National Institute of Technology Rourkela

Department of Computer Science and Engineering

Spring Mid Semester Examination (February), 2020
Subject: **Natural Language Processing** (CS 424)

Write neatly and legibly.
Answer **all** the questions.
Figures at the right hand margin indicate marks.
All parts of a question should be answered at one place.
Number your answers correctly according to the numbering system used in this question paper.

Full Marks: 30 Time: 2 Hours

1. Answer the following questions. [2 x 5 = 10]
   (a) Write short notes on (i). Lemmatization, (ii) Dependency Parsing, (iii). Named Entity Recognition, (iv). Co-reference Resolution.
   (b) What are the techniques used for evaluating a language model. Explain with suitable examples.
   (c) How do Finite State Transducers take an advantage over Finite State Machines? Explain with a suitable example.
   (d) Describe the ambiguity issues in various stages of Natural Language Processing.
   (e) Use the Penn Treebank tagset to tag each word in the following sentences. You may ignore punctuation.
      i. It is a nice night.
      ii. . . . I am sitting in Mindy's restaurant putting on the gefillte fish, which is a dish I am very fond of, . . .
      iii. . . . Nobody ever takes the newspapers she sells . . .
      iv. This crap game is over a garage in Fifty-second Street. . .

2. Consider the following documents (D_1, D_2, and D_3) and the query Q. [6]
   D_1: Shipment of gold damaged in a fire
   D_2: Delivery of silver arrived in a silver truck
   D_3: Shipment of gold arrived in a truck
   Q: gold silver truck
   Compute the similarity coefficient between the query and documents and retrieve the documents ranks using Vector Space Model

3. Reuters Dataset comprises of 54716 Sentences. Given below (table) is the raw bigram count between the occurrences of few words. The unigrams of the words in the table are given as: but (2429), we (905), must (322), be (6288), very (421), careful (23). Compute the raw bigram probabilities, Laplace smoothed bigram counts, Laplace-smoothed bigrams probabilities, and Reconstituted counts for the dataset. Consider the vocabulary size as 50000. [4]

	but	we	must	be	very	careful
but	0	60	0	2	1	0
we	0	0	17	2	0	0
must	0	0	0	113	0	0
be	1	0	0	2	31	4
very	0	0	0	0	1	3
careful	0	0	0	0	0	0

4. The n-gram language modeling has the problem of zero occurrences in the n-gram counts. Deduce various techniques to overcome these zero-occurrence problem. [4]

5. For a given language the probabilities given in the table holds with respect to the occurrences of Noun, Adjective and Verb for machine translation to another language. [6]

	Adjective	Noun	Verb
Adjective	0.8	0.05	0.15
Noun	0.2	0.6	0.2
Verb	0.2	0.3	0.5

(a) Given that the first word is Adjective, what is the probability that the next word is Adjective and the word after is Noun?

(b) Given that current word is a Verb, what is the probability that an Noun will occur two words from now?

(c) Give the third word is a Noun, what is the probability that the sixth word is also a Noun?