STA 3100 Programming With Data in R: Text Mining

Text Mining in STA 3100 Programming With Data in R

Text mining is a powerful tool for extracting meaningful information from large amounts of unstructured text data. It involves the use of natural language processing (NLP) and machine learning algorithms to identify patterns, trends, and relationships in text data. Text mining can be used to gain insights into customer sentiment, product reviews, and other types of text-based data.

Key Concepts

- Natural Language Processing (NLP): NLP is a field of computer science that focuses on understanding and manipulating human language. It involves the use of algorithms to analyze and interpret text data.

- Machine Learning: Machine learning is a type of artificial intelligence that uses algorithms to learn from data and make predictions. It can be used to identify patterns and trends in text data.

- Text Mining: Text mining is the process of extracting meaningful information from text data using natural language processing and machine learning algorithms.

Coding Examples

Example 1: Text Pre-processing

```
Start of Code
# Load libraries
library(tm)
library(SnowballC)
# Create corpus
text <- c("This is a sample sentence.", "This is another sample sentence.")
corpus <- Corpus(VectorSource(text))
# Clean corpus
corpus <- tm_map(corpus, content_transformer(tolower)) # convert to
lowercase
corpus <- tm_map(corpus, removePunctuation) # remove punctuation
corpus <- tm_map(corpus, removeNumbers) # remove numbers
corpus <- tm_map(corpus, stripWhitespace) # remove extra whitespace
corpus <- tm_map(corpus, stemDocument) # stem words
End of Code
```

Example 2: Document Term Matrix

```
Start of Code
```

```
# Create document term matrix
dtm <- DocumentTermMatrix(corpus)
# Inspect document term matrix
inspect(dtm)
End of Code
```

Practice Multiple Choice Questions

Q1. What is text mining?

A. Text mining is the process of extracting meaningful information from text data using natural language processing and machine learning algorithms.

Q2. What is natural language processing?

A. Natural language processing (NLP) is a field of computer science that focuses on understanding and manipulating human language. It involves the use of algorithms to analyze and interpret text data.