Text Files in Python

The first two minutes of the video below are dedicated to the ***glob*** library, which makes opening files with similar path structure (like our folder of Roger Ebert review text files) simple.

Quiz

So we have 88 Roger Ebert reviews to open and read, which you can see in the Jupyter Notebook dashboard below (click *jupyter* in the top lefthand corner to access the dashboard) in the *ebert_reviews* folder. If you want to work outside of the Udacity classroom, click **this link** to download a zipped version of that folder.

We'll need to loop to iterate through all of the files in this folder to open and read each, then extract the bits of text that we need as separate pieces of data:

- the first line, which is the movie title (to merge to the master dataset with)
- the second line, which is the review URL (not necessary for the word cloud but nice to have)
- everything from the third line onwards, which is the review text

- ~~Creates an empty list, *df_list*, to which dictionaries will be appended. This list of~~ dictionaries will eventually be converted to a pandas DataFrame (this is the **most efficient way of building a DataFrame row by row**).
- Loops through each movie's Roger Ebert review text file in the *ebert_reviews* folder.
- Opens each text file using a path generated by **glob** and passes it into a file handle called *file*.
- Creates a DataFrame called *df* by converting *df_list* using the **`pd.DataFrame` constructor**.

Your task is to extract the movie title, Roger Ebert review URL, and the review in each text file and append each trio as a dictionary to *df_list*.

The file methods required for this task are:

- `readline()`
- `read()`

Quiz

In [2]:
```
import glob
import pandas as pd
```

In [9]:
```
# List of dictionaries to build file by file and later convert to a
df_list = []
for ebert_review in glob.glob('ebert_reviews/*.txt'):
    with open(ebert_review, encoding='utf-8') as file:
        title = file.readline()[:-1]
        review_url = file.readline()[:-1]
        review_text = file.read()
        # Append to list of dictionaries
        df_list.append({'title': title,
                        'review_url': review_url,
                        'review_text': review_text})
df = pd.DataFrame(df_list, columns = ['title', 'review_url', 'revie
```

Solution Test

Run the cell below the see if your solution is correct. If an `AssertionError` is thrown, your solution is incorrect. If no error is thrown, your solution is correct.

In [10]:
```
df_solution = pd.read_pickle('df_solution.pkl')
pd.testing.assert_frame_equal(df, df_solution)
```

```
---------------------------------------------------------------------
----------
AssertionError                            Traceback (most recent
 call last)
<ipython-input-10-3734e7da8818> in <module>()
      1 df_solution = pd.read_pickle('df_solution.pkl')
----> 2 pd.testing.assert_frame_equal(df, df_solution)

/opt/conda/lib/python3.6/site-packages/pandas/util/testing.py in
assert_frame_equal(left, right, check_dtype, check_index_type, ch
eck_column_type, check_frame_type, check_less_precise, check_name
s, by_blocks, check_exact, check_datetimelike_compat, check_categ
```

MENU EXPAND

Solution

More Information

- Stack Overflow: Best Practices for Opening Files in Python
- Stack Overflow: The Correct, Fully Pythonic Way to Read a File
- Stack Overflow: Iterables and Iterators
- Wikipedia: Glob programming

NEXT