KATHMANDU UNIVERSITY
End Semester Examination
July/August, 2017

Level : B.E./B. Sc.
Year : IV
Time : 2 hrs. 30 mins.

Course : COMP 473
Semester : II
F. M. : 40

SECTION "B"

Attempt *ANY SIX* questions.

1. What is the difference between a Finite State Automata (FSA) and a Finite State Transducer (FST)? [4]

2. Differentiate between "inflectional" and "derivational" morphology with suitable examples. [4]

3. What is parts-of-speech tagging? Shed it's importance in NLP applications with suitable examples. [4]

4. What is the primary difference between the bag-of-words model and the n-grams model? Which of these models is employed by Information Retrieval? [4]

5. Explain the following terms in the WordNet with appropriate examples: [4]
   a. Hypernymy
   b. Hyponymy
   c. Meronymy
   d. Homonymy

6. What are the different levels of Sentiment Analysis in texts? Explain aspect level analysis in the context of product reviews. [4]

7. What is parsing? What are the basis for phrase structure based parsing and dependency grammar based parsing? [4]

SECTION "C"

[2 Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. Define context-free grammar. What are the terminal and non-terminal symbols in a context-free grammar.Consider the following grammar:

S -> NP VP VP -> Verb NP VP -> Verb PP NP -> NP PP NP -> NP and NP PP -> P NP	NP -> Kathy NP -> London NP -> Paris NP -> February	Verb -> flew P -> in P -> to CONJ -> and

Draw a parse treethat would be derived for the sentence "Kathy flew to London and Paris in February."

[4+4]

9. What are semantic roles and semantic role labeling in Natural Language Processing? In each of the following sentences, identify the semantic roles selecting from agent, patient, theme, experiencer, stimulus, goal, recipient, source, instrument, location, temporal. Justify your choice. [3 + 5]

a. The company wrote me a letter.
b. Jack opened the lock with a paper clip.
c. The river froze during the night.
d. Kathy ran to class every day at Columbia.
e. I felt the warmth of the fire.

10. Describe the linguistic as well as technical challenges of an automatic machine translation system. What are the well-known approaches and current trends of developing a Machine Translation system? [4+4]

KATHMANDU UNIVERSITY
End Semester Examination
July/August, 2017

Marks Scored:

Level : B.E. / B. Sc.
Year : IV
Exam Roll No. :
Registration No. :

Course : COMP 473
Semester : II
F. M. : 10
Time : 30 mins.
Date :

SECTION "A"
[20 Q. × 0.5 = 10 marks]

Tick the most appropriate answer.

1. Many words have more than one meaning; we have to select the meaning which makes the most sense in context. This can be resolved by
   a. Fuzzy Logic
   b. Word Sense Disambiguation
   c. Shallow Semantic Analysis
   d. All of the mentioned

2. In linguistic morphology, what is the process for reducing inflected words to their root form?
   a. Rooting
   b. Stemming
   c. Text-Proofing
   d. Both a & b

3. One of the main challenge/s of NLP is:
   a. Handling Ambiguity of Sentences
   b. Handling Tokenization
   c. Handling POS-Tagging
   d. All of the mentioned

4. Machine Translation
   a. Converts one human language to another
   b. Converts human language to machine language
   c. Converts any human language to English
   d. Converts Machine language to human language

5. Morphological Segmentation
   a. Does Discourse Analysis
   b. Separate words into individual morphemes and identify the class of the morphemes
   c. Is an extension of propositional logic
   d. None of the mentioned

6. Which of the following techniques can be used for the purpose of keyword normalization, the process of converting a keyword into its base form?
   i. Lemmatization
   ii. Levenshtein
   iii. Stemming
   iv. Soundex
   a. i and ii
   b. ii and iv
   c. i and iii
   d. i, ii and iii

7. N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence?
   "This book is a great source to learn data science"
   a. 7
   b. 8
   c. 9
   d. 10

8. In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times. What is the correct value for the product of TF (term frequency) and IDF (inverse-document frequency), if the term "data" appears in approximately one-third of the total documents?
a. K T * Log(3) b. K * Log(3) / T c. T * Log(3) / K d. Log(3) / K T

9. Google Search's feature – "Did you mean", is a mixture of different techniques. Which of the following techniques are likely to be ingredients?
i. Collaborative Filtering Model to detect similar user behaviors (queries).
b. Model that checks for Levenshtein distance among the dictionary terms
c. Translation of sentences into multiple languages
a. i b. ii c. i, ii d. i, ii, iii

10. While working with text data obtained from news sentences, which are structured in nature, which of the grammar-based text parsing techniques can be used for noun phrase detection, verb phrase detection, subject and object detection?
a. Part of speech tagging b. Dependency Parsing and Constituency Parsing
c. Skip Gram and N-Gram extraction d. Continuous Bag of words

11. Which of the following character represents zero or one of the preceding character in regular expressions?
a. ? b. * c. + d. ^

12. Which of the following string does the regular expression /[abc]/ match?
i. 'abc' ii. 'a', 'b' or 'c' iii. 'ab' or 'c' iv. 'a' or 'bc'

13. Which module in Python supports regular expressions?
a. re b. regex c. pyregex d. none of the mentioned

14. The Kleene star or the '*' symbol means:
a. One or more of the previous character
b. zero or more occurrences of the immediately previous character or expression
c. More than one of the previous characters
d. Just one previous character

15. To which of the following does the verb "should" belong to?
a. Main b. Primary c. Modal d. Auxiliary or Helping

For questions 16-18, consider the following context:

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.

16. Which of the following models can perform tweet classification with regards to the context mentioned above?
a. Naïve Bayes b. SVM c. Decision Tree d. None of the above

17. You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, with regards to the document term matrix?
i. Removal of stop words from the data will affect the dimensionality of data
ii. Normalization of words in the data will reduce the dimensionality of data
iii. Converting all the words in lowercase will not affect the dimensionality of the data
a. Only i b. Only ii c. i and ii d. i, ii and iii

18. Which of the following features can be used for accuracy improvement of a classification model?
a. Frequency count of terms
b. Part of Speech Tag
c. Grammar Structure
d. All of the above

19. What is Unicode?
a. Standard Font
b. Software
c. Character Encoding System
d. Keyboard Layout

20. While working with content extraction from a text data, you encountered two different sentences:
i. The tank is full of soldiers.
ii. The tank is full of nitrogen.
Which of the following measures can be used to remove the problem of word sense disambiguation in the sentences?
a. Compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood
b. Co-reference resolution in which one resolute the meaning of ambiguous word with the proper noun present in the previous sentence
c. Use dependency parsing of sentence to understand the meanings
d. None of the above