Published in Towards Data Science

Artem Follow

Jul 17, 2019 · 3 min read · Listen

Scrape and Summarize News Articles in 5 Lines of Python Code

Good programers write the code, great search github first.

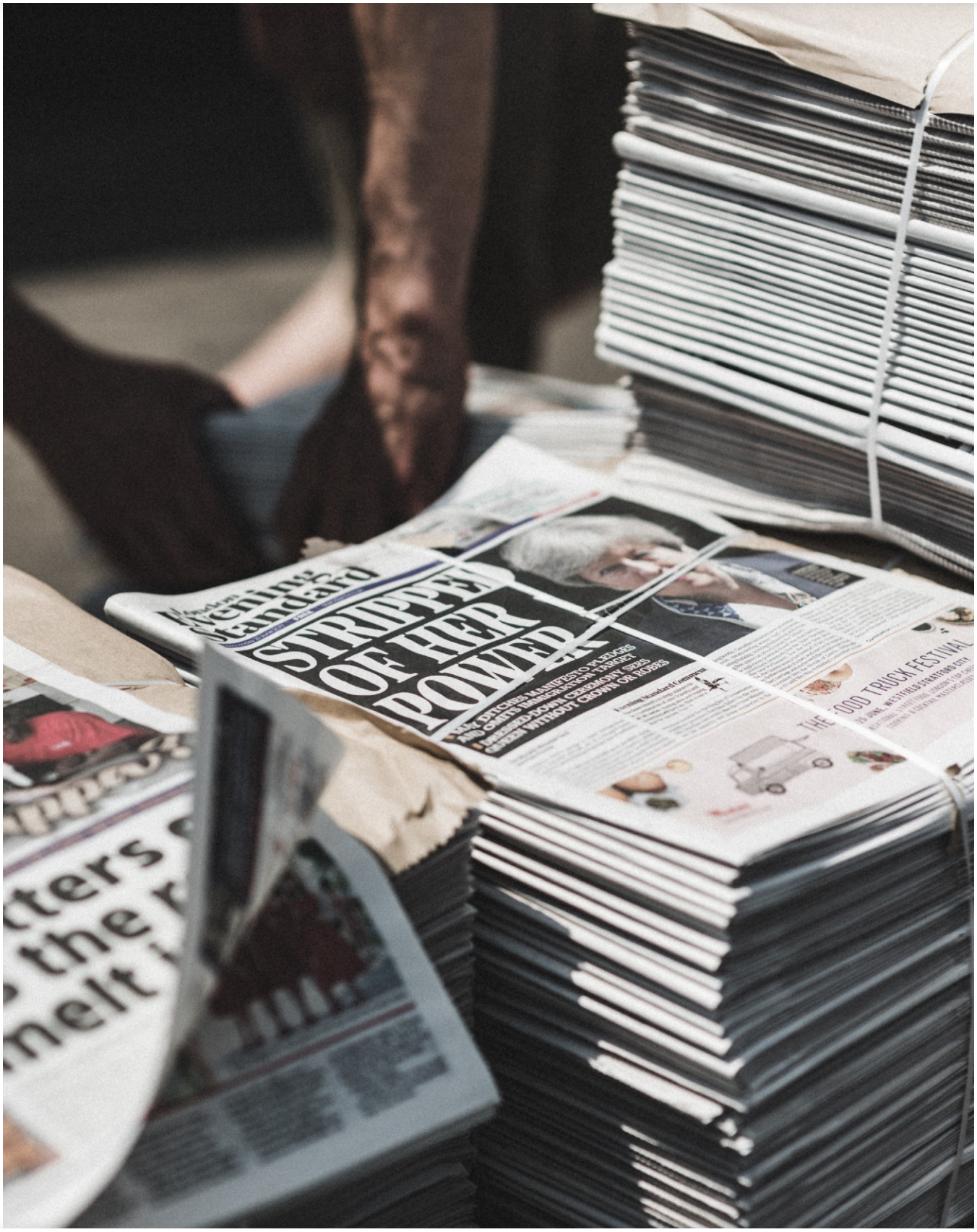

Photo by Thomas Charters on Unsplash

Want to stand out from the crowd of data scientists who just do machine learning and visualization? Then, you can begin one step earlier by collecting your own dataset instead of using outdated CSV files from Kaggle.

In this post I will show you how to collect lots of news data from many sources in a unified way. Therefore, instead of spending months on writing a script for each news website, you will use newspaper3k to automatically extract structured information.

Install the package:

```
$ pip install newspaper3k
```

Now, let's ask newspaper3k to scrape the article, extract information and summarize it for us.

```
>>> from newspaper import Article

>>> article = Article('https://www.npr.org/2019/07/10/740387601
/university-of-texas-austin-promises-free-tuition-for-low-income-
students-in-2020')

>>> article.download()

>>> article.parse()

>>> article.nlp()
```

That's all folks, 5 lines of code including package importing.

```
>>> article.authors

['Vanessa Romo', 'Claire Mcinerny']

>>> article.publish_date

datetime.datetime(2019, 7, 10, 0, 0)

>>> article.keywords

['free', 'program', '2020', 'muñoz', 'offering', 'loans',
'university', 'texas', 'texasaustin', 'promises', 'families',
'lowincome', 'students', 'endowment', 'tuition']
```

Concerning the text itself, you have an option to access full text:

```
>>> print(article.text)

University of Texas-Austin Promises Free Tuition For Low-Income
Students In 2020

toggle caption Jon Herskovitz/Reuters

Four year colleges and universities have difficulty recruiting...
```

In addition to that you get the built-in summary:

```
>>> print(article.summary)

University of Texas-Austin Promises Free Tuition For Low-Income
Students In 2020toggle caption Jon Herskovitz/ReutersFour year
colleges and universities have difficulty recruiting talented
students from the lower end of the economic spectrum who can't
afford to attend such institutions without taking on massive debt.
```


```
The endowment — which includes money from oil and gas royalties
earned on state-owned land in West Texas — more than doubles an
existing program offering free tuition to students whose families
make less than $30,000.

It also expands financial assistance to middle class students whose
families earn up to $125,000 a year, compared to the current
$100,000.

In 2008, Texas A&M began offering free tuition to students whose
families' income was under
```


Not bad for a built-in feature.

To profit from all features including automation of the feed of a magazine and accessing trending topics, please, refer to the official documentation.

Using newspaper3k you can collect your unique dataset to train your models. More importantly, you will have a real data feed after model is ready so you will also be able to see the real performance.

Define a problem first and only then search for data, not vice versa. Try to be a re lver and think how your model can resolve real business problems bec that is what you are going to be paid for.

st, I would emphasize you to read the one that inspired me.

Words that will inspire, a data science project on TED Talks

Machine Learning 8 min read

Shibsankar Das · Jul 17, 2019

Image similarity using Triplet Loss

Machine Learning 5 min read

Read more from Towards Data Science