LNMIIT

The LNM Institute of Information Technology
Department of Computer Science & Engineering
CSE 3201 Natural Language Processing
Exam Type: Mid Term

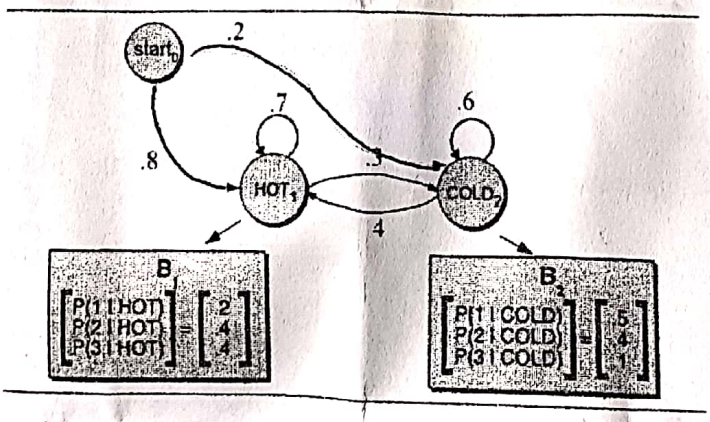
Time: 90 min 30/09/2019 Max. Marks: 30

Answer all questions **in the same order as it appears in the question paper.** If there are any assumptions to be made for your answer write clearly the assumption that you are making before answering. Only if the assumption is reasonable it will be considered. **No doubt clarifications in the examination hall!** All the best!

1. Find the regular expression for strings $w \in \{0,1\}^*$ where w has exactly a pair of consecutive zeros or a pair of consecutive ones but not both. (2)

2. Compute the minimum edit distances by making the memoisation table for the minimum edit distance algorithm and figure out whether the word *drive* is closer to *brief* or to *divers*, and what the edit distances are? You may use costs as 1 for insertion, 1 for deletion, 2 for substitution. (6)

3. Use the Viterbi algorithm and the given HMM below to compute the most likely weather sequences for each of the two observation sequences 33111 and 33113. Draw the Trellis and fill up the Trellis by using Viterbi's algorithm. (6)



4. Consider the following grammar $G = (V, T, P, S)$ where $V = \{E, I\}; T = a, b, 0, 1; S = E$; and P is defined as $E \rightarrow I|E + E|E * E|(E)$ and $I \rightarrow a|b|Ia|Ib|I0|I1$. Using CKY parser illustrate and check if the string $(a1 + b0 * a1)$ is generated by G or not. Find the set of rules applied to get the same. Also by the same method check if the string $(a1 + 1ba)$ is generated by G or not. (6)

5. We are conducting a draw of lots using a bowl. Bowl contains bits of paper where each bit of paper has a word written on it. These words are collected from a corpus. There are seven tags that are available in this corpus, namely, *pronoun*, *proper noun*, *adjective*, *preposition*, *determiner*, *adverb*, and *verb*. Assume that we have already done some lots (with replacement) and seen 7 prepositions, 3 pronouns, determiners ~~nouns~~, 2 verbs, and 1 proper noun. Adjective and adverb are yet to be seen. Answer the following questions:

   (a) What is the probability that the next draw will result either in an *adverb* or an *adjective* – compute this by using MLE method and Good Turing method. (2)

All the best!

(b) For the same observation what will be the probability that the next word will be a *verb*? Use both MLE and Good Turing method. (2)

6. The Soundex algorithm is a method given below:

   (a) Keep the first letter of the name, and drop all occurrences of non-initial a, e, h, i, o, u, w, y.

   (b) Replace the remaining letters with the following numbers:

   i. $b, f, p, v \rightarrow 1$
   ii. $c, g, j, k, q, s, x, z \rightarrow 2$
   iii. $d, t \rightarrow 3$
   iv. $l \rightarrow 4$
   v. $m, n \rightarrow 5$
   vi. $r \rightarrow 6$

   (c) Replace any sequences of identical numbers, only if they derive from two or more letters that were adjacent in the original name, with a single number (e.g., $666 \rightarrow 6$).

   (d) Convert to the form `Letter Digit Digit Digit` by dropping digits past the third (if necessary) or padding with trailing zeros (if necessary).

   For example if we apply the above algorithm on the name *Jurafsky* we will get $J612$ and for the name *Sakthi* we shall get $S23$. Answer the following questions:

   (a) Take your first name and apply the Soundex algorithm on it and find the equivalent code. (1)

   (b) Find an application of this algorithm and discuss it briefly with your supporting arguments. (3)

7. Fill in the blank by a word (your appropriate guess) by using two different methods mentioned below:

   I did my exam _____.

   (a) by using absolute-discounting interpolation (b) using Kneser-Ney method. (2)