Registration No :

Total Number of Pages : 03   **B.Tech.**
                                                                              PIT6J004

6th Semester Regular Examination 2017-18

 6th Semester Regular Examination 2017-18

6th Semester Regular Examination 2017-18
NATURAL LANGUAGE PROCESSING
BRANCH : IT
Time : 3 Hours
Max Marks : 100
Q.CODE : C370
Answer Part-A which is compulsory and any four from Part-B.
The figures in the right hand margin indicate marks.

Part – A (Answer all the questions)

Q1   **Answer the following questions :** *multiple type or dash fill up type :*   **(2 x 10)**

a) N-grams are defined as the combination of N keywords together. How many
bi-grams can be generated from given sentence:
"Analytics Sandhya is a great source to learn data science"
(A) 7   (B) 8
(C) 9   (D) 10
(E) 11

b) What is the right order for a text classification model components
1. Text cleaning
2. Text annotation
3. Gradient descent
4. Model tuning
5. Text to predictors
(A) 12345   (B) 13425
(C) 12534   (D) 13452

c) Which of the following technique is not a part of flexible text matching?
(A) Soundex   (B) Metaphone
(C) Edit Distance   (D) Keyword Hashing

d) 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 detection
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

e) While creating a machine learning model on text data, you created a
document term matrix of the input data of 100K documents. Which of the
following remedies can be used to reduce the dimensions of data –
 1. Latent Dirichlet Allocation
 2. Latent Semantic Indexing
 3. Keyword Normalization
(A) Only 1   (B) 2, 3
(C) 1, 3   (D) 1, 2, 3

f) While working with context extraction from a text data, you encountered two
different sentences : The tank is full of soldiers. 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

g) Retrieval based models and Generative models are the two popular techniques used for building chatbots. Which of the following is an example of retrieval model and generative model respectively?
(A) Dictionary based learning and Word 2 vector model
(B) Rule-based learning and Sequence to Sequence model
(C) Word 2 vector and Sentence to Vector model
(D) Recurrent neural network and convolutional neural network

h) What is the major difference between CRF (Conditional Random Field) and HMM (Hidden Markov Model)?
(A) CRF is Generative whereas HMM is Discriminative model
(B) CRF is Discriminative whereas HMM is Generative model
(C) Both CRF and HMM are Generative model
(D) Both CRF and HMM are Discriminative model

i) You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, in regards to document term matrix?
1. Removal of stopwords from the data will affect the dimensionality of data
2. Normalization of words in the data will reduce the dimensionality of data
3. Converting all the words in lowercase will not affect the dimensionality of the data
(A) Only 1   (B) Only 2
(C) Only 3   (D) 1 and 2
(E) 2 and 3   (F) 1, 2 and 3

j) Which of the following features can be used for accuracy improvement of a classification model?
(A) Frequency count of terms   (B) Vector Notation of sentence
(C) Part of Speech Tag   (D) Dependency Grammar
(E) All of these

Q2 **Answer the following questions : *Short answer type :***   (2 x 10)

a) A sentence can easily have more than one parse tree that is consistent with a given CFG. How do PCFGs and non-probability-based CFGs differ in terms of handling parsing ambiguity?

b) What is the difference between phrase and sentence level construction?

c) Draw the top-ranked parse tree for the sentence below by applying the given PCFG.

production rule	probability
S → VP	1.0
VP → Verb NP	0.7
VP → Verb NP PP	0.3
NP → NP PP	0.3
NP → Det Noun	0.7
PP → Prep Noun	1.0
Det→ the	0.1
Verb → Cut | Ask | Find| ...	0.1
Prep → with | in | ...	0.1

Noun → envelope | grandma | scissors | men | suits | summer | ... 0.1
Does the result seem reasonable to you? Why or why not? "Ask the grandma with scissors."

d) Describe why production rules with zero probability are problematic.

e) Describe one method to avoid zero probabilities for lexicalized PCFGs.

f) 4-grams are better than trigrams for part-of-speech tagging. Is it true or false? Explain your answer.

g) Noun phrase coreference resolution includes **pronoun** resolution, **proper noun** resolution, and **common noun** resolution. Which of the three would you expect to be the most difficult to handle computationally? Explain why.

h) Information extraction is harder than text categorization. Is it true or false? Explain your answer.

i) State two advantages of partial parsers over parsers that provide in-depth syntactic information.

j) What is the difference between phrase-based and feature-based NLP?

Part – B (Answer any four questions)

Q3 **a)** Explain the issues in computational morphology with suitable example. **(10)**

b) Discuss the applications of natural language processing. **(5)**

Q4 **a)** Given the grammar and lexicon below, show the final chart for the following sentence after applying the bottom-up chart parser. Remember that the final chart contains all edges added during the parsing process. You may use either the notation from class (i.e. nodes/links) or the notation from the book to depict the chart. **(10)**

S → VP

VP → Verb NP NP → NP PP NP → Det Noun

PP → Prep Noun

Det → the

Verb → Find

Prep → in

Noun → men | suits

Find the men in suits.

b) Give an account of CYK parser. **(5)**

Q5 **a)** Discuss language as a rule-based system. **(10)**

b) Discuss stochastic part-of-speech tagging. **(5)**

Q6 **a)** Write an algorithm for simple top-down parser with an example. **(10)**

b) Explain the five verb forms. **(5)**

Q7 **a)** Describe unification method with suitable examples. **(10)**

b) Explain how unification is implemented. **(5)**

Q8 **a)** Between the words eat and find which would you expect to be more effective in selecting restriction-based sense disambiguation. **(10)**

b) Explain the application of semantics. **(5)**

Q9 **a)** **Consider the following article for this problem.** **(10)**

"I bought my wireless keyboard/mouse set several months ago, and, like a lot of new products, it has some unanticipated issues. On the plus side, obviously, is the styling. The design is fresh, clean, and interesting. The keyboard can tilt at different angles, which was important because I had some difficulty typing with it flat. The bluetooth receiver in the charger was functional, and I appreciated having a bluetooth hub for my cellphone. The mouse and the keyboard have both proved durable and reliable despite a number of mishaps."

For each of inferences (a) through (d) below,

(a) The reviewer owns the keyboard.

(b) The charger is part of the keyboard.

(c) The reviewer had difficulty typing with the keyboard.

(d) The reviewer likes the keyboard.

1. State whether the inference depends on the **discourse context, knowledge about actions,** and/or **general world knowledge;** and

2. Describe what natural language processing techniques, if any, might enable a system to make the inference automatically.

b) What is text summarization? Explain with an example. **(5)**