Fuzzy 1.1

Fast Python phonetic algorithms

Download
Fuzzy-1.1.tar.gz

Fuzzy is a python library implementing common phonetic algorithms quickly. Typically this is in string similarity exercises, but they're pretty versatile.

It uses C Extensions (via Pyrex) for speed.

The algorithms are:

- **Soundex**
- **NYSIIS**
- **Double Metaphone** Based on Maurice Aubrey's C code from his perl implementation.

Copyright

Fuzzy is Copyright 2011 YouGov, Plc. and released under the MIT license except where noted in specific source files.

Installation

Installation should be easy if you have a C compiler such as gcc. All you should need to do is *easy_install/pip install* it. If you have Pyrex it will regenerate the C code, otherwise it will use the pre-generated code. Here's a basic installation on a clean virtualenv:

```
(fuzzy_cean)Kotai:~ chmullig$ pip install https://bitbucket.org/yougov/fuzzy/get/1.0.tar.gz
Downloading/unpacking https://bitbucket.org/yougov/fuzzy/get/1.0.tar.gz
  Downloading 1.0.tar.gz
  Running setup.py egg_info for package from https://bitbucket.org/yougov/fuzzy/get/1.0.tar.gz
Installing collected packages: Fuzzy
  Running setup.py install for Fuzzy
    building 'fuzzy' extension
    gcc-4.2 -fno-strict-aliasing -fno-common -dynamic -DNDEBUG -g -fwrapv -Os -Wall -Wstrict-prototypes
        -DENABLE_DTRACE -arch i386 -arch ppc -arch x86_64 -pipe -I/System/Library/Frameworks/Python.framework/Versions/2.6/include/python2.6
        -c src/fuzzy.c -o build/temp.macosx-10.6-universal-2.6/src/fuzzy.o
    gcc-4.2 -fno-strict-aliasing -fno-common -dynamic -DNDEBUG -g -fwrapv -Os -Wall -Wstrict-prototypes
        -DENABLE_DTRACE -arch i386 -arch ppc -arch x86_64 -pipe -I/System/Library/Frameworks/Python.framework/Versions/2.6/include/python2.6
        -c src/double_metaphone.c -o build/temp.macosx-10.6-universal-2.6/src/double_metaphone.o
    gcc-4.2 -Wl,-F. -bundle -undefined dynamic_lookup -arch i386 -arch ppc -arch x86_64
        build/temp.macosx-10.6-universal-2.6/src/fuzzy.o build/temp.macosx-10.6-universal-2.6/src/double_metaphone.o
        -o build/lib.macosx-10.6-universal-2.6/fuzzy.so
Successfully installed Fuzzy
Cleaning up...
(fuzzy_cean)Kotai:~ chmullig$
```

Usage

The functions are quite easy to use!

```
>>> import fuzzy
>>> soundex = fuzzy.Soundex(4)
>>> soundex('fuzzy')
'F200'
>>> dmeta = fuzzy.DMetaphone()
>>> dmeta('fuzzy')
['FS', None]
>>> fuzzy.nysiis('fuzzy')
'FASY'
```

Performance

Fuzzy's Double Metaphone was ~10 times faster than the pure python implementation by **Andrew Collins** in some recent **testing**. Soundex and NYSIIS should be similarly faster. Using iPython's timeit:

```
In [3]: timeit soundex('fuzzy')
1000000 loops, best of 3: 326 ns per loop

In [4]: timeit dmeta('fuzzy')
100000 loops, best of 3: 2.18 us per loop

In [5]: timeit fuzzy.nysiis('fuzzy')
100000 loops, best of 3: 13.7 us per loop
```

Distance Metrics

We recommend the **Python-Levenshtein** module for fast, C based string distance/similarity metrics. Among others functions it includes:

- **Levenshtein** edit distance
- **Jaro** distance
- **Jaro-Winkler** distance
- **Hamming distance**

In testing it's been several times faster than comparable pure python implementations of those algorithms.

File	Type	Py Version	Uploaded on	Size
Fuzzy-1.1.tar.gz (**md5**)	Source		2015-06-04	19KB

Author: YouGov, Plc.
Home Page: https://bitbucket.org/yougov/fuzzy
Categories
Development Status :: 4 - Beta
License :: OSI Approved :: Artistic License
License :: OSI Approved :: MIT License
Operating System :: POSIX
Programming Language :: Python :: 2.7

Topic :: Text Processing
Topic :: Text Processing :: General
Topic :: Text Processing :: Indexing
Topic :: Text Processing :: Linguistic
Package Index Owner: chmullig, jaraco
DOAP record: Fuzzy-1.1.xml