
probablepeople Documentation

Release 0.3.1

Cathy Deng, Forest Gregg

February 09, 2017

1	Installation	3
2	Usage	5
3	Details	7
4	Important links	9
5	Indices and tables	11

probablepeople is a python library for parsing unstructured romanized name or company strings into name components, using advanced NLP methods.

Installation

```
pip install probablepeople
```

Usage

The `parse` method will split your string into components, and label each component.

```
>>> import probablepeople
>>> probablepeople.parse('Mr George "Gob" Bluth II')
[('Mr', 'PrefixMarital'),
 ('George', 'GivenName'),
 ('"Gob"', 'Nickname'),
 ('Bluth', 'Surname'),
 ('II', 'SuffixGenerational')]
>>> probablepeople.parse('Lucille & George Bluth')
[('Lucille', 'GivenName'),
 ('&', 'And'),
 ('George', 'GivenName'),
 ('Bluth', 'Surname')]
>>> probablepeople.parse('Sitwell Housing Inc')
[('Sitwell', 'CorporationName'),
 ('Housing', 'CorporationName'),
 ('Inc', 'CorporationLegalType')]
```

The `tag` method will return an `OrderedDict` with distinct labels as keys & parts of your string as values, as well as a string type

```
>>> import probablepeople
>>> probablepeople.tag('Mr George "Gob" Bluth II')
(OrderedDict([
 ('PrefixMarital', 'Mr'),
 ('GivenName', 'George'),
 ('Nickname', '"Gob"'),
 ('Surname', 'Bluth'),
 ('SuffixGenerational', 'II')]),
 'Person')
>>> probablepeople.tag('Lucille & George Bluth')
(OrderedDict([
 ('GivenName', 'Lucille'),
 ('And', '&'),
 ('SecondGivenName', 'George'),
 ('Surname', 'Bluth')]),
 'Household')
>>> probablepeople.tag('Sitwell Housing Inc')
(OrderedDict([
 ('CorporationName', 'Sitwell Housing'),
 ('CorporationLegalType', 'Inc')]),
 'Corporation')
```

Because the `tag` method returns an `OrderedDict` with labels as keys, it will throw a `RepeatedLabelError` error when multiple areas of a name have the same label, and thus can't be concatenated. When `RepeatedLabelError` is raised, it is likely that either (1) the input string is not a valid person/corporation name, or (2) some tokens were labeled incorrectly.

`RepeatedLabelError` has the attributes `original_string` (the input string) and `parsed_string` (the output of the pa

```
try:
    tagged_name, name_type = probablepeople.tag(string)
except probablepeople.RepeatedLabelError as e :
    some_special_instructions(e.parsed_string, e.original_string)
```

If you already know that the string refers to a person or a company, you can indicate that to probable people by using the `type` argument of the `parse` and `tag` methods. Valid options are `'person'` and `'company'`.

Details

probablepeople has the following labels for parsing names & companies:

- PrefixMarital
- PrefixOther
- GivenName
- FirstInitial
- MiddleName
- MiddleInitial
- Surname
- LastInitial
- SuffixGenerational
- SuffixOther
- Nickname
- And
- CorporationName
- CorporationNameOrganization
- CorporationLegalType
- CorporationNamePossessiveOf
- ShortForm
- ProxyFor
- AKA

Important links

- Documentation: <http://probablepeople.rtf.d.org/>
- Repository: <https://github.com/datamade/probablepeople>
- Issues: <https://github.com/datamade/probablepeople/issues>
- Distribution: <https://pypi.python.org/pypi/probablepeople>
- Blog Post: <http://datamade.us/blog/parse-name-or-parse-anything-really/>
- Web Interface: <http://parserator.datamade.us/probablepeople>

Indices and tables

- `genindex`
- `modindex`
- `search`