# Notes on Regular Expressions (Chapter 2.1 - 2.1.5)

## Overview

ELIZA, a chatbot developed in 1966. ELIZA used simple pattern-matching to mimic a Rogerian psychotherapist. This basic pattern rules can be effective in certain contexts.

## 2.1 Regular Expressions

Regular expressions (regex) are powerful tools for pattern matching in text. They are useful for extracting data (e.g., prices), text normalization, tokenization, and intent recognition in NLP.

## 2.1.1 Basic Regular Expression Patterns

- Simple matching: /woodchuck/  
- Character sets: /[abc]/ matches a, b, or c  
- Ranges: /[A-Z]/ for uppercase, /[a-z]/ for lowercase  
- Negation: /[^a]/ matches any character except 'a'  
- Optionality: /colou?r/ matches 'color' and 'colour'  
- Repetition: \* (zero or more), + (one or more), ? (zero or one)  
- Wildcard: . matches any character  
- Anchors: ^ (start), $ (end), \b (word boundary)

## 2.1.2 Disjunction, Grouping, and Precedence

- Disjunction (|): /cat|dog/ matches either 'cat' or 'dog'  
- Grouping: () used to group patterns  
- Example: /gupp(y|ies)/ matches 'guppy' or 'guppies'  
- Operator precedence: (), \*, +, ?, sequences, | (in order)  
- Greedy matching: by default matches longest possible string  
- Non-greedy: \*? or +?

## 2.1.3 A Simple Example

- Goal: match word 'the' (case-insensitive, whole word only)  
- Regex: /\b[tT]he\b/  
- \b ensures word boundary, [tT] handles case  
- Avoids false matches like 'other' or 'there'

## 2.1.4 More Operators

- Shortcuts:  
 \d = digit [0-9], \w = word character [a-zA-Z0-9\_], \s = whitespace  
 \D, \W, \S are negations  
- Quantifiers:  
 {n} = exactly n, {n,} = n or more, {n,m} = between n and m  
- Escaping special characters: use backslash \., \\*, \?, etc.

## 2.1.5 A More Complex Example

- Example: match prices like '$999.99'  
- Regex: (ˆ|\W)\$[0-9]{0,3}(\.[0-9][0-9])?\b  
- Matches dollar sign followed by up to 3 digits and optional cents  
  
- Example: match disk space like '500 GB'  
- Regex: \b[0-9]+(\.[0-9]+)? \*(GB|[Gg]igabytes?)\b  
- Matches whole/decimal numbers with GB or gigabytes