NLP regular expressions

Joanna Kolodziejczyk

2022

1 Exercise 1

Work with the following text:

```
  10 WORK \t lecture
11 PRIV    Breakfast
12   WORK  laboratories
```

Write regular expressions for the following:

1. extract all hours
2. the starting and ending positions of the first occurrence of the hour in text
3. remove the extra spaces and put all the words in one single line.
4. extract all codes (text in capitals)
5. extract all tasks
6. define the text pattern groups and extract them all (groups are inside ())

2 Exercise 2

Work with the following text:

```
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
```

Write regular expressions for the following:

1. extract all tags
2. extract all end tags
3. extract all text inside tags

3 Exercise 3

This exercise comes from the textbook (ex 2.2 pp.28) `https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf`

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. write a pattern that places the first word of an English sentence in a register. Deal with punctuation.