CADT

Cambodia Academy of Digital Technology

Text Processing

Lecturer: Soklay HENG

Academic Year: 2024-2025

Python Environment and Libraries

- Python Review (from Standford CS231n by Justin Johnson): https://cs231n.github.io/python-numpy-tutorial/
- Python Environment: Google Colab (https://colab.research.google.com)
- Libraries:
  - Regular Expression Operations (re) https://docs.python.org/3/library/re.html
  - Natural Language Toolkits (NLTK) https://www.nltk.org/

Regular Expressions:

"A formal language for specifying text strings"

Regular Expressions

- Character Classes

Pattern	Matches
.	any character except newline
\w \d \s	word, digit, whitespace
\W \D \S	not word, digit, whitespace
[abc]	either a, b, or c
[^abc]	not a, b, or c
[a-g]	character between a & g

Regular Expressions

- Quantifiers and Alternations

Pattern	Matches
`a* a+ a?`	0 or more, 1 or more, 0 or 1
`a{5} a{2,}`	exactly five, two or more
`a{1,3}`	between one & three
`a+? a{2,}?`	match as few as possible
ab|cd	ab or cd

Regular Expressions

- Escaped Character

Pattern	Matches
`\.  \*  \\`	escaped special characters
`\t  \n  \r`	tab, linefeed, carriage return

- Anchors

Pattern	Matches
`^abc$`	start / end of the string
`\b`	word boundary

Regular Expressions

- Groups and Lookaround

Pattern	Matches
`(abc)`	capture group (useful with **replace**)
`\1`	backreference to group #1
`(?:abc)`	non-capturing group
`(?=abc)`	positive lookahead
`(?!abc)`	negative lookahead

Exercise:

- Write a regular expression to match each of the following patterns:

1. Punctuation
2. String of letters whose length is at most 3
3. String of digits whose length is at least 3
4. String of word characters containing at least one a and one b
5. Anything enclosed by square brackets
6. String of word characters whose the 5th character from the right end is a digit
7. Date format: yyyy-mm-dd

Exercise

8. String of 10 digits that starts and ends with the same 3- digit sequence
9. Password satisfying the following conditions:
   - at least 1 lowercase letter
   - at least 1 uppercase letter
   - at least 1 digit
   - six characters or more

Any Questions?