NATURAL LANGUAGE PROCESSING
(CSEN 4231)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) What is the following transformation called in NLP?
The boy's cars are different colors → The boy car be different color
(a) Unification (b) Lemmatization
(c) NER (d) Pruning

(ii) What is the Levenshtein distance between "drive" and "brief" assuming all operations cost 1
(a) 2 (b) 3 (c) 4 (d) 5

(iii) Deciding whether the full stop at the end of "Inc." is the end of a sentence or not is a problem of
(a) Sentence segmentation (b) Word Segmentation
(c) Tokenization (d) Word Embedding

(iv) In HMM based POS tagging using bigrams, tag transition probabilities measure
(a) The likelihood of a POS tag given a word
(b) The likelihood of a POS tag given the preceding tag
(c) The likelihood of a word given a POS tag
(d) The likelihood of a POS tag given all preceding tags

(v) An n-gram is
(a) A contiguous sequence of n words
(b) A contiguous sequence of (n-1) words
(c) A sequence of n words but not necessarily contiguous
(d) A sequence of n-1 words but not necessarily contiguous

(vi) A dependency tree which can be drawn with no crossing arcs is called
(a) Projective (b) Non-Projective (c) Planar (d) Non-Crossing

(vii) Which of the following is not a vectorization method?
(a) One-hot encoding (b) One-hot hashing
(c) Word Embedding (d) Regularization

(viii) The Chu-Liu-Edmonds algorithm for Graph based dependency parsing finds
(a) Minimum spanning tree in a directed graph
(b) Maximum spanning tree in a directed graph
(c) Maximum spanning tree in a undirected graph
(d) Minimum spanning tree in a undirected graph

(ix) Tagging of words/phrases in a sentence as shown below into "Agent", "Predicate", "Theme", "Location" etc. is called
(a) Selectional Restriction (b) Semantic Role Labelling
(c) NER Tagging (d) POS Tagging

(x) Summarization based on phrases and sentence in the source document is called
(a) abstractive (b) extractive
(c) transformative (d) discriminative

Fill in the blanks with the correct word

(xi) The sentence "*We saw the Eiffel Tower flying to Paris"*, has _____ ambiguity.

(xii) Deciding whether the word "*bank*" refers to the financial institution or the sloping side of a river is called _____.

(xiii) _____ is the decomposition of a text into sentences.

(xiv) If we a apply MaxPooling on an array (0.1, 1, 0.2, 0.1, 0, 0, 1) the output assuming a pool size of 3 and a stride of 2 is _____.

(xv) In Artificial Neural Networks, ReLU refers to the function f(x) = _____.

Group - B

2. (a) Define the minimal edit distance problem, briefly describing applications in NLP.
[(CO6)(Remember/LOCQ)]

(b) Given two strings X[1..n] and Y[1..m], let the Minimum Edit Distance D(i,j) be the edit distance between X[1..i] and Y[1..j]. Define the recurrence for D(i, j). The operations insertions and deletions cost 1 and substitutions cost 2. A part of the table is shown. Compute the value of D(i, j) for the cell labelled * if i values increase to the right and j values increase downwards?

22	23	24
21	22	23
22	23	*

[(CO1)(Apply/IOCQ)]

(c) Explain how the optimal sequence of edits is obtained by the Minimal Edit Distance dynamic programming algorithm, assuming that the allowed operations are insert, delete, substitute and transposition of adjacent characters.
[(CO1)(Analyse/HOCQ)]

4 + 4 + 4 = 12

3. (a) Which of the following will the bigram noisy channel model choose and why?
A: "Swimming across the river"
B:"Swimming access the river"

Given:
P(across | Swimming) = 0.3, P(Swimming | across) = 0.1
P(access | Swimming) = 0.2, P(Swimming | access) = 0.2
P(the | across) = 0.3, P(across | the) = 0.3
P(the | access) = 0.4, P(access | the) = 0.4 *[(CO3)(Apply/IOCQ)]*

(b) In n-gram smoothing, what is "*stupid backoff*", and what are its advantages? *[(CO1)(Remember/LOCQ)]*

(c) Show that the smoothed probabilities add up to 1 in Witten-Bell Smoothing. *[(CO3)(Analyze/HOCQ)]*

4 + 4 + 4 = 12

Group - C

4. (a) Consider the text classification problem into two classes Chinese and Japanese of which there are 3 and 2 documents respectively. Total count of word "Kunming" in the Chinese docs is 8. Total word count of all Chinese docs together is 24. The total vocabulary size V is 12. What is P("Kunming"| Chinese)? Assume that we are using Multinomial Naïve Bayes with add-1 smoothing. *[(CO3)(Apply/IOCQ)]*

(b) Give 4 applications of NER (Named Entity Recognition). *[(CO6)(Remember/LOCQ)]*

(c) Why are Naïve Bayes calculations often done in log space? *[(CO1)(Understand/LOCQ)]*

4 + 4 + 4 = 12

5. (a) Why is the inverse document frequency IDF important in text processing? In a collection of 100 documents, 20 documents contain the word *wind* and 4 documents contain the word *cyclone*. What is the ratio $IDF_{wind}/IDF_{cyclone}$? *[(CO3)(Analyse/HOCQ)]*

(b) In the Viterbi Algorithm for POS tagging to find the most likely sequence of tags for a given sequence of words, if
$max_{1 <= i <= N} v_{t-1}(i)a(i, j) = 0.5$ and $v_t(j) = 0.2$, what is the value of the emission probability in state j for observation o_t $b_j(o_t)$?
Here $v_t(k)$ is the viterbi value for state k after t observations and a(i, j) is the transition probability from state i to state j. *[(CO3)(Apply/IOCQ)]*

(c) Explain briefly two fundamental sequence modelling problems in language processing which are solved using HMMs. *[(CO3)(Remember/LOCQ)]*

4 + 4 + 4 = 12

Group - D

6. (a) Consider the CFG
S → NP VP
VP →ate NP
NP→NP R
R → that NP ate
NP→ the N
N→ hawk | snake | mouse | corn
Show the parse tree that generates the following, or indicate that it is not generated:
(i) the mouse ate the corn

(ii) the mouse that the snake ate ate the corn
(iii) the mouse that the snake that the hawk ate ate ate the corn

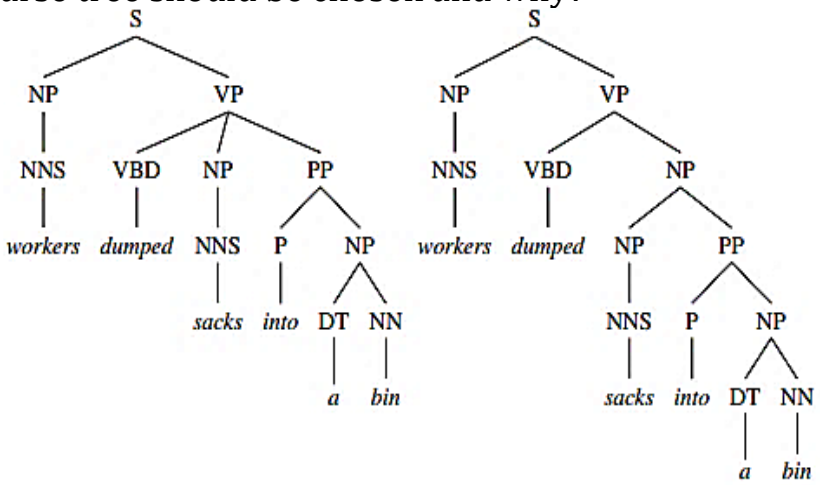
[(CO4)(Analyse/HOCQ)]

(b) Show the trace of a transition-based dependency parser and indicate the relation added at each step while parsing "*book me the morning flight*". *[(CO4)(Apply/IOCQ)]*

(2 + 2 + 2) + 6 = 12

7. (a) Consider the probabilistic CKY algorithm. At any point during the algorithm, table [2,8,A] = 0.009, P(A→XY) = 0.2, table[2,4,X] = 0.3, table[4,8,Y] = 0.1. Then when the algorithm terminates, what can we say about table[2,8,A] *[(CO5)(Apply/IOCQ)]*

(b) Which parse tree should be chosen and why? *[(CO5)(Apply/IOCQ)]*



(c) What are the problems with parsing *garden path sentences*? Explain with the following example:
The students forgot the solution was in the back of the book *[(CO1)(Apply/IOCQ)]*

4 + 4 + 4 = 12

Group - E

8. (a) A question answering system generates candidate answers with the first correct answer in position 2, 3, 2, 1, 4 for 5 queries.
Find the mean reciprocal rank. *[(CO3)(Apply/IOCQ)]*

(b) Give two examples each of single document summarization and multi-document summarization. *[(CO1)(Remember/LOCQ)]*

(c) Describe the ROUGE evaluation metric for document summarization. *[(CO6)(Remember/LOCQ)]*

4 + 4 + 4 = 12

9. (a) What are word embeddings and why are they important? *[(CO2)(Remember/LOCQ)]*

(b) What are pre-trained word embeddings? Give two examples. *[(CO6)(Remember/LOCQ)]*

(c) With code snippets show how to implement embedding layer in keras and tensorflow. Assume a 1000 word vocabulary is embedded in 5 dimensions. *[(CO2)(Apply/IOCQ)]*

4 + 4 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	37.5	43.75	18.75