Tokenize UK Documentation

Release 0.1.4

Vsevolod Dyomkin, Dmitry Chaplinsky

May 29, 2016

Contents

1 Tokenize UK 3

1.1 Features . . . 3

1.2 API . . . 3

2 Installation 5

2.1 Stable release . . . 5

2.2 From sources . . . 5

3 Usage 7

4 Contributing 9

4.1 Types of Contributions . . . 9

4.2 Get Started! . . . 10

4.3 Pull Request Guidelines . . . 10

4.4 Tips . . . 11

5 Credits 13

5.1 Development Lead . . . 13

5.2 Contributors . . . 13

6 History 15

6.1 0.1.0 (2016-05-29) . . . 15

7 Indices and tables 17

Contents:

CHAPTER 1

Tokenize UK

Simple python lib to tokenize texts into sentences and sentences to words. Small, fast and robust. Comes with ukrainian flavour

- Free software: MIT license
- Documentation: https://tokenize_uk.readthedocs.org.

1.1 Features

- Tokenize given text into sentences
- Tokenize given sentence into words
- Works well with accented characters (like stresses) and apostrophes
- Suitable also for other languages

1.2 API

Ukrainian tokenization script based on standard tokenization algorithm.

2016 (c) Vsevolod Dyomkin <vseloved@gmail.com>, Dmitry Chaplinsky <chaplinsky.dmitry@gmail.com>

`tokenize_uk.tokenize_uk.`**`tokenize_words`**`(`*string*`)`

Tokenize input text to words.

Parameters **`string`** (*`str or unicode`*) – Text to tokenize

Returns words

Return type list of strings

`tokenize_uk.tokenize_uk.`**`tokenize_text`**`(`*string*`)`

Tokenize input text to paragraphs, sentences and words.

Tokenization to paragraphs is done using simple Newline algorithm For sentences and words tokenizers above are used

Parameters **`string`** (*`str or unicode`*) – Text to tokenize

Returns text, tokenized into paragraphs, sentences and words

Return type list of list of list of words `tokenize_uk.tokenize_uk.`**`tokenize_sents`**`(`*string*`)`

Tokenize input text to sentences.

Parameters **`string`** (*`str or unicode`*) – Text to tokenize

Returns sentences

Return type list of strings

CHAPTER 2

Installation

2.1 Stable release

To install Tokenize UK, run this command in your terminal:

```
$ pip install tokenize_uk
```

If you don't have pip installed, this Python installation guide can guide you through the process.

2.2 From sources

The sources for Tokenize UK can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/dchaplinsky/tokenize_uk
```

Or download the tarball:

```
$ curl  -OL https://github.com/dchaplinsky/tokenize_uk/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```

CHAPTER 3

Usage

To use Tokenize UK in a project:

```
import tokenize_uk
```

CHAPTER 4

Contributing

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given.

You can contribute in many ways:

4.1 Types of Contributions

4.1.1 Report Bugs

Report bugs at https://github.com/dchaplinsky/tokenize_uk/issues.

If you are reporting a bug, please include:

- Your operating system name and version.
- Any details about your local setup that might be helpful in troubleshooting.
- Detailed steps to reproduce the bug.

4.1.2 Fix Bugs

Look through the GitHub issues for bugs. Anything tagged with “bug” is open to whoever wants to implement it.

4.1.3 Implement Features

Look through the GitHub issues for features. Anything tagged with “feature” is open to whoever wants to implement it.

4.1.4 Write Documentation

Tokenize UK could always use more documentation, whether as part of the official Tokenize UK docs, in docstrings, or even on the web in blog posts, articles, and such.

4.1.5 Submit Feedback

The best way to send feedback is to file an issue at https://github.com/dchaplinsky/tokenize_uk/issues.

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.
- Remember that this is a volunteer-driven project, and that contributions are welcome :)

4.2 Get Started!

Ready to contribute? Here's how to set up *tokenize_uk* for local development.

1. Fork the *tokenize_uk* repo on GitHub.
2. Clone your fork locally:

```
$ git clone git@github.com:your_name_here/tokenize_uk.git
```

3. Install your local copy into a virtualenv. Assuming you have virtualenvwrapper installed, this is how you set up your fork for local development:

```
$ mkvirtualenv tokenize_uk
$ cd tokenize_uk/
$ python setup.py develop
```

4. Create a branch for local development:

```
$ git checkout -b name-of-your-bugfix-or-feature
```

   Now you can make your changes locally.

5. When you're done making changes, check that your changes pass flake8 and the tests, including testing other Python versions with tox:

```
$ flake8 tokenize_uk tests
$ python setup.py test or py.test
$ tox
```

   To get flake8 and tox, just pip install them into your virtualenv.

6. Commit your changes and push your branch to GitHub:

```
$ git add .
$ git commit -m "Your detailed description of your changes."
$ git push origin name-of-your-bugfix-or-feature
```

7. Submit a pull request through the GitHub website.

4.3 Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. The pull request should include tests.
2. If the pull request adds functionality, the docs should be updated. Put your new functionality into a function with a docstring, and add the feature to the list in README.rst.
3. The pull request should work for Python 2.6, 2.7, 3.3, 3.4 and 3.5, and for PyPy. Check https://travis-ci.org/dchaplinsky/tokenize_uk/pull_requests and make sure that the tests pass for all supported Python versions.

4.4 Tips

To run a subset of tests:

```
$ py.test tests.test_tokenize_uk
```

CHAPTER 5

Credits

5.1 Development Lead

- Vsevolod Dyomkin, Dmitry Chaplinsky <chaplinsky.dmitry@gmail.com>

5.2 Contributors

None yet. Why not be the first?

CHAPTER 6

History

6.1 0.1.0 (2016-05-29)

- First release on PyPI.

CHAPTER 7

Indices and tables

- genindex
- modindex
- search

T

tokenize_sents() (in module tokenize_uk.tokenize_uk), 3
tokenize_text() (in module tokenize_uk.tokenize_uk), 3
tokenize_uk.tokenize_uk (module), 3
tokenize_words() (in module tokenize_uk.tokenize_uk), 3