Wiki-Class Documentation

Release v0.0.1

Aaron Halfaker & Morten Warncke-Wang

Aug 14, 2019

Contents

A library for performing automatic detection of assessment classes of Wikipedia articles.

Compatible with Python 3.x only. Sorry.

- **Install:** `pip install wikiclass`
- **Models:** https://github.com/halfak/Wiki-Class/tree/master/models
- **Repo:** https://github.com/halfak/Wiki-Class

CHAPTER 1

Basic usage

If you want to detect some assessment classes, you're going to need a model. You can download a prebuilt model or build one yourself.

1.1 Model from file

```
from wikiclass.models import RFTextModel

model = RFTextModel.from_file(open("enwiki.rf_text.model", "rb"))

assessment, probabilities = model.classify("Some article text")

print("I'm about {0}% ".format(probabilities[assessment]*100) + \
      "sure that this should be classified {0}".format(assessment))
```

1.2 Model building

```
from wikiclass.models import RFTextModel

# Gather a training and test set
train_set = [
    ("Stub", "Some article text"),
    ("Start", "Some more article text<ref>news</ref>.")
    # ...
]
test_set = [
    ("C", "The Lorem Ipsum dolored the sit amet."),
    ("FA", "'''Lorem Ipsum''', sit amet the dolor amer. {{InfoBox|...}}")
    # ...
]
```

(continues on next page)

(continued from previous page)

```
# Train a model
model = RFTextModel.train(
    train_set,
    assessments=assessments.WP10
)

# Run the test set & print the results
results = model.test(test_set)
print(results)

# Write the model to disk for reuse.
model.to_file(open("enwiki.rf_text.model", "wb"))
```

CHAPTER 2

Modules

wikiclass.models A set of classification models that can be trained and used to classify articles.

- `RFTextModel` – A random forrest classifier that extracts features from article text.

wikiclass.features A set of feature extractors used to organize a set of features for use in model training and classification.

- `WikitextAndInfonoise` – A text feature extractor that gathers wiki markup features and an information-based measure.

wikiclass.languages Some `FeatureExtractor` s require information about the language being processed. This module contains basic language info for common languages.

- **`get()`, gets a `Language` based on a name. Currently supported languages include:**
  - `"English"`
- `register()`, registers a new `Language` for access from `get()`.

CHAPTER 3

Authors

Aaron Halfaker

- ahalfaker@wikimedia.org
- http://halfaker.info

Morten Warcke-Wang

- http://www-users.cs.umn.edu/~morten

CHAPTER 4

Indices and tables

- genindex
- modindex
- search