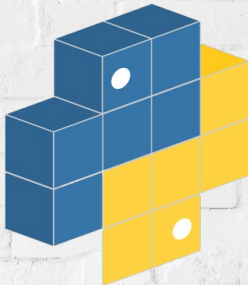


python modules and packages



André Santos, afs@inesctec.pt

Installing packages

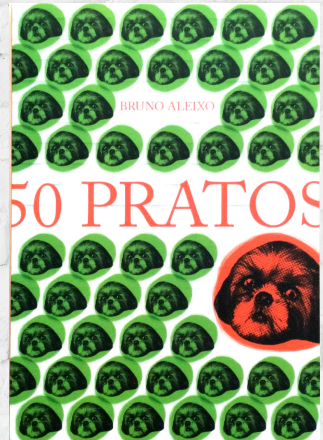
- pip
- Python Package Index (PyPI)

```
pip install --user setuptools wheel twine
```

```
pip install --user \  
git+git://github.com/andrefs/python-aleixo50.git@master
```

Example package: aleixo50

- **Package holding information regarding Bruno Aleixo's 50 culinary dishes**
- **Allows to list all the dishes, select a random one, ...**
- **Has a command line script**



Python package structure

```
aleixo50/  
├── README.md  
├── LICENSE  
├── MANIFEST.in  
├── setup.py  
└── aleixo50/
```

Python package structure

aleixo50/aleixo50/

- ├── `__init__.py`
- ├── `dish.py`
- ├── `recipes.json`
- ├── `dishes.py`
- ├── `rand.py`
- ├── `bin`
 - └── `aleixo50`
- ├── `test_dishes.py`
- └── `test_rand.py`

dish.py

```
class Dish(object):
    def __init__(self, name, ingrs=[], instrs=[]):
        self.name = name
        self.ingredients = ingrs
        self.instructions = instrs

    def __repr__(self):
        return ('Dish(                + \
                '{0.name!r}',         + \
                '{0.ingredients!r}',    + \
                '{0.instructions!r}))'  \
                .format(self))

    def __str__(self):
        return 'Dish({0.name})'.format(self)
```

recipes.json

```
[  
  {"name": "Arroz de pato"},  
  {"name": "Arroz de cabidela"},  
  {"name": "Leitão"},  
  {"name": "Dourada grelhada"},  
  (...)  
]
```

dishes.py

```
import os
import json
from .dish import Dish

_dirname = os.path.dirname(__file__)
_path = os.path.join(_dirname, "recipes.json")

with open(_path, 'r') as json_file:
    _recipes = json.loads(json_file.read())

dishes = [Dish(r['name']) for r in _recipes]
```


rand.py

```
import random
from .dishes import dishes

def rand():
    return random.choice(dishes)
```

aleixo50/___init___py

```
from .dishes import dishes  
from .rand import rand
```

bin/aleixo50

```
#!/usr/bin/python3
```

```
import getopt  
import sys  
import aleixo50
```

```
options, remainder = \  
    getopt.getopt(sys.argv[1:], 'lr', ['list', 'rand'])  
dict_opts = dict(options)
```

```
if '-r' in dict_opts or '--rand' in dict_opts:  
    print(aleixo50.rand().name)
```

```
else:  
    for i, r in enumerate(aleixo50.dishes):  
        print(str.rjust(str(i+1), 3) + '. ' + r.name)
```

Python package structure

```
aleixo50/  
├── README.md  
├── LICENSE  
├── MANIFEST.in  
├── setup.py  
└── aleixo50/
```

MANIFEST.in

```
include aleixo50/recipes.json
```


setup.py

```
import setuptools

with open("README.md", "r") as fh:
    long_description = fh.read()

setuptools.setup(
    name="aleixo50", version="0.0.1", author="André Santos",
    author_email="andrefs@andrefs.com",
    description="(...)", long_description=long_description,
    long_description_content_type="text/markdown",
    url="https://github.com/andrefs/python-aleixo50",
    packages=setuptools.find_packages(),
    scripts=['bin/aleixo50'],
    include_package_data=True,
    classifiers=[
        "Programming Language :: Python :: 3",
        "Operating System :: OS Independent",
        "Topic :: Utilities"
    ],
)
```

Linting

```
pylint aleixo50
```

- **By default, it follows PEP8 conventions**
- **Behavior can be modified with a configuration file**

test_rand.py

```
import unittest
import aleixo50
print(aleixo50)
```

```
from aleixo50 import rand
from aleixo50.dish import Dish
```

```
class RandTest(unittest.TestCase):
    def test_rand_type(self):
        self.assertIsInstance(rand(), Dish)
```

```
if __name__ == '__main__':
    unittest.main()
```

test_dishes.py

```
import unittest
import aleixo50
print(aleixo50)

from aleixo50 import dishes

class RandTest(unittest.TestCase):
    def test_rand_type(self):
        self.assertEqual(len(dishes), 50)

if __name__ == '__main__':
    unittest.main()
```

Testing

```
python -m unittest discover
```

```
..
```

```
Ran 2 tests in 0.000s
```

```
OK
```


Building the distribution

```
python3 setup.py sdist bdist_wheel
```

Publishing

Uploading to TestPyPI

```
twine upload \  
  --repository-url https://test.pypi.org/legacy/ \  
  dist/*
```

Installing from TestPyPI

```
pip install \  
  --index-url https://test.pypi.org/simple/ \  
  aleixo50
```

Publishing

Uploading

```
twine upload dist/*
```

Installing

```
pip install --user aleixo50
```

Making a new release

- 1. increment the version number in your `setup.py` file,**
- 2. update your `CHANGES.txt` file,**
- 3. if necessary, update the “Contributors” and “Thanks also to” sections of your `README.txt` file.**
- 4. run `twine upload` again.**

Paragraph and sentence delimitation

329 Mas o silêncio prolongado de Ron e Hermione tinham-no feito
330 sentir-se tão longe do mundo da magia que até divertir-se à
331 custa de Dudley perdera o interesse. E agora o Ron e a
332 Hermione tinham-se esquecido do seu aniversário.
333
334 Quanto não daria ele por uma mensagem de Hogwarts? De qualquer
335 feiticeiro ou feiticeira? Quase ficaria satisfeito com um
336 sinal do seu inimigo feroz, Draco Malfoy, só para ter a
337 certeza de que tudo aquilo não fora apenas um sonho...
338
339 Não que tudo durante o ano em Hogwarts tivesse sido divertido.
340 Mesmo no fim do último período, Harry :, confrontara-se nem
341 mais nem menos do que com o próprio Lord Voldemort. Voldemort

Paragraph and sentence delimitation

331 Mas o silêncio prolongado de Ron e Hermione tinham-no
feito sentir-se tão longe do mundo da magia que até
divertir-se à custa de Dudley perdera o interesse.

332 E agora o Ron e a Hermione tinham-se esquecido do seu
aniversário.

333

334 Quanto não daria ele por uma mensagem de Hogwarts?

335 De qualquer feiticeiro ou feiticeira? Quase ficaria
satisfeito com um sinal do seu inimigo feroz, Draco
Malfoy, só para ter a certeza de que tudo aquilo não fora
apenas um sonho...

336

337 Não que tudo durante o ano em Hogwarts tivesse sido
divertido. Mesmo no fim do último período, Harry :,
confrontara-se nem mais nem menos do que com o próprio
Lord Voldemort.

Paragraph and sentence delimitation

329 Mas o silêncio prolongado de Ron e Hermione tinham-no feito
330 sentir-se tão longe do mundo da magia que até divertir-se à
331 custa de Dudley perdera o interesse. E agora o Ron e a
332 Hermione tinham-se esquecido do seu aniversário.

333 Quanto não daria ele por uma mensagem de Hogwarts? De
334 qualquer feiticeiro ou feiticeira? Quase ficaria satisfeito
335 com um sinal do seu inimigo feroz, Draco Malfoy, só para ter a
336 certeza de que tudo aquilo não fora apenas um sonho...

337 Não que tudo durante o ano em Hogwarts tivesse sido
338 divertido. Mesmo no fim do último período, Harry confrontara-
339 se nem mais nem menos do que com o próprio Lord Voldemort.

340 Voldemort podia ter arruinado o seu ego inicial mas

Paragraph and sentence delimitation

One approach is to:

- **Find paragraph break candidates**
- **Analyze whether they're actual paragraph breaks**
- **Add a break mark**
- **Normalization**
- **Repeat for sentences**

Named entity recognition (NER)

Identification of all the instances of a name for a given type of entity within a given set of texts

- **lexicon-based:** use a dictionary (and maybe Regex)
- **rules-based:** define rules and patterns
- **statistical approach:** train a classifier with manually annotated documents

Exercises I

https://github.com/andrefs/spln-2018-i/tree/master/data/livros/harry_potter

- 1. Create a function to mark paragraphs in a text**
- 2. Create a function to mark sentences in a text**
- 3. Create a new package spln-utils (just boilerplate code)**
- 4. Add the functions to the module**
- 5. Create a command line script which splits paragraphs and sentences**

Exercises II

- 1. Create a function to mark proper names in a text**
- 2. Add the function to the previously created package**
- 3. Add options to:**
 - 3.1 print the list of names**
 - 3.2 print the text with marks**
 - 3.3 print the number of occurrences of each name**