# Regex with Pandas and Named Groups

September 26, 2020

---

*You are currently looking at **version 1.0** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the Jupyter Notebook FAQ course resource.*

---

## 1 Working with Text Data in pandas

```
In [ ]: import pandas as pd

        time_sentences = ["Monday: The doctor's appointment is at 2:45pm.",
                          "Tuesday: The dentist's appointment is at 11:30 am.",
                          "Wednesday: At 7:00pm, there is a basketball game!",
                          "Thursday: Be back home by 11:15 pm at the latest.",
                          "Friday: Take the train at 08:10 am, arrive at 09:00am."]

        df = pd.DataFrame(time_sentences, columns=['text'])
        df
```

```
In [ ]: # find the number of characters for each string in df['text']
        df['text'].str.len()
```

```
In [ ]: # find the number of tokens for each string in df['text']
        df['text'].str.split().str.len()
```

```
In [ ]: # find which entries contain the word 'appointment'
        df['text'].str.contains('appointment')
```

```
In [ ]: # find how many times a digit occurs in each string
        df['text'].str.count(r'\d')
```

```
In [ ]: # find all occurances of the digits
        df['text'].str.findall(r'\d')
```

```
In [ ]: # group and find the hours and minutes
        df['text'].str.findall(r'(\d?\d):(\d\d)')
```

```
In [ ]: # replace weekdays with '???'
        df['text'].str.replace(r'\w+day\b', '???')
```

```
In [ ]: # replace weekdays with 3 letter abbrevations
        df['text'].str.replace(r'(\w+day\b)', lambda x: x.groups()[0][:3])
```

```
In [ ]: # create new columns from first match of extracted groups
        df['text'].str.extract(r'(\d?\d):(\d\d)')
```

```
In [ ]: # extract the entire time, the hours, the minutes, and the period
        df['text'].str.extractall(r'((\d?\d):(\d\d) ?([ap]m))')
```

```
In [ ]: # extract the entire time, the hours, the minutes, and the period with group names
        df['text'].str.extractall(r'(?P<time>(?P<hour>\d?\d):(?P<minute>\d\d) ?(?P<period>[ap]m)
```