109550198 卜銳凱

HW8
Report

This report describes the analysis of the dataset from file ‘Restaurant_Reviews.tsv’. It generates various graphs to examine and analyze data, and provides insights into the dataset. This report summarizes a Python code snippet for performing text mining and various Natural Language Processing (NLP) tasks using text data. The code demonstrates techniques for text cleaning, tokenization, feature extraction, and text classification.

Code Walkthrough

Importing Libraries:

The code begins by importing key libraries such as NumPy, Matplotlib, and Pandas, which are often used for data manipulation and visualization.

Text manipulation:

A sample text, "The fries were great too," is prepared for demonstrative purposes.
Splitting Text:
The split() method is used to divide the text into separate words, resulting in a list of tokens. Every word is referred to as a token.

List Comprehension:

List comprehensions are used to filter individual words based on attributes such as length or capitalization.

Finding Words:

To locate words that fulfill specified requirements, various methods are used, including:

- Using endswith() to find words that end with "o"
- using startswith() to find words that begin with "w"
- Using istitle() to identify capitalized terms.
- Using in to check for the presence of a letter ("f") in a word ("fries").
- Using isupper() and islower() to determine whether a word is totally uppercase or lowercase.

Checking for Digits:

The isdigit() method is used to determine whether a word is entirely composed of digits.

Finding Letters:

The find () method is used to locate the index of a particular letter ("o") within the text.

Replacing Letters:

Letters ("o") are replaced with other letters ("3") using the replace () method.

Cleaning Text:
A more extensive text cleaning process is demonstrated, where unnecessary white spaces are removed from a text string before splitting it into words.

Conclusion

The Python code snippet supplied demonstrates a variety of text processing and NLP approaches, providing a realistic foundation for working with text data in a variety of applications. It also shows how to preprocess and clean text input, prepare it for machine learning tasks, and train a text classification Naive Bayes classifier.
Discuss possible problems you plan to investigate for future studies

• Clustering(UnsupervisedLearning) • NeuralNetworkandDeepLearning