All the questions in this worksheet have one or more than one correct answers. Choose all the correct options to answer the questions:

1. Which of the following are steps in NLP?

Ans. **A) Lexical Processing B) Syntactic processing** C**) Vectorizer processing** **D) Semantic processing**

2. Which of the following tasks can be completed with only lexical processing?

Ans. **A) Spam- Ham classification**

3. Which of the following are steps in lexical processing?

Ans. A) **Breaking the text in to words**

4. Which of the following tokenizers are available in NLTK?

Ans. A) **word\_tokenize**() B) **sent\_tokenize()**

5. What will be the output of the following lines of code? from nltk.tokenize import word\_tokenize doc = “I love #food #pasta” print( word\_tokenize(doc))

Ans. **B) [“I”, “love”, “#food”, “#pasta”]**

6. What will be the output of the following lines of code? from nltk.tokenize import TweetTokenizer tknz = TweetTokenizer() doc = “I love #food #pasta” print( tknz.tokenize(doc))

Ans. **A) [“I”, “love”, “#”, “food”, “#”, “pasta”]**

7. Which of the following is/ are true regarding to stopwords?

**Ans. A)** They provide no useful information, especially in applications such as spam detector or search engine.

B**) Since the frequency of stopwords is very high, removing stopwords results in a much smaller data.**

**C) removing stopwords results in faster computation.**

8. In which of the NLP tasks we can remove stopwords?

**Ans. A) spam-ham classifier building B) Language Translation task C) Chat- Bot building**

9. which of the following is/are true regarding bag of words model of text?

Ans.. **B) It takes in to consideration both the words present as well as the order of the words.**

**C) It captures the semantics of the text**.

10. Consider the following two documents we create a bow representation using Count Vectorizer of NLTK library. What will the shape of the resultant data? from sklearn.feature\_extraction.text import CountVectorizer Doc1 = "HE love python" Doc2 = "HE love eating healthy" vectrz = CountVectorizer() Bow\_array = vectrz.fit\_transform([Doc1, Doc2]) print(Bow\_array.shape)

Ans. A) (2,3) **B) (2,5)** C) (5,2) D) (5,3)

11. Which of the following are true regarding Tf-Idf?

**Ans** .A) The importance of a word in a document becomes more if it is present exclusively only in this document

**For questions Q12-Q15, Consider the following Documents and answer the Questions**

**Document1: "**Vapour, Bangalore has a really great terrace seating and an awesome view of the Bangalore skyline**"**

**Document2: "**The beer at Vapour, Bangalore was amazing. My favorites are the wheat beer and the ale beer.**"**

**Document3: "**Vapour, Bangalore has the best view in Bangalore."

Please remove the stopwords from the above documents before answering the below questions:

12. What will be the tf-idf score of word “Bangalore” in Document 1?

**Ans. B) 1**

13. What will be the tf-idf score of the word “beer” in document 2?

**Ans .B) 0.89**

14. Which of the following statements are true regarding the above documents?

**Ans.** B) The tf-idf score of “vapour” is less than tf-idf score of “Bangalore” in document 1

D) tf-idf of both “vapour” and “Bangalore” are equal and non-zero

15. Which of the following are advantages of using tf-idf model over BOW model?

**Ans. A) The bow model gives equal importance to all the words while tf-idf model gives more importance to those**

**words in a document which occurs exclusively only I this document .**