

Scraping for stories

@PaulBradshaw

Course leader, MA Data Journalism, Birmingham City University
Data journalist, BBC Shared Data Unit

You are planning a story...

1. ...on zero hours contracts via agencies. How can you prove that this is a **systemic problem**?
2. ...on discrimination in housing that prevents people finding rooms to rent - but how do you establish **how big the problem is**?
3. ...on MPs paying relatives — but how do you get an **exclusive angle**?

What is scraping?

Automating the gathering of information from **online sources**

Typically, multiple **pages, documents or updates**

...across **space** (e.g. a website) **and/or time** (e.g. every day or hour).



Print Sample



1 of 6



Hallelujah

Words & Music by Leonard Cohen

Freely ♩ = 66

N.C.

The musical score for 'Hallelujah' is presented in a four-staff format. The first staff is a grand staff (treble and bass clef) with a melody line and a bass line. The second staff is a single treble clef staff with a melody line. The third staff is a single bass clef staff with a bass line. The fourth staff is a single treble clef staff with a melody line. The score includes various musical notations such as notes, rests, and dynamic markings. Chord diagrams are provided for several chords: Am, Am(5), Am(5), Dm7(5), F/C, G, C, and Am7. The score is marked with 'mp' (mezzo-piano) and 'p' (piano). The tempo is marked 'Freely ♩ = 66'. The key signature is one flat (B-flat).

Hallelujah

BY JEFF BUCKLEY - DIGITAL SHEET MUSIC

Price: £4.20

Save Up To 25% with Volume Discounts [?](#)

Includes digital copy + 1 print.

Each additional print is £3.04

Transpose (9) ▾

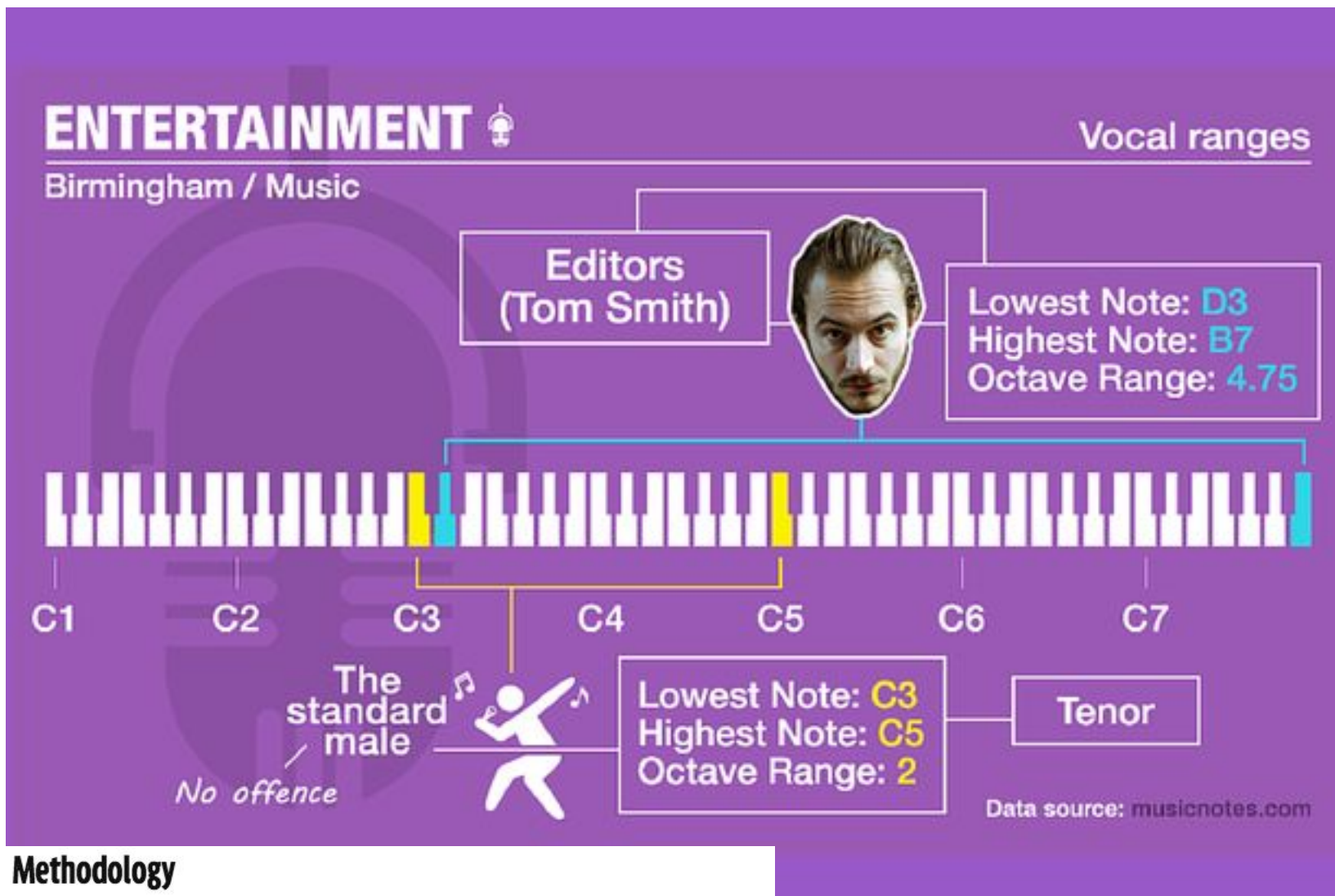
[See other arrangements of this song](#)

QUICK DETAILS

Scoring:	Piano/Vocal/Guitar
Instruments:	Voice, range: C4-F5 • Piano • Guitar
Pages:	6
Avg. Rating:	★★★★★
Product #:	MN0053340
Lyrics:	Contains complete lyrics

[View Full Product Details](#)

Which singer has the best vocal range in the UK - No, it's not who you think



Methodology

In case you're wondering, this list is based on sheet music data from musicnotes.com, a website with over 260,000 sheet music arrangements. We took vocal melodies from UK bands and artists and plotted the highest and lowest notes each singer has recorded.

Why scraping?

1. Proof
2. Context
3. Exclusive leads

Why scraping?

1. Proof

2. Context

3. Exclusive leads

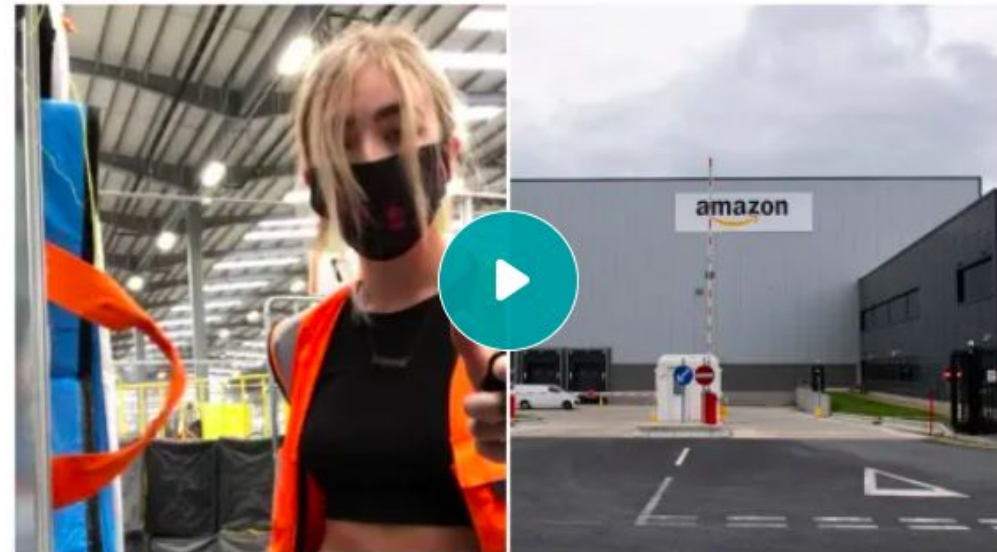
itv NEWS

REALITY OF WORK AT AMAZON

Zero-hour contracts and wrong wages -
the reality of being an Amazon agency
worker

HEALTH | BUSINESS | AMAZON | CORONAVIRUS

🕒 Thursday 18 February 2021, 11:03pm



<https://www.youtube.com/watch?v=xZfxWGZdHMs>



Matt Chase for BuzzFeed News

T
H
E
T
E
N
N
I
S
R
A
C
K
E
T

No DSS: Most flat shares refuse benefit claimants

🕒 9 March 2017 | 💬 Comments



Benefit claimants face landlord discrimination despite ruling

By Paul Bradshaw & Alex Homer
BBC Shared Data Unit

🕒 28 August 2020



<https://github.com/search?q=topic%3Adss+org%3ABBC-Data-Unit&type=Repositories>

Leicester House Share - Refurbished & Bills Paid

Ad Details

Message

New Ad ref# 8549557

Mark as unsuitable Save Share



House share
Leicester
LE3

£395 pcm (double)

£395 pcm (double)



Availability

Available	Now
Minimum term	6 months
Maximum term	None

Extra cost

Deposit (Room 1)	£350.00
Deposit (Room 2)	£350.00
Bills included?	Yes

Amenities

Furnishings	Furnished
Parking	No
Garage	No



The video gives an accurate reflection of the house as it currently is. NOTE, THE ROOM AVAILABLE IS NOT THE ROOM AT THE TOP OF THE STAIRS, the two rooms available are ground floor rooms. See photos for actual room. If you are interested having viewed the video then we can now arrange a viewing. A deposit will be required to secure the room.

We have NO HIDDEN FEES - A house which has been

New housemate preferences

Couples OK?	No
Smoking OK?	No
Pets OK?	No
Occupation	Professional
References?	Yes
Min age	18
Gender	Males or females

Why scraping?

1. Proof

2. Context

3. Exclusive leads



<https://www.youtube.com/watch?v=DWRGqmywNYs&t=160s>

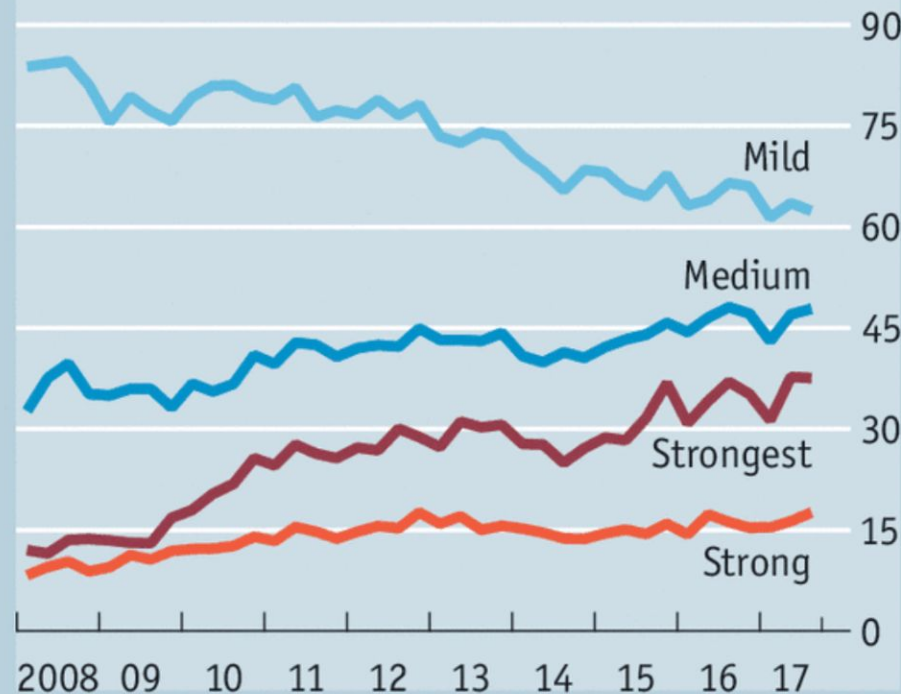
Foul-mouthed mothers are causing problems for Mumsnet

Sponsors fear that their adverts sit alongside ever-ruder content—and they are right

Mothercussers

Britain, offensive language* on Mumsnet, occurrences per 100,000 words

By swear-word strength, on the “Am I being unreasonable” discussion thread



Sources: Mumsnet; *The Economist*

By topic, most recent 50 discussions, Sept 1st 2017



*Defined by Ofcom, based on survey data

Why scraping?

1. Proof
2. Context
- 3. Exclusive leads**

Give MPs deadline on hiring relatives, campaigners urge

By Pete Sherlock and Paul Bradshaw
BBC News

🕒 31 July 2017 | 💬 Comments



One in five MPs continue to employ a member of their family using taxpayers' money despite the practice being banned for new members of Parliament.

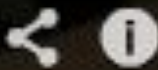
Of the 589 returning MPs, 122 have declared the employment of a relative in the **latest Register of Members' Financial Interests.**

None of the 61 new MPs who secured their seats at the general election on 8 June are allowed to do so.

<https://www.bbc.co.uk/news/uk-england-40709220>

<https://github.com/BBC-Data-Unit/mps-registers-of-interest>

Why is government website carrying fake jobs?



WATCH LIVE 7p

News

UK WORLD POLITICS BUSINESS SCIENCE TECHNOLOGY

FRIDAY | 07 FEBRUARY 2014 UK

Why is government website carrying fake jobs?

<https://www.youtube.com/watch?v=Efr-VEkwWoM>



Moment to shine

Olympic Games 27 July - 12 August
Official London 2012 website

Last gold medal



Asadauskaite L.

All medals



Spectators



Paralympic Games

[Results](#) | [Medals](#) | [Sports](#) | [Athletes](#) | [Countries](#) | [Join In](#) | [Venues](#) | [Torch Relay](#) | [Ceremonies](#) | [News](#) | [About us](#)

[About](#) | [Torchbearers](#) | [The Torch](#) | [History and tradition](#) | [News](#) | [Photos](#) | [Presenting Partners](#) | [Cymraeg](#)

Olympic Torchbearers



8,000 inspirational people carried the Olympic Flame across the UK. Nominated by someone they knew, it was their moment to shine, inspiring millions of people watching in their community, in the UK and worldwide.

Spotlight



Daley Thompson



Nathan Robertson



Steve Backley

Search Torchbearers

Date

Location

Select date



Search Torchbearers

Date

Location

Select date



NEWS | OPINION | **SPORT** | OLYMPICS | LIFE | PROPERTY | ARTS & ENTS | TRAVEL | MONEY

Athletics | Cricket | Football ▼ | Golf | Motor Racing | Olympics | Racing | Rugby League | Rugby Union ▼ | Sailing | Tennis

Hot Topics | [Euro 2012](#) | [Wimbledon](#) | [Olympics](#)

Olympics | Sport > Olympics

Seb Coe promised an 'uplifting torch relay to inspire a generation'. So are these really the role models to do it?

The intention was that 8,000 local heroes with tales to motivate young people would carry the flame around Britain. But the 2012 sponsors had their own ideas. Tom Peck reports

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science |

[RightMinds Home](#) | [News Board](#) | [Sport Boards](#) | [Showbiz Boards](#) | [Femail Boards](#) | [H](#)

Flames of Greed: How the Olympic been hijacked by commercialism

london2012

London 2012: torchbearers picked by sponsors keep flame of commerce alive

Companies sponsoring the Games have been awarded hundreds of slots in the torch relay. Their nominations include directors, clients – and a steel billionaire

James Ball

[guardian.co.uk](#), Wednesday 6 June 2012 14:09 BST

 [Comments \(28\)](#)



8,000 Holes

How the 2012
Olympic
Torch Relay
Lost its Way

All proceeds go to the Brittle Bone Society

Paul Bradshaw, Carol Miers
and users of Help Me Investigate

leanpub.com/8000holes

Tools

1. Chrome/Firefox extensions, e.g. [Instant Data Scraper](#), [Web Scraper](#)
2. Google Sheets: [IMPORTXML function](#)
3. [OutWit Hub](#) (e.g. job ads, tables on webpages)
4. [Octoparse](#), [Apify](#), [Browse AI](#)
5. API connectors: [Zapier](#), [IFTTT](#)
6. Code: Python or R or JavaScript or command line, etc. (Hacks/Hackers - find a developer)

If it doesn't work (block strategies)

It's not necessarily *you*.

1. Page might be dynamic (look for an undocumented API in Network tab)
2. Site might block basic scrapers (add User-Agent/Rotate User-Agents; add additional headers)
3. Site might have a limit on visits from IP address (use proxies; reconnect in Colab) or frequency of requests (slow down scraper)
4. Other options: 2captcha; become whitelisted

leanpub.com/u/paulbradshaw
@paulbradshaw
onlinejournalismblog.com
helpmeinvestigate.com
slideshare.net/onlinejournalist
linkedin.com/in/onlinejournalist