**TEXT AND SPEECH ANALYSIS-CCS369**

**Expt 1- Create Regular Expressions in python for detecting word patterns,and tokenizing texts:**

**1)a) Email Address Detection**

**Code:**

import re  
email\_pattern = r'\b[a-zA-Z0-9\_.+-]+@[a-zA-Z0-9-]+(\.[a-zA-Z0-9-]+)+\b'  
text = "Contact us at [hello@srmcet.edu.in](mailto:hello@srmcet.edu.in)"    
emails = re.findall(email\_pattern, text)  
valid\_emails = []  
matches = re.finditer(email\_pattern, text)  
for match in matches:  
    email = match.group()  
    domain = email.split('@')[-1]  
    tld\_parts = domain.split('.')  
    if len(tld\_parts[-1]) >= 2:  
        valid\_emails.append(email)  
for valid\_email in valid\_emails:  
    print(valid\_email)

**Output:**

h[ello@srmcet.edu.in](mailto:hello@srmcet.edu.in" \t "_blank)

### 1)b) ****URL Detection****

**Code:**

import re

url\_pattern = r'https?://(?:www\.)?[a-zA-Z0-9./?=\_-]+'

text = "Visit [https://www.openai.com](https://www.openai.com/) or [http://example.com](http://example.com/)"

urls = re.findall(url\_pattern, text)

print(urls)

**Output:**

['https://www.openai.com', 'http://example.com']

### 1)c) ****Date Detection:****dd****,****mm****,****yy****,****yyyy****Format****

### ****Code:****

### import re

### text = "Today's date is 04/08/2025 or 04-08-25"

### date\_pattern = r'\d{2}[-/]\d{2}[-/]\d{2,4}'

### dates = re.findall(date\_pattern, text)

### simple\_dates = [f"{d[:2]}/{d[3:5]}/{d[-2:]}" for d in dates]

### print(simple\_dates)

### Output:

### ['04/08/25', '04/08/25']

### 1)d) ****Tokenize Sentence into Words****

**Code:**

import re

sentence = "Hi girl! I'm learning Python, and it's awesome."

words = re.findall(r'\b\w+\b', sentence)

print(words)

**Output:**

['Hi', 'girl', 'I', 'm', 'learning', 'Python', 'and', 'it', 's', 'awesome']

### 1)e) ****Tokenize Paragraph into Sentences****

**Code:**

import re

paragraph = "Hello there! How are you doing? This is a test. Let's split this paragraph."

sentences = re.split(r'(?<=[.!?])\s+', paragraph)

print(sentences)

**Output:**

['Hello there!', 'How are you doing?', 'This is a test.', "Let's split this paragraph."]