CEN 807 Roll No…….

B.Tech (Computer Engineering) 8th Semester Examinations 2021

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>Apply the Max-Match word segmentation algorithm on the following:
“wecanonlyseeashortdistanceahead”
Assume you have the English all vocabulary set.</td><td>6</td><td>1</td></tr>
<tr><td>1(b)</td><td>Consider Count(w), Count(v,w) be unigram and bigram counts taken from a training corpus, where w is the single word, v,w is a bigram. Let N be te total number of words in the Corpus. What are the maximum likelihood estimates for unigram and bigram model?</td><td>6</td><td>1</td></tr>
<tr><td>1(c)</td><td>Write the training and prediction algorithms for Bernoulli Naïve Bayes Text Classification. Consider the following dataset:
<table>
<tr><td>Feature</td><td>1</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>Label</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>a</td><td>b</td></tr>
</table>
Predict the label values a and b using Bernoulli Naïve Bayes Text Classification</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>
</table>

2(b)	Given a corpus w1,w2,w3,…wN such that each word is labelled with three POS tags: NOUN, VERB, ADJECTIVE, OTHER. Construct the graphical sequence labelling model using HMM for POS tagging and define all the transition and emission probabilities.	6	2
2(c)	Define Precision, Recall and F-measure.	6	2
3(a)	Given the following Dictionary entry for **line**. **line**[2] a length of cord, rope, wire, or other material serving a particular purpose: *wring the clothes and hang them on the line	a telephone line*. • one of a vessel's mooring ropes. • a telephone connection: *she had a crank on the line*. • a railroad track. • a branch or route of a railroad system: *the Philadelphia to Baltimore line*. **line**[3] a horizontal row of written or printed words. • a part of a poem forming one such row: *each stanza has eight lines*. • (lines) the words of an actor's part in a play or film. • a particularly noteworthy written or spoken sentence: *his speech ended with a line about the failure of justice*. 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.	6	3
3(b)	Assume the following sentence L in which the word ***line*** is in focus: **L** = you must wait in a long ***line*** at the checkout counter Give a collocation feature vector (including n-gram) for in the word ***line*** in L, given a window size of 3 words to the left and 3 words to the right.	6	3
3(c)	**C** = 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 ***line*** overhead. For the word ***line*** in the above text L, generate the bag-of-words feature vector for window size = +-2, assume **C** as the whole corpus.	6	3

<table>
<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><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>6</td><td>4</td></tr>
<tr><td>4(b)</td><td>Describe the Logistic Regression based skip gram model with the help of suitable diagram.</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>