# Working With Text

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

```
In [ ]: text1 = "Ethics are built right into the ideals and objectives of the United Nations "

        len(text1) # The length of text1
```

```
In [ ]: text2 = text1.split(' ') # Return a list of the words in text2, separating by ' '.

        len(text2)
```

```
In [ ]: text2
```

List comprehension allows us to find specific words:

```
In [ ]: [w for w in text2 if len(w) > 3] # Words that are greater than 3 letters long in text2
```

```
In [ ]: [w for w in text2 if w.istitle()] # Capitalized words in text2
```

```
In [ ]: [w for w in text2 if w.endswith('s')] # Words in text2 that end in 's'
```

We can find unique words using `set()`.

```
In [ ]: text3 = 'To be or not to be'
        text4 = text3.split(' ')

        len(text4)
```

```
In [ ]: len(set(text4))
```

```
In [ ]: set(text4)
```

```
In [ ]: len(set([w.lower() for w in text4])) # .lower converts the string to lowercase.
```

```
In [ ]: set([w.lower() for w in text4])
```

### 1.0.1 Processing free-text

```
In [ ]: text5 = '"Ethics are built right into the ideals and objectives of the United Nations" \
        #UNSG @ NY Society for Ethical Culture bit.ly/2guVelr'
        text6 = text5.split(' ')

        text6
```

Finding hastags:

```
In [ ]: [w for w in text6 if w.startswith('#')]
```

Finding callouts:

```
In [ ]: [w for w in text6 if w.startswith('@')]
```

```
In [ ]: text7 = '@UN @UN_Women "Ethics are built right into the ideals and objectives of the Uni
        #UNSG @ NY Society for Ethical Culture bit.ly/2guVelr'
        text8 = text7.split(' ')
```

We can use regular expressions to help us with more complex parsing.

For example `'@[A-Za-z0-9_]+'` will return all words that: * start with `'@'` and are followed by at least one: * capital letter (`'A-Z'`) * lowercase letter (`'a-z'`) * number (`'0-9'`) * or underscore (`'_'`)

```
In [ ]: import re # import re - a module that provides support for regular expressions

        [w for w in text8 if re.search('@[A-Za-z0-9_]+', w)]
```