MANIPAL ACADEMY OF HIGHER EDUCATION

B.Tech Vth Semester First Sessional Examination September 2022

NATURAL LANGUAGE PROCESSING [DSE 3155]

Marks: 15, Duration: 60 mins

MCQ

Answer all the questions. Section Duration: 20 mins

1. N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence: "Coursera videos are a great source to learn engineering courses" (0.5)

1. 6
2. 7
3. 8
4. 9

2. Consider the following simple bigram language model, where the vocabulary consists of the single word x, and the parameters of the model are;

$$q(a|*) = 1.0;\ q(a|a) = 0.4;\ q(END|a) = 0.6$$

Which of the following are the probabilities of the string '* a a' with and without END? (0.5)

1. 1, 0.4
2. 0.24, 0.4
3. 0.24, 0.6
4. 1.4, 1.0

3. Assume a corpus with 350 tokens in it. We have 20 word types in that corpus (V = 20). The frequency (unigram count) of word types "short" and "fork" are 25 and 15 respectively. Which of the following is the probability of "short" without smoothing and probability of "fork" after smoothing? (0.5)

1. 25/350 and 16/370
2. 26/370 and 15/20
3. 26/350 and 45/370
4. 25/370 and 16/20

4. Find the probability P(Alice |name is) as per the tri-gram model. Use the corpus given below;

<s> My name is Alice </s>

<s> Alice my name is</s>

<s> A girl said that her name is Alice </s>

<s> My daughter's name is Alice </s>

1. 1
2. 0.75
3. 0.25
4. 0.5

5. Which of the following equations is used to find the unigram probabilities using Add-1 smoothing? (0.5)

1. Count (wi)/N
2. Count (wi)/(N+1)
3. (Count (wi)+1)/(N+1)
4. *(Count (wi)+1)/(N+V)

6. *Rule-based POS taggers doesn't possess which of the following properties* (0.5)

1. The rules in Rule-based POS tagging are built auto
2. These taggers are knowledge-driven taggers
3. These taggers consist of many hand written rules
4. The information is coded in the form of rules.

7. What is the output of the following code:

```
str='This is the story about Pandu, 29, joined on 23-02-2008'
str1=word_tokenize(str)
set([w.lower() for w in str1 if w.isalpha()])
```

(0.5)

1. {'about', 'is', 'joined', 'on', 'pandu', 'story', 'the', 'this'}
2. {'the', 'story', 'about', 'is', 'joined', 'on', 'pandu', 'this'}
3. {'This',' is '.'the',' story',' about',' Pandu', 'joined ',' n '}
4. {'about', 'is', 'joined', 'on', 'pandu', 'story', 'this', 'the'}

8. Consider the following statements :

i. Sentence tokenization and word tokenization should be addressed separately
ii. Question marks and exclamation points are relatively ambiguous markers of sentence boundaries
iii. Period as a sentence boundary is more ambiguous
iv. Sentence tokenization method work by building a binary classifier

(0.5)

1. iii,ii and iv are true
2. i,ii,iii and iv are true
3. iii and iv are true
4. i and iv are true

9. Select the incorrect statement:(0.5)

1. Writing *break* instead of *brake* can be a non-word error detection
2. Correcting *pleasure* as *pressure* with reference to nearby words is context – dependent error correction.
3. Correcting the word *pleaure* as *pleasure* by only looking at the word is isolated-word correction
4. Clerk who does know the correct spelling types *beautful* instead of *beautiful* is a cognitive error

10. Which of the following is true? (0.5)

1. *Distance between drive* and *brief* is less than distance between *drive* and *divers*
2. *Distance between drive* and dive is more than distance between *drive* and *divers*
3. *Distance between drive* and dive is equal to distance between *drive* and *divers*
4. *Distance between drive* and *brief* is more than distance between *drive* and *diver*

DESCRIPTIVE

Answer all the questions. Section Duration: 40 mins

11. Let us suppose that we are given a mini-corpus consisting of four sentences. Compute probabilities for each given sentence as per bigram model. (2)

s1: <s>Paris is beautiful</s>

s2: <s>Is Paris in Europe</s>

s3: <s>Paris is in France</s>

s4: <s>Paris is a city</s>

12. Explain any four closed word class .(2)

13. Describe the dynamic programming approach to find the minimum edit distance between the words *adventure* and *advantages* .(3)

14. List the regular expression character class symbols along with its function.

i) Write the code in nltk to find the count of vowels in any given sentence.
ii) Write the code in nltk to look for all sequences of two or more vowels in any given statement. (3)