Class Assignments: Regular Expressions and Text Preprocessing

15.04.2024

1. Create a Python program that reads a text file and prints all the email addresses found in the file using regular expressions. Extend the program to extract and print domain names of the email addresses found.

2. Write a Python program that reads a text file containing a list of phone numbers in various formats (e.g., (123) 456-7890, 123-456-7890, 123.456.7890) and converts them to a standardized format (e.g., 1234567890) using regular expressions. Handle country codes and different variations of phone number formats.

3. Develop a Python script that extracts all the URLs from a given HTML document using regular expressions.

4. Write a Python program that extracts all the hashtags from a given tweet using regular expressions.

5. Collect a research paper containing texts from a specific domain (e.g., medical, legal, technical). Apply stemming and lemmatization to this domain-specific text and analyze how it differs from processing general-purpose text.

6. Develop an interactive Python application using Tkinter or PyQt that allows users to input text and view the stemmed or lemmatized output along with the original words.

7. Write a Python script that continuously monitors a text stream (e.g., news articles from RSS feeds) and performs stemming and lemmatization on new incoming texts in real-time. Display the processed text along with some statistics (e.g., most common stems or lemmas).