Q1) A corpus consisting of only two documents. Given below are two documents and their term (word) count

|  |  |
| --- | --- |
| **Document 1** | |
| **Term** | **Term Count** |
| this | 1 |
| is | 1 |
| a | 2 |
| sample | 1 |

|  |  |
| --- | --- |
| **Document 2** | |
| **Term** | **Term Count** |
| this | 1 |
| is | 1 |
| another | 2 |
| example | 3 |

What is the value of tf–idf for the term "this"?

1. 0
2. log(2)
3. log(3)
4. None of the above

Your Answer: A

Q2) Which of the following are the techniques of keyword normalization?

1. Lemmatization
2. Stemming
3. Both Lemmatization and Stemming
4. Neither Lemmatization nor Stemming

Your Answer: C

Q3) Bag of Words is an example of?

1. Feature scaling technique
2. Feature selection technique
3. Feature extraction technique
4. Feature normalization technique

Your Answer: C

Q4) What is the purpose of topic modeling?

1. Converting text into vectors
2. Clustering the documents
3. Understanding the frequency of term (word)
4. Understanding the importance of term (word)

Your Answer: B

Q5) Word2vec is used to?

1. Create corpus by combining dictionaries
2. Develop a dictionary from documents
3. Generate vectors out of words
4. Train data mining models

Your Answer: C

Q6) Which of the following gate is used to decide how much past information to forget?

1. Reset gate
2. Update gate
3. Forget gate
4. Output gate

Your Answer: A

Q7) Vanishing Gradient is a challenge for which of the following networks?

1. Machine Learning
2. RNN
3. GRU
4. LSTM

Your Answer: B

Q8) Which of the following NLP library provides has a built-in spell checker?

1. NLTK
2. Gensim
3. Spacy
4. All of the above

Your Answer: B

Q9) Which of the following is not a stemmer?

1. Porter
2. Snowball
3. Reverb
4. Lancaster

Your Answer: C

Q10) A document term matrix of the data was created by treating every tweet as one document. Which of the following reduces the dimensionality of that data?

1. Removal of stopwords from the data
2. Converting all the words to lowercase
3. Normalization of words in the data
4. All of the above

Your Answer: D