

Assignment-1

Natural Language Processing

Instructor: Muhammad Umair

Points: 10

Deadline: 14-09-2023

Instructions:

- Late submissions would not be entertained.
- Plagiarized work would not be considered.
- Use Python as a programming language to demonstrate the assignment work.
- Submit PDF print of Jupyter Notebook with an output of Regex operations on sample text string Named StudentName_Reg#
- Submit 2nd PDF of handwritten Regex operations named H_StudentName_Reg#

Q1: Write regular expressions for the following languages.

1. the set of all alphabetic strings;
2. the set of all lower-case alphabetic strings ending in a b;
3. the set of all strings from the alphabet a, b such that each a is immediately preceded by and immediately followed by a b;

Q2: Write regular expressions for the following languages. By “word”, we mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth.

1. the set of all strings with two consecutive repeated words (e.g., “Humbert Humbert” and “the the” but not “the bug” or “the big bug”);
2. all strings that start at the beginning of the line with an integer and that end at the end of the line with a word;
3. all strings that have both the word grotto and the word raven in them (but not, e.g., words like grottos that merely contain the word grotto);
4. Demonstrate a pattern to check if the email address is valid.
5. Demonstrate a pattern to verify the validity of a phone number, and check if it belongs to the Pakistani Mobile Network.
6. Demonstrate a RE to remove symbols and non-alphanumeric characters from the string paragraph.
7. Demonstrate a RE to remove URLs and HTML tags from the text string.

Reference Material:

- [Google Colaboratory as Python IDE](#)
- [Regular Expression Documentation](#)
- [W3School for Sample RE Implementation](#)

- [Build and Test RE Online](#)