NLP Mock Test https://bit.ly/2ZOaSX8

1) Read the given statements carefully and choose the correct option.

S1: The '?' matches the preceding character zero or one time.
S2: The '*' quantifier is used to mark the presence of the preceding character zero or more times.

- S1 is true and S2 is false
- S1 false and S2 is true
- Both S1 and S2 are false
- Both S1 and S2 are true

2) Which of these kinds of words are present in any text corpus when performing Lexical Processing?

1. Highly Frequent words
2. Significant words
3. Rarely occurring words

- All 1, 2 and 3
- Only 1 and 2
- Only 2 and 3
- Only 1 and 3

3) You are performing Lemmatization on a dataset. In the given context, which of the following statements are valid?

- Lemmatization works on incorrectly spelt words
- Lemmatization expects the POS tag of a word to be passed along with the word.
- Both 1 and 2
- None of these

4) Which of the following regular expression can you use to extract the country code +91 from a valid phone number "(+91)-1234567890"?

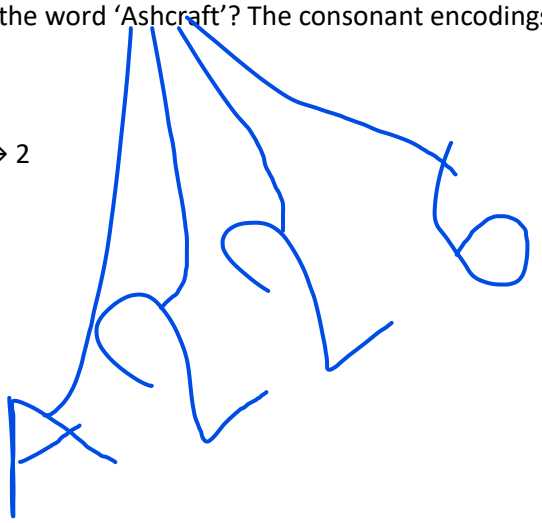
- (+\d{2})-\d{10}
- \+\d{2}-\d{10}
- +\d{2}-\d{10}
- (\+\d{2})-\d{10}

5) Which of the following statement is true about Unicode character encoding?

- UTF-8 uses 8-bit encoding for all characters
- UTF-16 uses 16-bit encoding some of the characters
- UTF-8 uses 8-bit encoding to store the English character set
- UTF-8 occupies less space per character than UTF-16 for all languages

6) What is the Soundex of the word 'Ashcraft'? The consonant encodings scheme of the Soundex are given below:

b, f, p, v → 1
c, g, j, k, q, s, x, z → 2
d, t → 3
l → 4
m, n → 5
r → 6



- A261
- A221
- A232
- A982

7) You are trying to reduce the words to its base form when working with a text corpus that contains misspelt words. What can you do to reduce the misspelt words to its base form?

- Perform Stemming on the words to produce different stems
- Perform Lemmatization on the words to produce different lemmas
- Correct the spellings of the words and perform Stemming or Lemmatization
- Correct the spellings of the words and perform Stemming only

8) Which of these regular expressions is equivalent to the regular expression "ab?cb*" ?

1. ab{0,1}cb{0, }
2. ab{0, }cb{0,1}
3. ab{0,1}cb{1, }

- Only 1
- Only 2
- Only 3
- None of these

9) You have a string 234AAAA and you only want a substring 234A to be matched when a pattern is run on the string. Which of the following patterns can you use in order to do this?

- \w*A
- \w*?A
- w*?A
- w*?A

10) You want to label each word in the sentence given alongside using the inside-outside-beginning (IOB) tags. Which of the following IOB labels is the most appropriate for the given sentence?

"*New Delhi is the capital of India*"

- New: B Delhi: I is: O the: O capital: B of: I India: O
- New: B Delhi: I is: O the: O capital: O of: I India: O
- New: B Delhi: I is: O the: O capital: I of: I India: O
- None of these

11) Which of the following sentences is an example of a Noun phrase?

1. A crazy white cat
2. By the river
3. The morning flight

- Only 1
- Only 1 and 3
- Only 1 and 2
- Only 2 and 3

12) Which of these would you consider to be a valid limitation of using lesk algorithm for word sense disambiguation?

1. The absence of a certain word can radically change the results.
2. The algorithm determines overlaps only among the glosses of the senses being considered.

- Only 1
- Only 2
- Both 1,2
- None of these

13) You want to visualize the set of word vectors in your dataset. If you are using the StandardScaler in scikit-learn, which of these steps would you perform to increase efficiency of the ML algorithm being applied?

- scaling of the features in your data
- dimensionality reduction
- Trimming of outliers
- None of these

14) Consider the three documents given alongside. The tf-idf score for the word ginger is 0.025 for Document 1. What will the tf-idf score for the word ginger be for Document 3?

Document 1: Rahul likes to have ginger tea with biscuits in the evening.
Document 2: Tea contains caffeine.
Document 3: Ginger is used as a spice. Ginger boosts the immune system.

- 0.05
- 0.005
- 0.025
- 0.25

15) What will be the output of:

Pattern='\w+ed'
String= "He played and won the match when he was injured"
re.search(Pattern,String)

- Played, injured
- Injured
- Played
- It will throw an error

16) Which option is true about the Recursive Descent Parser and Shift reduce parser?

- The recursive descent parser starts from the start symbol and uses production rules to parse the sentence until the last word is parsed
- The shift-reduce parser starts from the start symbol and uses production rules to parse the sentence until the last word is parsed
- The recursive descent parser starts from the sentence and reduces the sentence to a non-terminal symbol until we reach the start symbol of the grammar
- None of the above

17) You are given a paragraph and asked to represent it in the form of a co-occurrence matrix using the 3-skip-2gram technique. To do so, you remove the stopwords, punctuations and retain the words given in the vocabulary resulting in the paragraph given alongside.

In the given context, which of the following co-occurrence pairs are valid for the given paragraph?

Cat jumped narrow hallways cat ran fast behind basement stairs.

- (Cat, jumped) (Cat, narrows) (Cat, hallways) (Cat, cat) (jumped, narrow) (jumped, hallways) (jumped, cat) (jumped, ran) (hallways, cat) (hallways, ran) (hallways, fast) (hallways, behind) (cat, ran) (cat, fast) (cat, behind) (cat, basement) (ran, fast) (ran, behind) (ran, basement) (ran, stairs) (fast, behind) (fast, basement) (fast, stairs) (behind, basement) (behind, basement) (basement, stairs)
- (Cat, jumped) (Cat, narrows) (Cat, hallways) (jumped, narrow) (jumped, hallways) (jumped, cat) (hallways, cat) (hallways, ran) (hallways, fast) (cat, ran) (ran, fast) (fast, behind) (ran, fast) (ran, behind) (ran, basement) (fast, behind) (fast, basement) (fast, stairs) (behind, basement) (behind, basement) (basement, stairs)
- (Cat, jumped) (Cat, narrows) (Cat, hallways) (jumped, narrow) (jumped, hallways) (jumped, cat) (hallways, cat) (hallways, ran) (hallways, fast) (cat, ran) (ran, fast) (fast, behind) (ran, fast) (ran, behind) (ran, basement) (fast, behind) (fast, basement) (fast, stairs)
- None of the above

18) Consider the sentence given below.

"*And now for something completely different*"

Which of these tag sequence is the most appropriate for the given sentence?

- CC And/VB now/IN for/NN something/RB completely/JJ different
- CC And/RB now/IN for/NN something/RB completely/JJ different
- DT And/RB now/IN for/NN something/AD completely/JJ different
- CC And/RB now/IN for/NN something/VBD completely/JJ different

19) You have a string 'flabbergasted' that is misspelled as 'flaburgasted'. You are using the Levenshtein algorithm to calculate the edit distance for these words.

What will the value of the cell where row is 'g' and column is 'l' be? What will the edit distance for the given word be?

- The value of the cell will be 6 and the edit distance will be 3
- The value of the cell will be 8 and the edit distance will be 2
- The value of the cell will be 8 and the edit distance will be 3
- The value of the cell will be 6 and the edit distance will be 2

20) For the following sentence, identify the word2vec skipgram training sets: "*Dog with the bone*"

- (Dog,[with]),(with,[dog,with the]),(the,[with,the bone]),(bone,[the])
- ([with],Dog),([dog,the],with),([with,bone],the),([the],bone)
- ([with,the], Dog), ([dog,the,bone],with), ([dog,with,bone],the), ([dog,with,the],bone)
- None of the above

21) Which of the following define Hidden Markov Model?

1. Emission Probability;
2. Transition probability;
3. Initial state probability;
4. Terminal state probability

- Only 1 & 2
- 1, 2 & 3
- Only 3 & 4
- 1, 2 ,3 & 4

22) Use the following grammar to answer the question:

```
S -> NP VP [1.0]
PP -> IN NP [1.0]
VP -> VB NP [0.3]| VP PP [0.4] | VB [0.3]
NP -> AT N [0.4] | NP PP [0.6]
IN -> 'on'[0.4] | 'with' [0.6]
VB -> 'jumped' [1.0]
N -> 'bed' [0.2]| 'child' [0.4] | 'bottle' [0.4]
AT -> 'the' [0.5] | 'a' [0.5]
```

Sentence: "the child jumped on the bed with a bottle". Which of the following is incorrect parse structure?

```
(S
(NP (AT the) (N child))
(VP
 (VP (VB jumped))
 (PP
  (IN on)
  (NP
   (NP (AT the) (N bed))
   (PP (IN with) (NP (AT a) (N bottle)))))))
```

```
(S
(NP (AT the) (N child))
(VP
(VP
(VP (VB jumped))
(PP (IN on) (NP (AT the) (N bed))))
(PP (IN with) (NP (AT a) (N bottle)))))
```

```
(S
(NP (AT the) (N child))
(VP
(VP (VP (VB jumped)))
(PP (IN on) (NP (AT the) (N bed)))
(PP (IN with) (NP (AT a) (N bottle)))))
```

All of the above

23) Why sequence models (like HMM, CRFs) are preferred over conventional ML models (decision trees, SVMs, naive Bayes etc.) in sequential modelling tasks?

- Sequence models do not use features but rather only use transition probabilities, which helps them reduce unnecessary complexity
- Sequence models learn dependencies among consecutive labels in the sequence, whereas, conventional models consider each observation independently
- Sequence models show lesser bias than conventional models
- All of the above

24) Calculate the pointwise mutual information of "New Delhi" in the following paragraph, based on its occurrences across sentences:

"New Delhi is emerging as the latest growth center in the new and emerging economies of the world, mostly coming from Asia. Delhi is often compared with how New York grew in the early years of the 20th century. New Delhi and Mumbai feature in a list of top 25 global growth centres across the world, in a study conducted by the New Age Economist."

- log(1)
- log(3/2)
- log(⅔)
- log(1/3)

25) Which of the following statements is true for word2vec model?

- It is a prediction based embedding model
- Maximizes probability of generating the context given term
- Number of dimensions of input is same as number of dimensions of output
- None of the above
- All of the above