<tr><td>Chinese</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>82</td><td>1</td><td>0</td></tr>
<tr><td>food</td><td>15</td><td>0</td><td>15</td><td>0</td><td>1</td><td>4</td><td>0</td><td>0</td></tr>
<tr><td>lunch</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>spend</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td><td>6</td><td>1</td></tr>
</table>

<table>
<tr><td></td><td>(1) Generate the Bigram probability matrix.
(2) Generate the Bigram probability matrix with add-1 smoothing.</td><td></td><td></td></tr>
<tr><td>1(c)</td><td>Define and give examples:
(1) Segmentation
(2) Lemmatization
(3) Stemming</td><td>6</td><td>1</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>2(a)</td><td>
<table>
<tr><td></td><td>Doc</td><td>Words</td><td>Class</td></tr>
<tr><td rowspan="4">Training</td><td>1</td><td>Chinese, Beijing, Chinese</td><td>C</td></tr>
<tr><td>2</td><td>Chinese, Chinese, Shanghai</td><td>C</td></tr>
<tr><td>3</td><td>Chinese, Macao</td><td>C</td></tr>
<tr><td>4</td><td>Tokyo, Japan, Chinese</td><td>J</td></tr>
<tr><td>Test</td><td>5</td><td>Chinese, Chinese, Tokyo, Japan</td><td>?</td></tr>
</table>
Compute the most likely class for Doc 5. Assume a multinomial naive Bayes classifier and use add-α La Place smoothing for the likelihoods. Suitable value of α may be chosen.</td><td>6</td><td>2</td></tr>
<tr><td>2(b)</td><td>Consider the following training data:
<table>
<tr><td>S.No.</td><td>Document</td><td>Class</td></tr>
<tr><td>1.</td><td>Natural Language Processing</td><td>A</td></tr>
<tr><td>2.</td><td>Language Model Learning</td><td>A</td></tr>
<tr><td>3.</td><td>Ngram Langauge Model</td><td>A</td></tr>
<tr><td>4.</td><td>Text Classification Model</td><td>A</td></tr>
<tr><td>5.</td><td>Text Processing Model</td><td>A</td></tr>
<tr><td>6.</td><td>Computer Vision</td><td>B</td></tr>
<tr><td>7.</td><td>Image Classification Model</td><td>B</td></tr>
<tr><td>8.</td><td>Object Segmentation</td><td>B</td></tr>
<tr><td>9.</td><td>Image Processing</td><td>B</td></tr>
<tr><td>10.</td><td>Object Recognition</td><td>B</td></tr>
</table>
And Test Data:
<table>
<tr><td>1.</td><td>Object Recognition Model</td><td>?</td></tr>
<tr><td>2.</td><td>Text Recognition Model</td><td>?</td></tr>
</table>
Predict the class for test samples using Multinomial Naïve Bayes and Bigram language model with La-Place smoothing.</td><td>6</td><td>2</td></tr>
<tr><td>2(c)</td><td>Define Precision, Recall and F-measure.</td><td>6</td><td>2</td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>3(a)</td><td>Given the following Dictionary entry for <b>line</b>.
<b>line</b>^{2} a length of cord, rope, wire, or other material serving a particular purpose: <i>wring the clothes and hang them on the line | a telephone line</i>.
• one of a vessel’s mooring ropes.
• a telephone connection: <i>she had a crank on the line</i>.
• a railroad track.
• a branch or route of a railroad system: <i>the Philadelphia to Baltimore line</i>.

<b>line</b>^{3} a horizontal row of written or printed words.
• a part of a poem forming one such row: <i>each stanza has eight lines</i>.
• (lines) the words of an actor's part in a play or film.
• a particularly noteworthy written or spoken sentence: <i>his speech ended with a line about the failure of justice</i>.

Which of these senses are related by homonymy, and which are related by polysemy? For any senses which are polysemous, give an argument as to how the senses are related.</td><td>6</td><td>3</td></tr>
<tr><td>3(b)</td><td>Assume the following sentence L in which the word <b><i>line</i></b> is in focus:
<b>L</b> = you must wait in a long <b><i>line</i></b> at the checkout counter
Give a collocation feature vector (including n-gram) for in the word <b><i>line</i></b> in L, given a window size of 3 words to the left and 3 words to the right.</td><td>6</td><td>3</td></tr>
<tr><td>3(c)</td><td><b>C</b> = About three years ago, he nearly gave up because he nearly had nothing to sell; Now his shelves are full, and towels and clothes hang from a <b><i>line</i></b> overhead.

For the word <b><i>line</i></b> in the above text L, generate the bag-of-words feature vector for window size = +-2, assume <b>C</b> as the whole corpus.</td><td>6</td><td>3</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>4(a)</td><td>For the following term document matrix:
<table>
<tr><td></td><td>Document 1</td><td>Document 2</td><td>Document 3</td><td>Document 3</td></tr>
<tr><td>digital</td><td>1</td><td>0</td><td>7</td><td>13</td></tr>
<tr><td>computer</td><td>114</td><td>80</td><td>62</td><td>89</td></tr>
<tr><td>information</td><td>36</td><td>58</td><td>1</td><td>4</td></tr>
</table></td><td>6</td><td>4</td></tr>
</table>

<table>
<tr><td></td><td><table><tr><td>data</td><td>20</td><td>15</td><td>2</td><td>3</td></tr></table>
Calculate the similarity between good and fool using (i) cosine similarity (ii) PPMI. Use add 2 smoothing if necessary.</td><td></td><td></td></tr>
<tr><td>4(b)</td><td>What are word vectors? Illustrate with the help of a suitable diagram. Define the cosine similarity between two word vectors.</td><td>6</td><td>4</td></tr>
<tr><td>4(c)</td><td>Construct the word co-occurrence matrix, window size+-1, for the following text:
Document: “Roses are red. Sky is blue.</td><td>6</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>5(a)</td><td>Define Information Extraction, Named Entity Recognition and Relation Extraction. Why is Information Retrieval task not sufficient to perform information extraction tasks?</td><td>6</td><td>5</td></tr>
<tr><td>5(b)</td><td>What are the different encoding schemes for Named Entity Recognition? Illustrate with examples.</td><td>6</td><td>5</td></tr>
<tr><td>5(c)</td><td>Construct a rule based e-mail extractor which can distinguish between sender and receiver e-mail addresses.</td><td>6</td><td>5</td></tr>
</table>