DS 100: Principles and Techniques of Data Science Date: September 19, 2018

Discussion #4

Name:

Regular Expressions

1. Which strings contain a match for the following regular expression, `"1+1$"`? The character `"␣"` represents a single space.

   ◯ `What␣is␣1+1` √ `Make␣a␣wish␣at␣11:11` ◯ `111␣Ways␣to␣Succeed`

   Solution: Recall that `1+` matches on *at least one* occurrence of the character `1`, and `$` marks the end of the string.

2. Given the text,

   ```
   "<record>␣Fernando␣Perez␣<fperez@berkeley.edu>␣Faculty␣</record>"
   "<record>␣Edward␣Fang␣<edward.fang@berkeley.edu>␣TA␣</record>"
   ```

   Which of the following matches exactly to the email addresses (including angle brackets)?

   ◯ `<.*@.*>` √ `<[^>]*@[^>]*>` ◯ `<.*@\w+\..*>`

   Solution: Greediness matches too much in the first and third choices.

3. For each pattern specify the starting and ending position of the first match in the string. The index starts at zero and we are using closed intervals (both endpoints are included).

	`abcdefg`	`abcs!`	`ab␣abc`	`abc,␣123`
`abc*`	$[0, 2]$	$[0, 2]$	$[0, 1]$	$[0, 2]$
`[^\s]+`	$[0, 6]$	$[0, 4]$	$[0, 1]$	$[0, 3]$
`ab.*c`	$[0, 2]$	$[0, 2]$	$[0, 5]$	$[0, 2]$
`[a-z1,9]+`	$[0, 6]$	$[0, 3]$	$[0, 1]$	$[0, 3]$

4. Write a regular expression that matches strings (including the empty string) that only contain lowercase letters and numbers.

Solution:

```
^[a-z0-9]*$
```

5. Write a regular expression that matches strings that contain exactly 5 vowels.

Solution:

```
^([^aeiouAEIOU]*[aeiouAEIOU]){5}[^aeiouAEIOU]*$
```

6. Given that `address` is a string, use `re.sub` to replace all vowels with a lowercase letter "o". For example `"123␣Orange␣Street"` would be changed to `"123␣orongo␣Stroot"`.

Solution:

```
re.sub(r"[aeiuAEIOU]",  "o", address)
```

7. Given `dates = "October␣10,␣November␣11,␣December␣12,␣January␣1"`, use `re.findall` to extract all the numbers in the string. The result should look like `["10", "11", "12", "1"]`.

Solution:

```
re.findall(r"[0-9]+",  dates)
```

8. Complete the following grid as you would a crossword puzzle. The clues are given as regular expressions on the top and the left. Each of the nine cells should be filled with exactly one letter such that each ROW and COLUMN taken as a whole match their corresponding regular expressions. Each box should only contain one letter.

	`[JUNDT]*`	`APA|OPI|OLK`	`(NA|FE|HE)[CV]`
`[DEF][MNO]*`	D	O	N
`[^DJNU]P[ABC]`	T	P	A
`[ICAN]*`	N	I	C