Documentation

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Adam Spannbauer
Machine Learning Engineer at Eastman

Unit testing

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Adam Spannbauer
Machine Learning Engineer at Eastman

Final Thoughts

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Adam Spannbauer
Machine Learning Engineer at Eastman

Readability counts

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Adam Spannbauer
Machine Learning Engineer

Documentation & testing in practice

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Adam Spannbauer
Machine Learning Engineer at Eastman

Documentation in Python

- Comments

```
# Square the number x
```

  - Docstrings

```
    """Square the number x

    :param x: number to square
    :return: x squared

    >>> square(2)
    4
    """
```

Looking Back

- Modularity

```
def function()
    ...
```

```
class Class:
    ...
```

Why testing?

- Confirm code is working as intended
- Ensure changes in one function don't break another
- Protect against changes in a dependency

Documenting projects with Sphinx

text_analyzer

Navigation

Classes

Utility Functions

Quick search

Go

Classes

class `text_analyzer.`**`Document`**(*text*)

Analyze text data

Parameters: **text** – text to analyze

Variables:
- **text** – Contains the text originally passed to the instance on creation
- **tokens** – Parsed list of words from `text`
- **word_counts** – `Counter` object containing counts of hashtags used in text

`plot_counts`(*attribute='word_counts', n_most_common=5*)

Plot most common elements of a `collections.Counter` instance attribute

Parameters:
- **attribute** – name of `Counter` attribute to use as object to plot
- **n_most_common** – number of elements to plot (using `Counter.most_common()`)

Returns: `None`; a plot is shown using `matplotlib`

```
>>> doc = Document("duck duck goose is fun")
>>> doc.plot_counts('word_counts', n_most_common=5)
```

The Zen of Python

```python
import this
```

```
The Zen of Python, by Tim Peters (abridged)

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
The complex is better than complicated.
Readability counts.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
```

Descriptive naming

Poor naming

```
def check(x, y=100):
    return x >= y
```

Descriptive naming

```
def is_boiling(temp, boiling_point=100):
    return temp >= boiling_point
```

Going overboard

```
def check_if_temperature_is_above_boiling_point(
        temperature_to_check
```

Comments

```
# This is a valid comment
x = 2
```

```
y = 3  # This is also a valid comment
```

```
# You can't see me unless you look at the source code

# Hi future collaborators!!
```

Documenting classes

```
class Document:
    """Analyze text data

    :param text: text to analyze

    :ivar text: text originally passed to the instance on creation
    :ivar tokens: Parsed list of words from text
    :ivar word_counts: Counter containing counts of hashtags used in text
    """
    def __init__(self, text):
        ...
```

Testing in Python

- `doctest`
- `pytest`

Looking Back

- Modularity
- Documentation

```
# Comments
```

```
"""docstrings"""
```

SPHINX
Python Documentation Generator

Effective comments

Commenting 'what'

```
# Define people as 5
people = 5

# Multiply people by 3
people * 3
```

Commenting 'why'

```
# There will be 5 people attending the party
people = 5

# We need 3 pieces of pizza per person
```

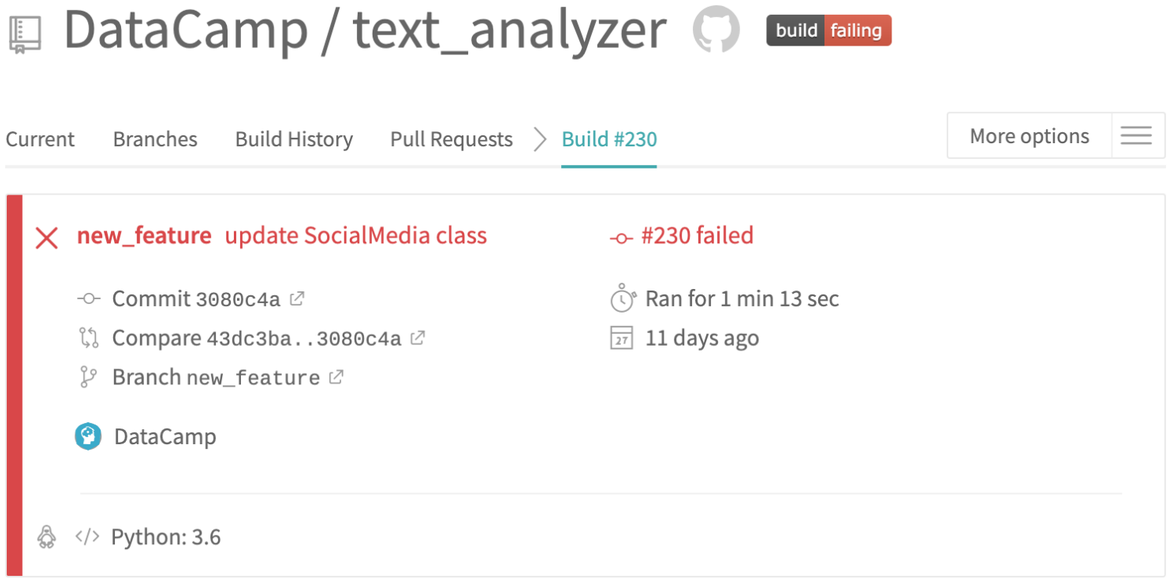
Looking Back

- Modularity
- Documentation
- Automated testing

```
def f(x):
    """
    >>> f(x)
    expected output
    """
    ...
```

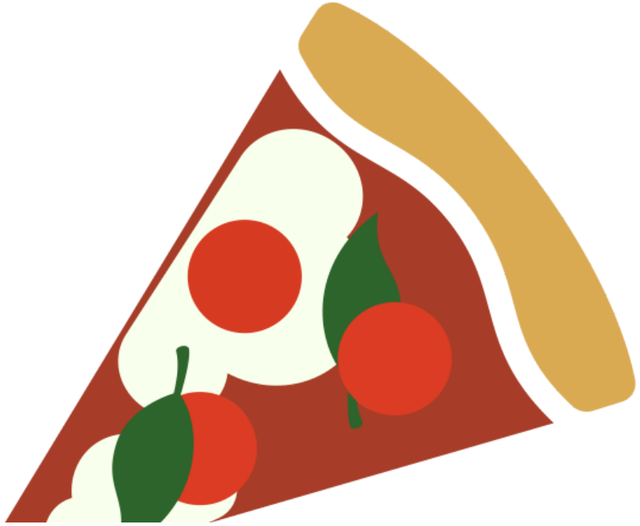
pytest

Continuous integration testing



Keep it simple

```
The Zen of Python, by Tim Peters (abridged)

Simple is better than complex.
Complex is better than complicated.
```



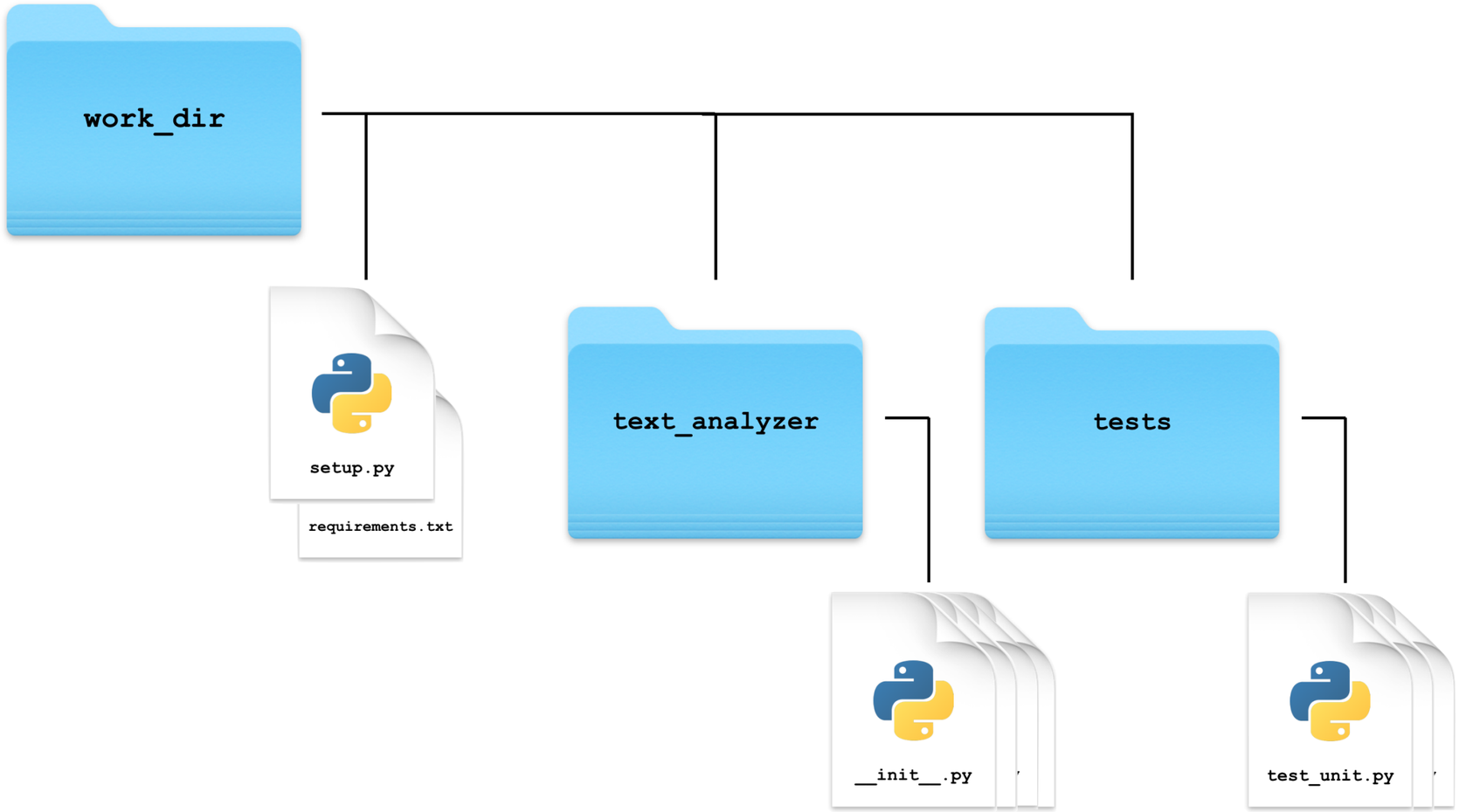
Using doctest

```
def square(x):
    """Square the number x

    :param x: number to square
    :return: x squared

    >>> square(3)
    9
    """
    return x ** x


import doctest
doctest.testmod()
```

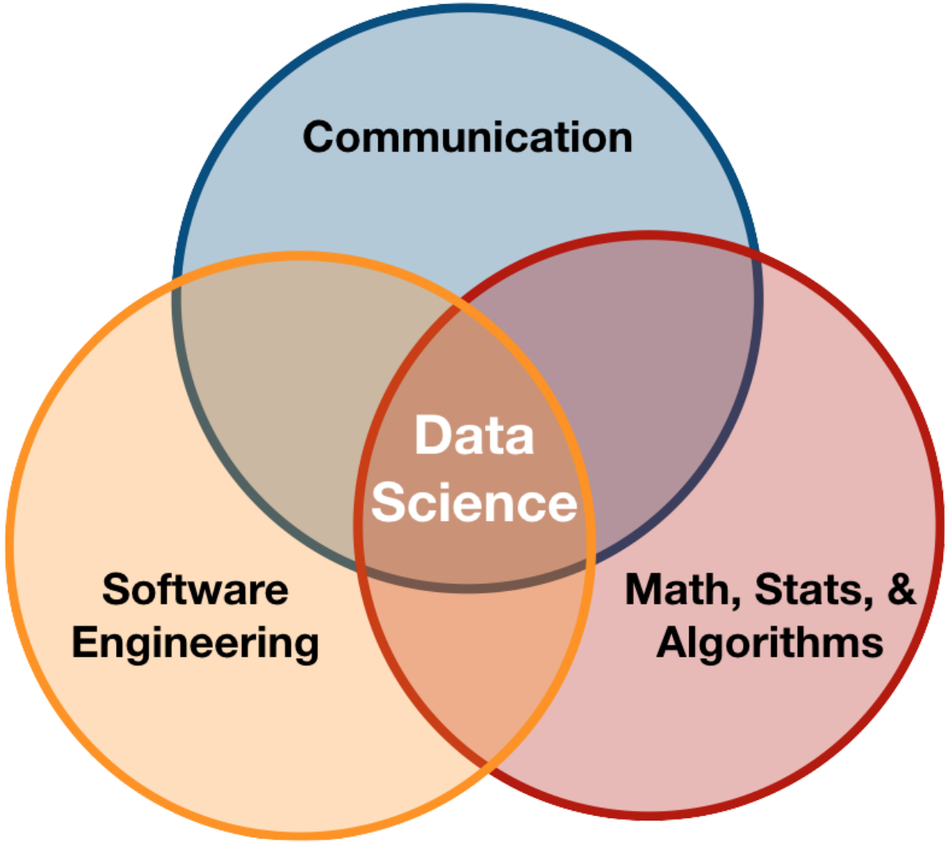
Making a pizza - complex

```
def make_pizza(ingredients):
    # Make dough
    dough = mix(ingredients['yeast'],
                ingredients['flour'],
                ingredients['water'],
                ingredients['salt'],
                ingredients['shortening'])

    kneaded_dough = knead(dough)
    risen_dough = prove(kneaded_dough)

    # Make sauce
    sauce_base = sautee(ingredients['onion'],
                                ingredients['garlic'],
```

pytest structure



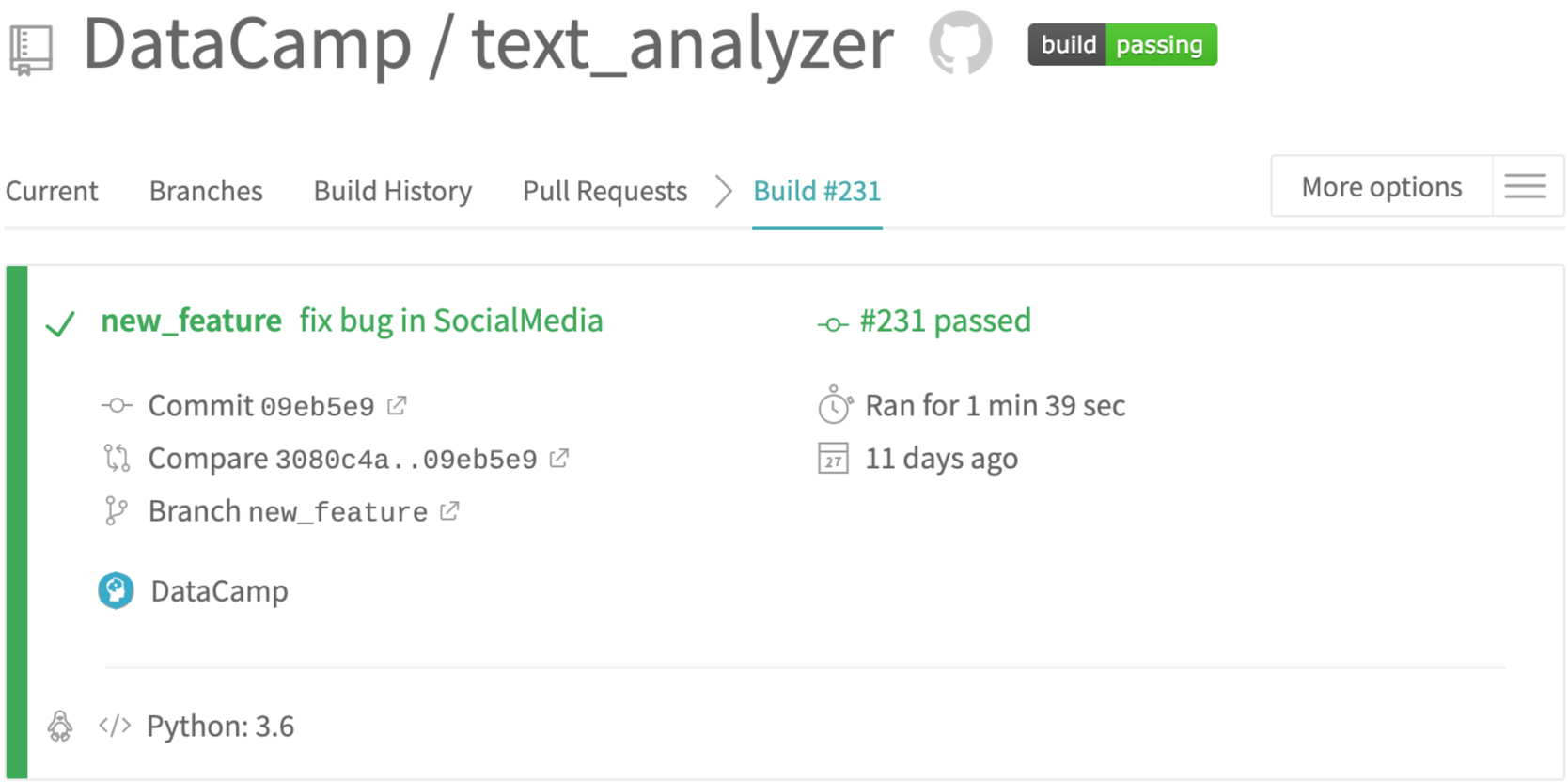
Data Science & Software Engineering



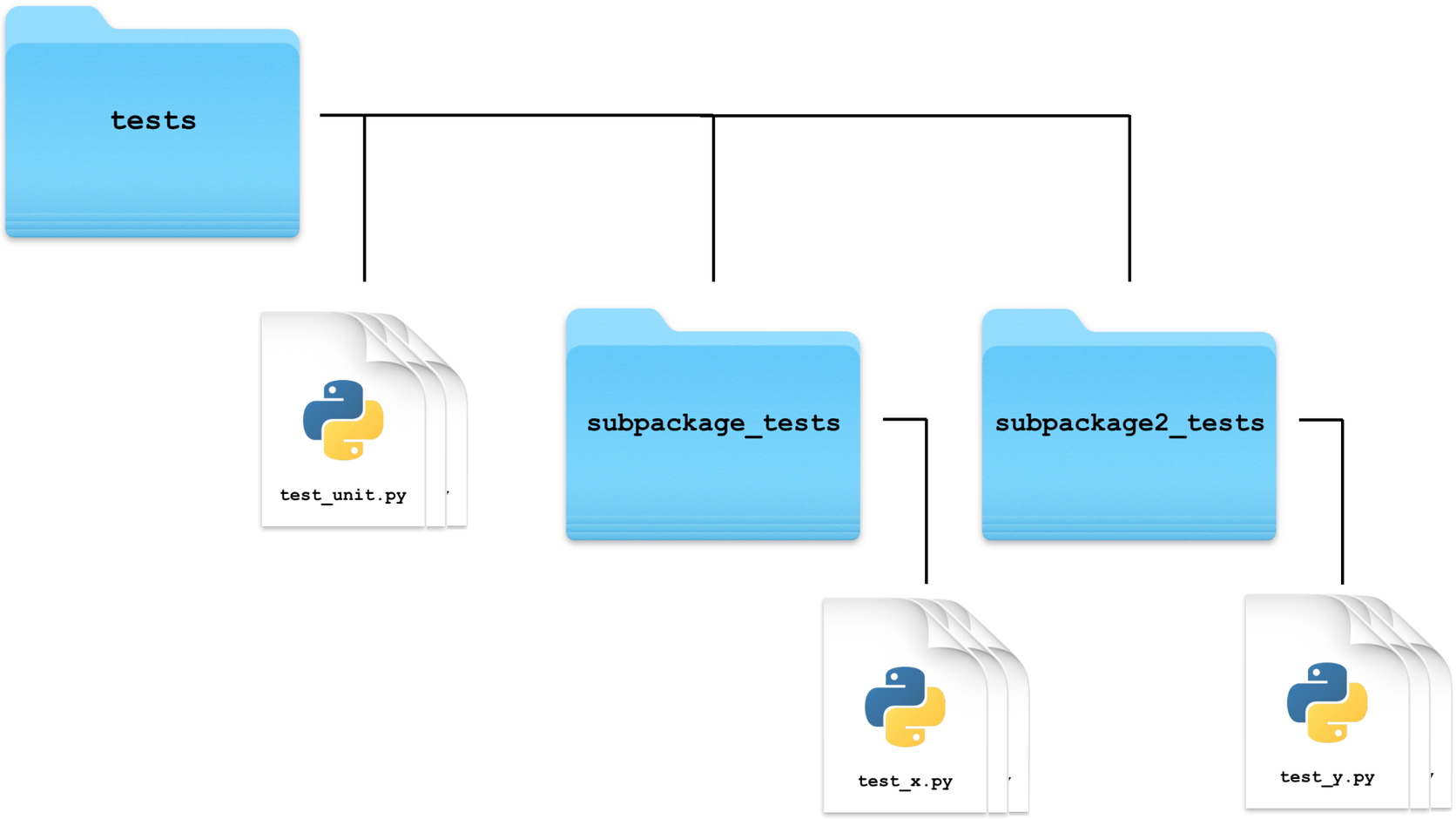
Docstrings

```
def function(x):
    """High level description of function

    Additional details on function
```

Continuous integration testing



pytest structure



Docstrings

```
def function(x):
    """High level description of function

    Additional details on function

    :param x: description of parameter x
    :return: description of return value
```

Example webpage generated from a docstring in the Flask package.

Links and additional tools

- **Sphinx** - Generate beautiful documentation
- **Travis CI** - Continuously test your code
- **GitHub** & **GitLab** - Host your projects with git
- **Codecov** - Discover where to improve your projects tests
- **Code Climate** - Analyze your code for improvements in readability

Good Luck!

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Making a pizza - simple

```
def make_pizza(ingredients):
    dough = make_dough(ingredients)
    sauce = make_sauce(ingredients)
    assembled_pizza = assemble_pizza(dough, sauce, ingredients)

    return bake(assembled_pizza)
```

Let's Practice

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Writing unit tests

working in `workdir/tests/test_document.py`

```python
from text_analyzer import Document


# Test tokens attribute on Document object
def test_document_tokens():
    doc = Document('a e i o u')

    assert doc.tokens == ['a', 'e', 'i', 'o', 'u']


# Test edge case of blank document
def test_document_empty():
    doc = Document('')
```

Docstrings

```
def function(x):
    """High level description of function

    Additional details on function

    :param x: description of parameter x
    :return: description of return value

    >>> # Example function usage
    Expected output of example function usage
    """
    # function code
```

When to refactor

- Function definition not fitting on screen
- Separable processes in single function
- Can't think of a good meaningful name for a function

Let's Practice

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Example docstring

```
def square(x):
    """Square the number x

    :param x: number to square
    :return: x squared

    >>> square(2)
    4
    """
    # `x * x` is faster than `x ** 2`
    # reference: https://stackoverflow.com/a/29055266/5731525
    return x * x
```

Writing unit tests

```
# Create 2 identical Document objects
doc_a = Document('a e i o u')
doc_b = Document('a e i o u')

# Check if objects are ==
print(doc_a == doc_b)

# Check if attributes are ==
print(doc_a.tokens == doc_b.tokens)
print(doc_a.word_counts == doc_b.word_counts)
```

```
False
True
True
```

Running pytest

working with `terminal`

```
datacamp@server:~/work_dir $ pytest
```

```
collected 2 items

tests/test_document.py ..

========== 2 passed in 0.61 seconds ==========
```

Example docstring output

```
help(square)
```

```
square(x)
    Square the number x

    :param x: number to square
    :return: x squared

    >>> square(2)
    4
```

Running pytest

working with `terminal`

```
datacamp@server:~/work_dir $ pytest tests/test_document.py
```

```
collected 2 items

tests/test_document.py ..

========== 2 passed in 0.61 seconds ==========
```

Let's Practice

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON

Failing tests

working with `terminal`

```
datacamp@server:~/work_dir $ pytest
```

```
collected 2 items

tests/test_document.py F.

============== FAILURES ==============
________ test_document_tokens ________

def test_document_tokens(): doc = Document('a e i o u')

assert doc.tokens == ['a', 'e', 'i', 'o']
E AssertionError: assert ['a', 'e', 'i', 'o', 'u'] == ['a', 'e', 'i', 'o']
```

Let's Practice

SOFTWARE ENGINEERING FOR DATA SCIENTISTS IN PYTHON