Question 1 [40 marks]

1) **Which of the following strings cannot be derived from the symbol S using the rules S ! SS | aaa | aaaaa ?**

(a) aaaaa
(c) aaaaaaa
(e) aaaaaaaaa
(b) aaaaaa
(d) aaaaaaaa

2) **Which of the following pairs of rules involves indirect recursion?**

(a) NP -> Det Adj N , NP -> Pron
(b) VP -> V NP , NP ->Det N that VP
(c) N' -> N N' , NP -> Det N'
(d) S -> AdvP S , AdvP -> Adj Adv
(e) VP -> V NP , NP -> Det N PP

3) **Which of the following strings is a member of the language over {a, b} defined by the regular expression (aa + ba)* (bb)*?**

(a) aabbba
(b) aaaabb
(c) babbaa
(d) bbaa
(e) None of the above.

4) **Consider the following context-free grammar, with start symbol S and terminals a, : , <, >.**

S -> <L | a L ->aR | <LR R -> > | ;L

How many different parse trees are there for the string << a >; a > ?

(a) 0
(b) 1
(c) 2
(d) 3
(e) More than three.

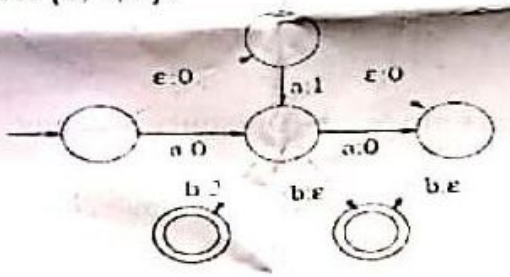
5) Which of the following English words ?

(a) drink
(b) time
(c) exam
(d) therefore
(e) can

6) Which of the following strings does not match the regular expression written in
machine syntax as (0*1?2+)*

(a) ε
(b) 02
(c) 12
(d) 01201
(e) 12222

7) Consider the following finite state transducer with input alphabet {a, b, c} and output alphabet {0, 1, 2}:



Which of the following is *not a possible* output string that could arise from processing the input string ab

(a) 0
(b) 00
(c) 01
(d) 02
(e) 012

8) Consider the following probabilistic context-free grammar:

S	—	N VP	(1.0)
VP	—	IV	(0.8)
VP	—	TV N	(0.2)
N	—	clocks	(0.7)
N	—	tables	(0.3)
IV	—	tick	(0.7)
IV	—	fly	(0.3)
TV	—	hate	(0.7)
TV	—	chase	(0.3)

Which of the following sentences is assigned the highest probability by this grammar?

(a) clocks tick
(b) tables fly
(c) clocks fly
(d) tables chase clocks
(e) tables hate clocks

9) Which of the following sets of productions is not in Chomsky normal form?

(a) G1
(b) G2
(c) G3
(d) G1 and G3
(e) G2 and G3

$$G1: \quad S \to AB \quad A \to AB|a \quad B \to Ba|b$$
$$G2: \quad S \to AB \quad A \to AB|a \quad B \to BA|b$$
$$G3: \quad S \to A|B \quad A \to AB|a \quad B \to BA|b$$

10) Consider the grammar below:

NP → art NP1
NP → ppro NP1
NP1 → num NP1
NP1 → NP2
NP2 → adj NP2
NP2 → adj NP3
NP3 → noun NP3
NP3 → noun

Which of the following part-of-speech tag sequence can be produced by grammar?

(a) art num noun
(b) art num num noun
(c) art adj noun
(d) art noun
(e) art adj num noun

11) What type of ambiguity exists in the word sequence "Time flies"?

(a) Syntactic
(b) Semantic
(c) Phonological
(d) Anaphoric
(e) none of the above

12) *computer vs computational* is an example of _____ morphology.

a) Inflectional
b) Derivational
c) Cliticization
d) None of the above

13) What is the number of trigrams in a normalized sentence of length n words?

(a) n
(b) n-1
(c) n-2
(d) n-3

14) **Assume that there are 10000 documents in a collection. Out of these, 50 documents contain the terms "difficult task". If "difficult task" appears 3 times in a particular document, what is the TFIDF value of the terms for that document?**

(a) 8.11 (b) 15.87
(c) 0 (d) 81.1

15) **Let us suppose that you have the following two 4-dimensional word vectors for two words w1 and w2 respectively:**

$w_1 = (0.2, 0.1, 0.3, 0.4)$ and $w_2 = (0.3, 0, 0.2, 0.5)$

What is the cosine similarity between w_1 and w_2?

(a) 0.948 (b) 0.832
(c) 0 (d) 0.5

16) **When training a language model, if we use an overly narrow corpus, the probabilities**

(a) Don't reflect the task (b) Reflect all possible wordings
(c) Reflect intuition (d) Don't generalize

17) **Morphotactics is a model of**

a) Spelling modifications that may occur during affixation
b) How and which morphemes can be affixed to a stem
c) All affixes in the English language
d) Ngrams of affixes and stems

18) **You have collected a data of about 10,000 rows of tweet text and no other information. You want to create a tweet classification model that categorizes each of the tweets in three buckets – positive, negative and neutral. Which of the following models can perform tweet classification with regards to context mentioned above?**

a) Naïve Bayes (b) Support Vector Machine
c) Language model d) None of the above

19) **_____ is the type of morphology that changes the word category and affects the meaning.**

a) Inflectional (b) Derivational
c) Cliticization d) None of the above

20) **"He lifted the beetle with red cap." contain which type of ambiguity ?**

a) Lexical ambiguity
b) Syntax Level ambiguity
c) Referential ambiguity
d) None of the Above

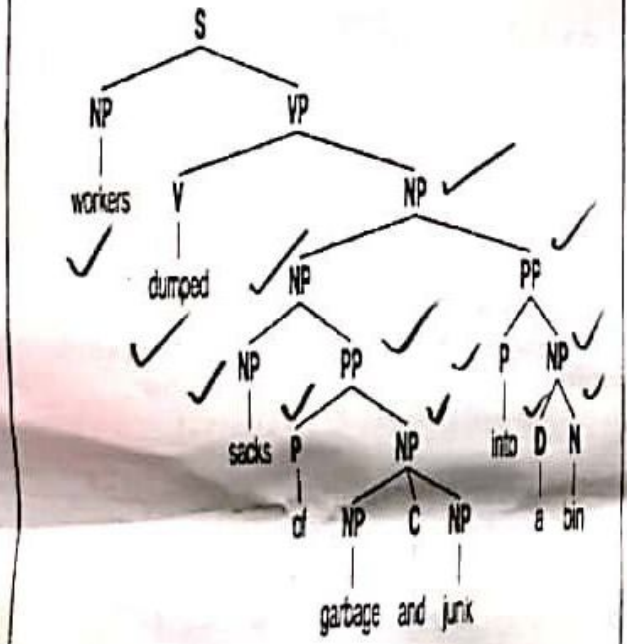
Question 2 [12 marks]

Assume we have the following context-free grammar G in Chomsky normal form:

```
S -> NP VP
VP -> V NP
NP -> D N
PP -> P NP
VP -> VP PP
NP -> NP PP
NP -> NP CNP
CNP -> C NP
NP -> "workers" | "sacks"
    | "garbage" | "junk"
N -> "bin" | "sack"
V -> "dumped"
P -> "of" | "into"
D -> "a" | "the"
C -> "and"
```

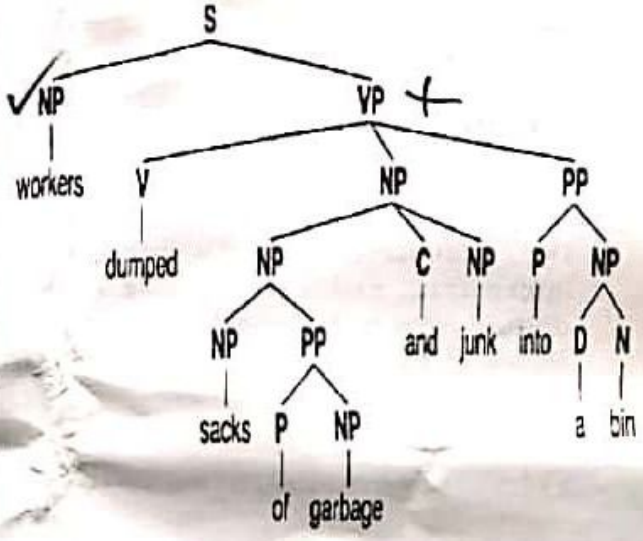
Which of the following trees for the sentence "workers dumped sacks of garbage and junk into a bin" are correct according to G?

21)



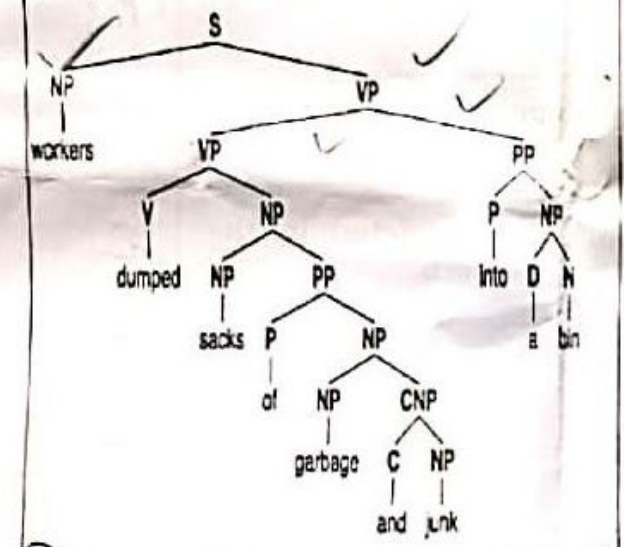
A) True B) False

22)



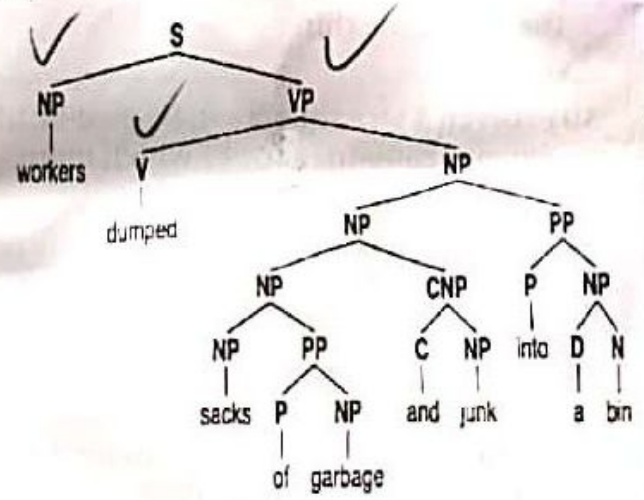
A) True B) False

23)



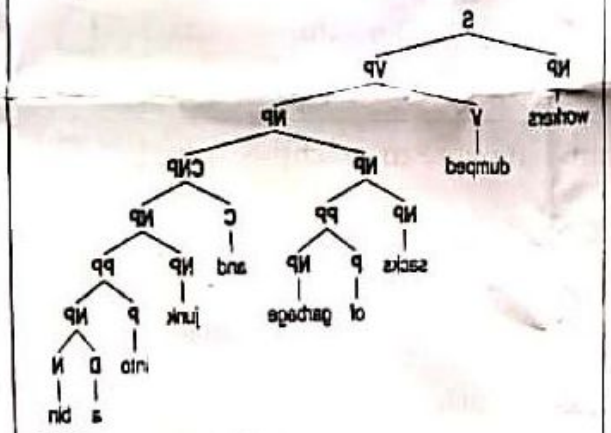
A) True B) False

24)



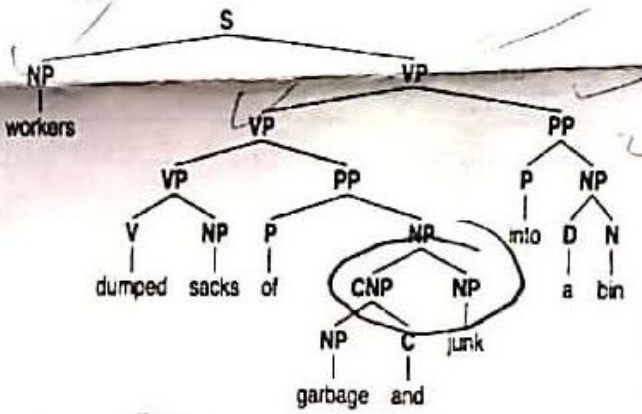
A) True B) False

25)



A) True B) False

26)



A) True B) False

Question 3 **State whether right or wrong (A for right and B for wrong in bubble sheet).** ***[28 marks]***

27) Stemming can be used for the purpose of keyword normalization, the process of converting a keyword into its base form.	(T)
28) 4-grams are better than trigrams for part-of-speech tagging.	(T)

29) IR (information Retrieval) and IE (Information Extraction) are the two same thing	(F)
30) Given a stream of text, Named Entity Recognition determines which pronoun maps to which noun.	(F)
31) Given a sound clip of a person or people speaking, determine the textual representation of the speech. "Text to speech"	(F)
32) "I am tired." Contain lexical ambiguity	(T)
33) Parts-of-Speech tagging determines all part-of-speech for a specific word given as input	(F)
34) Parsing determines Parse Trees (Grammatical Analysis) for a given sentence.	(T)
35) Co reference Resolution is Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities")	(T)
36) Machine Translation Converts human language to machine language	(F)
37) Natural Language generation is the main task of Natural language processing.	(T)
38) "Therefore" is not an open-class word in English.	(T)
39) Polysemy means different senses of same word (unrelated meaning)	(F)
40) An NLP system that identifies news articles about politics would be an information extraction system.	(T)