



SNS COLLEGE OF ENGINEERING



Kurumbapalayam(Po), Coimbatore – 641 97

Accredited by NAAC-UGC with 'A' Grade

Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

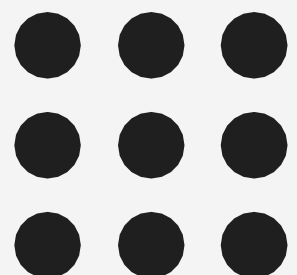
Department of Artificial Intelligence and Data Science

Course Name – 19AD601 – Natural Language
Processing

III Year / VI Semester

Unit 1 – Introduction

Topic 3- Regular Expression





Regular Expression



Regular Expression

- Regular expression (often shortened to regex), a language for specifying text search expression strings.
- This practical language is used in every computer language, word processor, and text processing tools like the Unix tools grep or Emacs.
- Formally, a regular expression is an algebraic notation for characterizing a set of strings. Regular expressions are particularly useful for searching in texts, when we have a pattern to search corpus for and a corpus of texts to search through.
- A regular expression search function will search through the corpus, returning all texts that match the pattern.
- The corpus can be a single document or a collection. For example, the Unix command-line tool grep takes a regular expression and returns every line of the input document that matches the expression.



Regular Expression



Basic Regular Expression Patterns

- The simplest kind of regular expression is a sequence of simple characters; putting concatenation characters in sequence is called concatenation.
- To search for woodchuck, we type `/woodchuck/`. The expression `/Buttercup/` matches any string containing the substring Buttercup.
- Regular expressions are case sensitive; lower case `/s/` is distinct from upper case `/S/` (`/s/` matches a lower case s but not an upper case S).
- This means that the pattern `/woodchucks/` will not match the string Woodchucks. We can solve this problem with the use of the square braces `[and]`.
- The string of characters inside the braces specifies a disjunction of characters to match



Regular Expression



- The regular expression `/[1234567890]/` specifies any single digit.
- While such classes of characters as digits or letters are important building blocks in expressions, they can get awkward (e.g., it's inconvenient to specify to mean “any capital letter”).
- `/[ABCDEFGHIJKLMNOPQRSTUVWXYZ]/`
- In cases where there is a well-defined sequence associated with a set of characters, the brackets can be used with the dash (-) to specify range any one character in a range.



THANK YOU