Package Index > **snownlp** > **0.11.1**

snownlp 0.11.1

Python library for processing Chinese text

Download
snownlp-0.11.1.tar.gz

Latest Version: 0.12.3

#SnowNLP: Simplified Chinese Text Processing

SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了[TextBlob](https://github.com/sloria/TextBlob)的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和TextBlob不同的是，这里没有用NLTK，所有的算法都是自己实现的，并且自带了一些训练好的字典。注意本程序都是处理的unicode编码，所以使用时请自行decode成unicode。

~~~~{python}
from snownlp import SnowNLP

s = SnowNLP(u'这个东西真心很赞')

s.words # [u'这个', u'东西', u'真心',
# u'很', u'赞']

s.tags # [(u'这个', u'r'), (u'东西', u'n'),
# (u'真心', u'd'), (u'很', u'd'),
# (u'赞', u'Vg')]

s.sentiments # 0.9769663402895832 positive的概率

s.pinyin # [u'zhe', u'ge', u'dong', u'xi',
# u'zhen', u'xin', u'hen', u'zan']

s = SnowNLP(u'「繁體字」「繁體中文」的叫法在臺灣亦很常見。')

s.han # u'「繁体字」「繁体中文」的叫法
# 在台湾亦很常见。'

text = u'''
自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。
它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。
~~~~

```
自然语言处理是一门融语言学、计算机科学、数学于一体的科学。
因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，
所以它与语言学的研究有着密切的联系，但又有重要的区别。
自然语言处理并不是一般地研究自然语言，
而在于研制能有效地实现自然语言通信的计算机系统，
特别是其中的软件系统。因而它是计算机科学的一部分。
'''

s = SnowNLP(text)

s.keywords(3) # [u'语言', u'自然', u'计算机']

s.summary(3) # [u'因而它是计算机科学的一部分',
# u'自然语言处理是一门融语言学、计算机科学、
# 数学于一体的科学',
# u'自然语言处理是计算机科学领域与人工智能
# 领域中的一个重要方向']
s.sentences

s = SnowNLP([[u'这篇', u'文章'],
[u'那篇', u'论文'],
[u'这个']])
s.tf
s.idf
s.sim([u'文章'])# [0.3756070762985226, 0, 0]
~~~~
```

Features

* 中文分词（[Character-Based Generative Model](http://aclweb.org/anthology//Y/Y09/Y09-2047.pdf)）
* 词性标注（[TnT](http://aclweb.org/anthology//A/A00/A00-1031.pdf) 3-gram 隐马）
* 情感分析（现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）
* 文本分类（Naive Bayes）
* 转换成拼音
* 繁体转简体
* 提取文本关键词（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf)算法）
* 提取文本摘要（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf)算法）
* tf，idf
* Tokenization（分割成句子）

* 文本相似（[BM25](http://en.wikipedia.org/wiki/Okapi_BM25)）
* 支持python3（感谢[erning](https://github.com/erning)）

Get It now

~~~~
$ pip install snownlp
~~~~

关于训练

现在提供训练的包括分词，词性标注，情感分析，而且都提供了我用来训练的原始文件
以分词为例
分词在`snownlp/seg`目录下
~~~~{python}
from snownlp import seg
seg.train('data.txt')
seg.save('seg.marshal')
#from snownlp import tag
#tag.train('199801.txt')
#tag.save('tag.marshal')
#from snownlp import sentiment
#sentiment.train('neg.txt', 'pos.txt')
#sentiment.save('sentiment.marshal')
~~~~
这样训练好的文件就存储为`seg.marshal`了，之后修改`snownlp/seg/__init__.py`里的`data_path`指向刚训练好的文件即可

License

MIT licensed.

File	Type	Py Version	Uploaded on	Size
snownlp-0.11.1.tar.gz (**md5**)	Source		2014-04-05	37MB

Author: isnowfy
Home Page: https://github.com/isnowfy/snownlp
Categories
Development Status :: 3 - Alpha
Intended Audience :: Developers
License :: OSI Approved :: MIT License

Programming Language :: Python

Package Index Owner: isnowfy

DOAP record: snownlp-0.11.1.xml