WORKSHEET-1

NLP

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?

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?

A) Spam- Ham classification
B) Machine Translation
C) Chat-Bot building
D) Question- Answering system

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

A) Breaking the text in to words
B) Removing Stopwords
C) Top-Down parsing
D) POS-tagging

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

A) word_tokenize()
B) sent_tokenize()
C) list_tokenizer()
D) Random_Tokenizer()

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))
```

A) [“I”, “love”, “#”, “food”, “#”, “pasta”]
B) [“I”, “love”, “#food”, “#pasta”]
C) [“I love”, “#food#pasta”]
D) error

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))
```

A) [“I”, “love”, “#”, “food”, “#”, “pasta”]
B) [“I”, “love”, “#food”, “#pasta”]
C) [“I love ”, “#food#pasta”]
D) error

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

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.
D)None of the above

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

A) spam-ham classifier building
B) Language Translation task
C) Chat- Bot building
D) None of them

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

A) It takes in to consideration of only the words present in the text and not the order of the words.
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.
D) All of the above

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)
```

A) (2,3)
B) (2,5)
C) (5,2)
D) (5,3)

11. Which of the following are true regarding Tf-Idf?
    A) The importance of a word in a document becomes more if it is present exclusively only in this document
    B) The importance of a word in a document becomes more if it is present in every other document also
    C) All the words are treated equally regardless of whether they are present in other documents or not
    D) None of the above

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?
    A) 0.2
    B) 1
    C) 0
    D)1.6
13. What will be the tf-idf score of the word “beer” in document 2?
    A) 0.25
    B) 0.89
    C) 0
    D) 0.159
14. Which of the following statements are true regarding the above documents?
    A) The tf-idf score of “vapour” is greater than tf-idf score of “Bangalore” in document 1
    B) The tf-idf score of “vapour” is less than tf-idf score of “Bangalore” in document 1
    C) tf-idf of both “vapour” and “Bangalore” are equal to zero
    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?
    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 .4
    B) The tf-idf model gives equal importance to all the words in a document regardless of whether that word occurs in other documents or not, while BOW model takes in to consideration whether a word occurs in other documents also.
    C) Both models work on same concept but have different names
    D) None of the above.