Enabling People

Education for Future Generations

AI Course

Chapter 7. Quiz

For students

Samsung Innovation Campus

1. Which of the following statements about text mining is incorrect?

   ① You can know the reaction to a specific group.

   ② Information on sales lead can be obtained.

   ③ Competitive strategies can be established by monitoring other brands.

   ④ The same method can be applied to various languages.

Answer: The same method can be applied to various languages.

2. Which of the following is not true about the use of social networks?

   ① It is possible to know how many groups the network is composed of.

   ② It is possible to know influential customers.

   ③ It is possible to know see the change over time.

   ④ It is possible to know if customers will leave next time.

Answer: It is possible to know if customers will leave next time.

3. Which term means separating a stem from a word whose form has been modified to extract the word that is the subject of morpheme analysis?

Answer: Stemming

4. Which term means a set of materials that can show the essential aspects of language as a research material required in each field of language research?

Answer: Corpu

5. Explain about TF (Term Frequency), and write how to calculate TF (formula).

Answer: TF (Term Frequency) measures how frequently a term appears in a document. It indicates the importance of a word within a specific document. Formula: TF(t,d) = (Number of times term t appears in document d) / (Total number of terms in document d)

6. Find the most similar statement using TF-IDF and cosine similarity with reference to the following code.

Answer: To find the most similar statements, calculate: distances = pairwise_distances(X, metric='cosine'), then find the minimum non-diagonal value. The pair with the smallest cosine distance represents the most similar statements.

```
import nltk
import numpy as np
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import pairwise_distances
```

```
doc = [
  "AI is a rapidly advancing field that involves the development of intelligent machines.",
  "Machine learning is a subset of AI that focuses on training machines to learn from data.",
  "Deep learning is a subfield of machine learning that utilizes neural networks with multiple layers.",
  "AI applications are found in various industries, including healthcare, transportation, and entertainment.",
  "Ethical consideration plays an important role in AI development.",
  "AI has the potential to revolutionize many aspects of society.",
  "AI systems such as chatbots and virtual assistants are becoming more common.",
  "Natural language processing is a branch of AI that allows machines to understand human language.",
  "The field of computer vision aims to enable machines to understand visual information.",
]

# pre-processing.
doc = [x.lower() for x in doc]

# parameters
max_features = 18
min_df = 1
max_df = 3
stop_words = stopwords.words('english')

vectorizer = TfidfVectorizer(max_features=max_features,
              min_df=min_df,
              max_df=max_df,
              stop_words=stop_words)
X = vectorizer.fit_transform(doc).toarray()
```