Nirma University
Institute of Technology
Semester End Examination, May-2021
B. Tech(CSE), Semester-VI
2CSDE70 Natural language processing

Roll No. [   ]   Supervisor's initial with date: [   ]

Time: 1:30 hours   Max. Marks: 40

Instructions: 1. Attempt all questions.
2. Figures to right indicate full marks.
3. Draw neat sketches wherever necessary.
4. Assume suitable data wherever necessary and clearly indicate it.

Q:1   Explain porter stemmer algorithm with example.   [04]

Q:2   Apply bigram probability and find out highest probability pair from   [04]
following data.
<s> This is book </s>
<s> book This is </s>
<s> This cannot be boring and sad</s>

Q:3   Write regular expression for following definitions.   [08]
1. Regular expression for strong password.
2. Regular expression to filter binary numbers with odd number of
   0's and1's.
3. Regular expression to remove past participles from words.
4. Regular expression to recognize C programming tokens.

Q:4   For the following document collection, Use and depict the Probabilistic   [08]
model and what are the returned results for query "gold silver truck"
D1 : "shipment of gold damaged in a fire"
D2 : " delivery of silver arrived in a silver truck "
D3 : "Shipment of gold in a truck"

Q:5   Given the following email classification, each labeled, either spam or   [08]
not spam.
1. sale, discount, shop, shop   **spam**
2. fees, exam, study **not spam**
3. discount, free, fees, sale, sale **spam**
4. exam, study, study, sale **not spam**
5. free, fees, study, shop **not spam**
Test data D: fees, discount, study, free
compute the most likely class for D. Assume a naive Bayes classifier
and use add-1 smoothing for the likelihoods.

Q:6   Explain direct relevance feedback and indirect relevance feedback with   [08]
       example. Discuss Rocchio method for relevance feedback.