Newspaper: 新闻文章元数据抽取的开源Python库

来自于requests的灵感，因为它很简单；并且由lxml驱动，因为它速度很快。

Newspaper是一个惊人的新闻、全文以及文章元数据抽取开源的Python类库，这个类库支持10多种语言，所有的东西都是用unicode编码的。我们可以使用下面命令查看：

```
/**
 * User: 过往记忆
 * Date: 2015-05-20
 * Time: 下午23:14
 * bolg:
 * 本文地址：/archives/1363
 * 过往记忆博客，专注于hadoop、hive、spark、shark、flume的技术博客，大量的干货
 * 过往记忆博客微信公共帐号：iteblog_hadoop
 */

>>> import newspaper
>>> newspaper.languages()

Your available langauges are:
input code      full name

  ar           Arabic
  ru           Russian
  nl           Dutch
  de           German
  en           English
  es           Spanish
  fr           French
  it           Italian
  ko           Korean
  no           Norwegian
  pt           Portuguese
  sv           Swedish
  hu           Hungarian
  fi           Finnish
  da           Danish
  zh           Chinese
  id           Indonesian
  vi           Vietnamese
```

来看看这个如何使用：

```
>>> import newspaper

>>> cnn_paper = newspaper.build('http://cnn.com')

>>> for article in cnn_paper.articles:
>>>     print(article.url)
u'http://www.cnn.com/2013/11/27/justice/tucson-arizona-captive-girls/'
u'http://www.cnn.com/2013/12/11/us/texas-teen-dwi-wreck/index.html'
...

>>> for category in cnn_paper.category_urls():
>>>     print(category)

u'http://lifestyle.cnn.com'
u'http://cnn.com/world'
u'http://tech.cnn.com'
...
>>> article = cnn_paper.articles[0]
>>> article.download()

>>> article.html
u'<!DOCTYPE HTML><html itemscope itemtype="http://...'
>>> article.parse()

>>> article.authors
[u'Leigh Ann Caldwell', 'John Honway']

>>> article.text
u'Washington (CNN) -- Not everyone subscribes to a New Year's resolution...'

>>> article.top_image
u'http://someCDN.com/blah/blah/blah/file.png'

>>> article.movies
[u'http://youtube.com/path/to/link.com', ...]
>>> article.nlp()

>>> article.keywords
['New Years', 'resolution', ...]

>>> article.summary
u'The study shows that 93% of people ...'
```

Newspaper拥有无缝的语言提取和检测功能。如果使用的时候没有指定语言，Newspaper将会自动地检测语言：

```
>>> from newspaper import Article
>>> url = 'http://www.bbc.co.uk/zhongwen/simp/chinese_news/2012/12/121210_hongkong_politics.shtml'

>>> a = Article(url, language='zh') # Chinese

>>> a.download()
>>> a.parse()

>>> print(a.text[:150])
香港行政长官梁振英在各方压力下就其大宅的违章建
筑（僭建）问题到立法会接受质询，并向香港民众道歉。
梁振英在星期二（12月10日）的答问大会开始之际在其
演说中道歉，但强调他在违章建筑问题上没有隐瞒的意
图和动机。 一些亲北京阵营议员欢迎梁振英道歉，
且认为应能获得香港民众接受，但这些议员也质问梁振英有

>>> print(a.title)
港特首梁振英就住宅违建事件道歉
```

如果你知道你抓取文章源的语言编码，那么可以在API中加上这个：

```
>>> import newspaper
>>> sina_paper = newspaper.build('http://www.sina.com.cn/', language='zh')

>>> for category in sina_paper.category_urls():
>>>     print(category)
u'http://health.sina.com.cn'
u'http://eladies.sina.com.cn'
u'http://english.sina.com'
...

>>> article = sina_paper.articles[0]
>>> article.download()
>>> article.parse()

>>> print(article.text)
新浪武汉汽车综合 随着汽车市场的日趋成熟，传统的“集
全家之力抱得爱车归”的全额购车模式已然过时，另一种轻
松的新兴 车模式——金融购车正逐步成为时下消费者购买
```

爱车最为时尚的消费理 念，他们认为，这种新颖的购车模
式既能在短期内
...

>>> print(article.title)
两年双免0手续0利率 科鲁兹掀背金融轻松购_武汉车市_武汉
汽车网_新浪汽车_新浪网

Newspaper的功能

1、能够支持10多种语言(English, Chinese, German, Arabic, ...)；
2、多线程文章下载框架；
3、新闻URL识别；
4、从HTML抽取文本；
5、从HTML抽取热门图片；
6、从HTML中抽取所有的图片；
7、从文本中抽取关键字；
8、从文本中抽取摘要；
9、从文本中抽取出作者；
10、Google趋势term提取。

下载地址：http://newspaper.readthedocs.org/en/latest/
使用文档：http://newspaper.readthedocs.org/en/latest/user_guide/install.html