SCS4209 / IS4108 / CS4113 - Natural Language Processing
Lab session - 23rd of June 2021

1. Describe the class of strings that is matched by the following regular expressions.
    a. [a-zA-Z]+
    b. [A-Z] [a-z] *
    c. p [aeiou] {,2} t
    d. \d+ (\ . \d+) ?
    e. ([^aeiou] [aeiou] [^aeiou]) *
    f. \w+ | [^\w\s]+
    g. [aeiou]{2,5}

    Test your answers using **nltk.re_show()**.

2. Write regular expressions to match the following classes of strings.
    a. A single determiner. (Consider **a, an, the** as the determiners)
    b. An arithmetic expression using integers, addition, and multiplication such as 3*4+9.

3. Copy and paste some text from an online news article. Apply the following tokenizations.
    a. Use **nltk.regexp_tokenize()** to create a tokenizer that tokenizes the various kinds of punctuations in this text. Use one multi-line expression, with inline comments, using the verbose flag (?x) when creating the regular expressions.
    b. Use **nltk.regexp_tokenize()** to create a tokenizer that tokenizes the following kinds of expressions.
        i. Monetary amounts
        ii. Dates
        iii. Names of People and organizations
    c. Use **nltk.regexp_tokenize()** to select the capitalized words in the selected text.

4. What do you understand by the terms “**Collocations**” and “**Bigrams**”? Explain briefly.
    a. Apply the **collocations()** function of nltk to the **text5 (Chat corpus)** and the **text7 (Wall Street Journal)** of nltk’s **book** module.

5. **International Phonetic Alphabet(IPA)** is a set of symbols intended as a universal system for transcribing speech sounds.
    a. Try to write your name using the IPA for Sinhala.
    b. Write the following words using the IPA.
        i. University
        ii. Computer
        iii. Information