# Newspaper3k: Article scraping & curation


Inspired by requests for its simplicity and powered by lxml for its speed:

> "Newspaper is an amazing python library for extracting & curating articles." – tweeted by Kenneth Reitz, Author of requests
>
> "Newspaper delivers Instapaper style article extraction." – The Changelog

**Newspaper is a Python3 library**! View on Github here, or, view our **deprecated and buggy** Python2 branch

## A Glance:

```
>>> from newspaper import Article

>>> url = 'http://fox13now.com/2013/12/30/new-year-new-laws-obamacare-pot-guns-and-drones/'
>>> article = Article(url)


>>> article.download()

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


>>> article.parse()

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

>>> article.publish_date
datetime.datetime(2013, 12, 30, 0, 0)

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

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

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


>>> article.nlp()

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

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


>>> import newspaper

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

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

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

http://lifestyle.cnn.com
http://cnn.com/world
http://tech.cnn.com
...

>>> cnn_article = cnn_paper.articles[0]
>>> cnn_article.download()
>>> cnn_article.parse()
>>> cnn_article.nlp()
...


>>> from newspaper import fulltext

>>> html = requests.get(...).text
>>> text = fulltext(html)
```

Newspaper has *seamless* language extraction and detection. If no language is specified, Newspaper will attempt to auto detect a language.

```
>>> 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
```

If you are certain that an *entire* news source is in one language, **go ahead and use the same api :)**

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

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

# Documentation

Check out The Documentation for full and detailed guides using newspaper.

Interested in adding a new language for us? Refer to: Docs - Adding new languages

# Features

- Multi-threaded article download framework
- News url identification
- Text extraction from html
- Top image extraction from html
- All image extraction from html
- Keyword extraction from text
- Summary extraction from text
- Author extraction from text
- Google trending terms extraction
- Works in 10+ languages (English, Chinese, German, Arabic, …)

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

Your available languages are:
input code      full name

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

```
tr              Turkish
el              Greek
uk              Ukrainian
```

## Get it now

Run ✅ `pip3 install newspaper3k` ✅

NOT ⛔ `pip3 install newspaper` ⛔

On python3 you must install `newspaper3k`, **not** `newspaper`. `newspaper` is our python2 library. Although installing newspaper is simple with pip, you will run into fixable issues if you are trying to install on ubuntu.

**If you are on Debian / Ubuntu**, install using the following:

- Install `pip3` command needed to install `newspaper3k` package:

  ```
  $ sudo apt-get install python3-pip
  ```

- Python development version, needed for Python.h:

  ```
  $ sudo apt-get install python-dev
  ```

- lxml requirements:

  ```
  $ sudo apt-get install libxml2-dev libxslt-dev
  ```

- For PIL to recognize .jpg images:

  ```
  $ sudo apt-get install libjpeg-dev zlib1g-dev libpng12-dev
  ```

NOTE: If you find problem installing `libpng12-dev`, try installing `libpng-dev`.

- Download NLP related corpora:

  ```
  $ curl https://raw.githubusercontent.com/codelucas/newspaper/master/download_corpora.py | pythor
  ```

- Install the distribution via pip:

  ```
  $ pip3 install newspaper3k
  ```

**If you are on OSX**, install using the following, you may use both homebrew or macports:

```
$ brew install libxml2 libxslt

$ brew install libtiff libjpeg webp little-cms2

$ pip3 install newspaper3k

$ curl https://raw.githubusercontent.com/codelucas/newspaper/master/download_corpora.py | python3
```

**Otherwise**, install with the following:



NOTE: You will still most likely need to install the following libraries via your package manager

- PIL: `libjpeg-dev zlib1g-dev libpng12-dev`
- lxml: `libxml2-dev libxslt-dev`
- Python Development version: `python-dev`

```
$ pip3 install newspaper3k

$ curl https://raw.githubusercontent.com/codelucas/newspaper/master/download_corpora.py | python3
```

Using python 2.X? We support python 2, however development work has stopped on the 2.X branch for a few years now so it is behind in features and is more buggy. See python 2 installation instructions here

# Development

If you'd like to contribute and hack on the newspaper project, feel free to clone a development version of this repository locally:

```
git clone git://github.com/codelucas/newspaper.git
```

Once you have a copy of the source, you can embed it in your Python package, or install it into your site-packages easily:

```
$ pip3 install -r requirements.txt
$ python3 setup.py install
```

Feel free to give our testing suite a shot, everything is mocked!:

```
$ python3 tests/unit_tests.py
```

Planning on tweaking our full-text algorithm? Add the `fulltext` parameter:

```
$ python3 tests/unit_tests.py fulltext
```

# User Guide

# Demo

View a working online demo here: http://newspaper-demo.herokuapp.com This is another working online demo: http://newspaper.chinazt.cc/

# LICENSE

Authored and maintained by Lucas Ou-Yang.

Parse.ly sponsored some work on newspaper, specifically focused on automatic extraction.

Newspaper uses a lot of python-goose's parsing code. View their license here.

Please feel free to email & contact me if you run into issues or just would like to talk about the future of this library and news extraction in general!