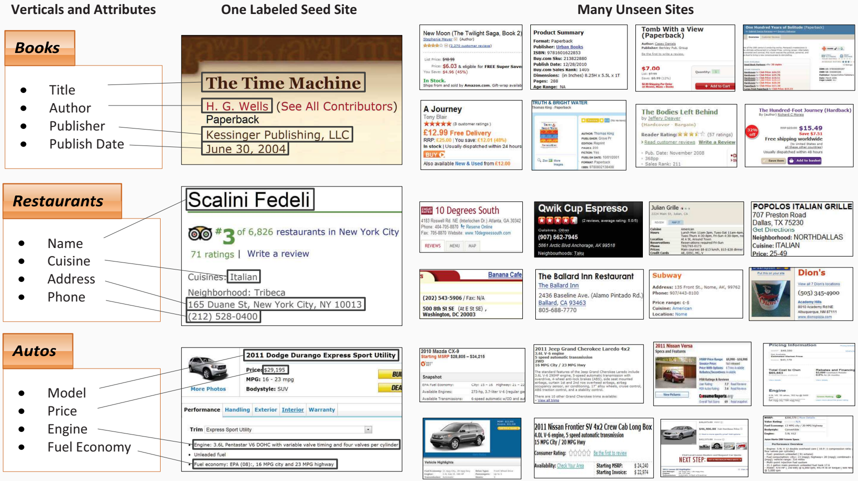
AI-based Structured Web Data Extraction

Jan Joneš

June 2022

Charles University

Web data extraction



Verticals and Attributes

One Labeled Seed Site

Many Unseen Sites

Books

- Title
- Author
- Publisher
- Publish Date

The Time Machine

H. G. Wells (See All Contributors)

Paperback

Kessinger Publishing, LLC

June 30, 2004

Restaurants

- Name
- Cuisine
- Address
- Phone

Scalini Fedeli

#3 of 6,826 restaurants in New York City

71 ratings | Write a review

Cuisines: Italian

Neighborhood: Tribeca

165 Duane St, New York City, NY 10013

(212) 528-0400

Autos

- Model
- Price
- Engine
- Fuel Economy

2011 Dodge Durango Express Sport Utility

Price: $29,195

MPG: 16 - 23 mpg

Bodystyle: SUV

More Photos

Performance | Handling | Exterior | Interior | Warranty

Trim: Express Sport Utility

- Engine: 3.6L Pentastar V6 DOHC with variable valve timing and four valves per cylinder
- Unleaded fuel
- Fuel economy: EPA (08):, 16 MPG city and 23 MPG highway

Q. Hao et al. “From One Tree to a Forest: a Unified Solution for Structured Web Data Extraction”. 2011

Motivation

- Scraping use-cases:
    - price comparison,
    - analysis,
    - creating datasets

Motivation

- Scraping use-cases:
  - price comparison,
  - analysis,
  - creating datasets
- Manual scrapers need manual labor

Approach

Text, DOM tree, visuals → Deep learning model

Datasets

SWDE (2011)

		price		economy		engine		model	
website	pages	values	nodes	values	nodes	values	nodes	values	nodes
aol	2,000	2,000	2,001	1,845	1,845	0	0	2,000	2,009
autobytel	2,000	1,994	1,994	1,816	1,816	2,000	2,000	3,998	4,023
automotive	1,999	1,999	1,999	1,999	1,999	1,999	1,999	1,999	1,999
autoweb	2,000	2,000	2,001	4,000	4,000	1,998	1,998	4,000	4,000
carquotes	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000
cars	657	647	741	607	607	1,168	1,168	657	683
kbb	2,000	2,000	4,732	2,000	2,000	2,000	2,938	4,000	4,000
motortrend	1,267	1,267	1,267	2,534	2,534	1,267	1,267	1,267	1,267
msn	2,000	3,901	4,007	3,623	4,081	2,000	2,000	2,000	3,177
yahoo	2,000	2,000	2,073	2,000	2,000	2,000	2,000	2,000	6,094
total	17,923	19,808	22,815	22,424	22,882	16,432	17,370	23,921	29,252

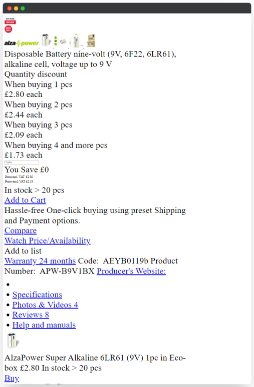
Apify e-commerce (2022)

website	pages	name	price	cat	images	short	long	spec
alza	2,493	2,493	2,478	2,493	2,493	2,493	2,493	2,477
asos	499	499	499	499	499	0	499	0
bestbuy	1,000	1,000	986	998	1,000	0	1,000	1
bloomingdales	448	462	447	448	448	0	1,087	0
conrad	1,500	1,500	1,499	1,500	1,485	1,500	1,500	1,495
etsy	250	250	250	250	250	0	1,000	0
ikea	1,972	1,972	1,972	1,959	1,972	2,261	1,917	0
notino	808	808	702	808	808	808	785	783
radioshack	499	499	499	499	498	0	6,204	4
tesco	1,500	1,499	1,465	1,499	1,499	0	17,205	21,526
total	10,969	10,982	10,797	10,953	10,952	7,062	33,690	26,286

Implementation

Visuals

Page without CSS or JavaScript evaluated is difficult to understand.



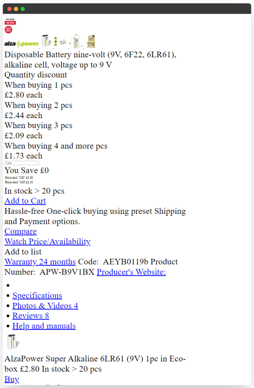
Disposable Battery nine-volt (9V, 6F22, 6LR61), alkaline cell, voltage up to 9 V
Quantity discount
When buying 1 pcs
£2.80 each
When buying 2 pcs
£2.44 each
When buying 3 pcs
£2.09 each
When buying 4 and more pcs
£1.73 each
1 pcs
You Save £0
Price incl. VAT £2.80
Price excl. VAT £2.33
In stock > 20 pcs
Add to Cart
Hassle-free One-click buying using preset Shipping and Payment options.
Compare
Watch Price/Availability
Add to list
Warranty 24 months Code: AEYB0119b Product Number: APW-B9V1BX Producer's Website:

-
- Specifications
- Photos & Videos 4
- Reviews 8
- Help and manuals

AlzaPower Super Alkaline 6LR61 (9V) 1pc in Eco-box £2.80 In stock > 20 pcs
Buy

Visuals

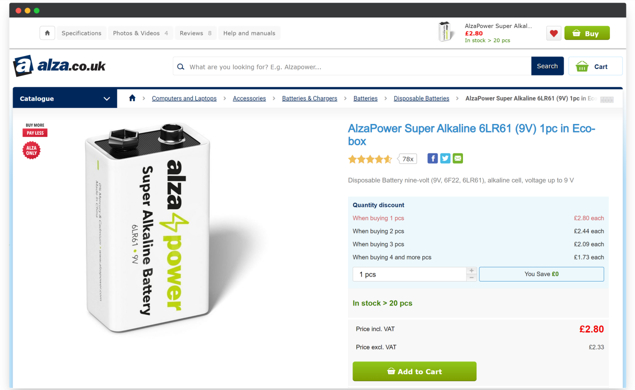
Page without CSS or JavaScript evaluated is difficult to understand.



Disposable Battery nine-volt (9V, 6F22, 6LR61), alkaline cell, voltage up to 9 V
Quantity discount
When buying 1 pcs
£2.80 each
When buying 2 pcs
£2.44 each
When buying 3 pcs
£2.09 each
When buying 4 and more pcs
£1.73 each
1 pcs
You Save £0
Price incl. VAT £2.80
Price excl. VAT £2.33
In stock > 20 pcs
Add to Cart
Hassle-free One-click buying using preset Shipping and Payment options.
Compare
Watch Price/Availability
Add to list
Warranty 24 months Code: AEYB0119b Product Number: APW-B9V1BX Producer's Website:

-
- Specifications
- Photos & Videos 4
- Reviews 8
- Help and manuals

AlzaPower Super Alkaline 6LR61 (9V) 1pc in Eco-box £2.80 In stock > 20 pcs
Buy



AlzaPower Super Alkaline 6LR61 (9V) 1pc in Eco-box

Disposable Battery nine-volt (9V, 6F22, 6LR61), alkaline cell, voltage up to 9 V

Quantity discount	
When buying 1 pcs	£2.80 each
When buying 2 pcs	£2.44 each
When buying 3 pcs	£2.09 each
When buying 4 and more pcs	£1.73 each

1 pcs
You Save £0

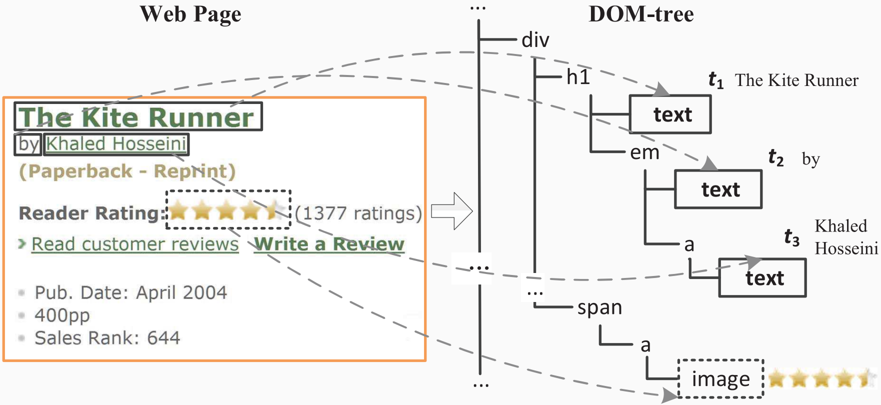
In stock > 20 pcs

Price incl. VAT	£2.80
Price excl. VAT	£2.33

Add to Cart

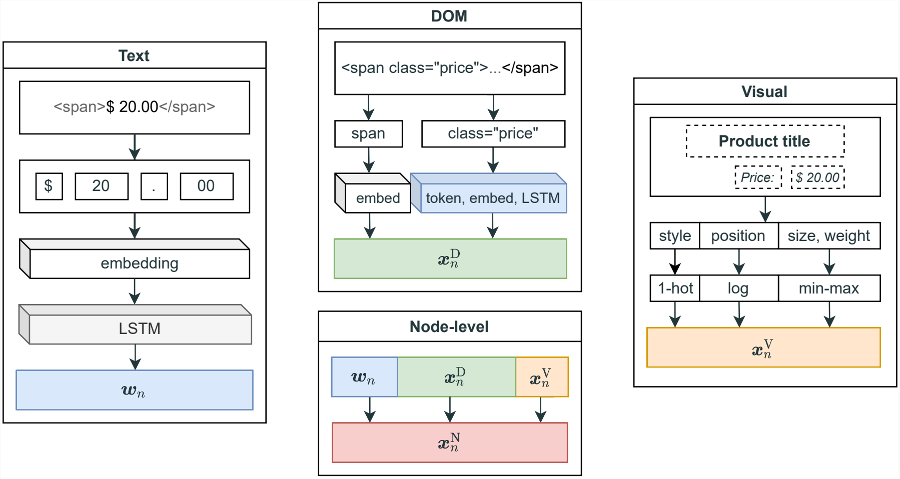
From www.alza.co.uk, archived at https://archive.ph/7xXoG

Extractor



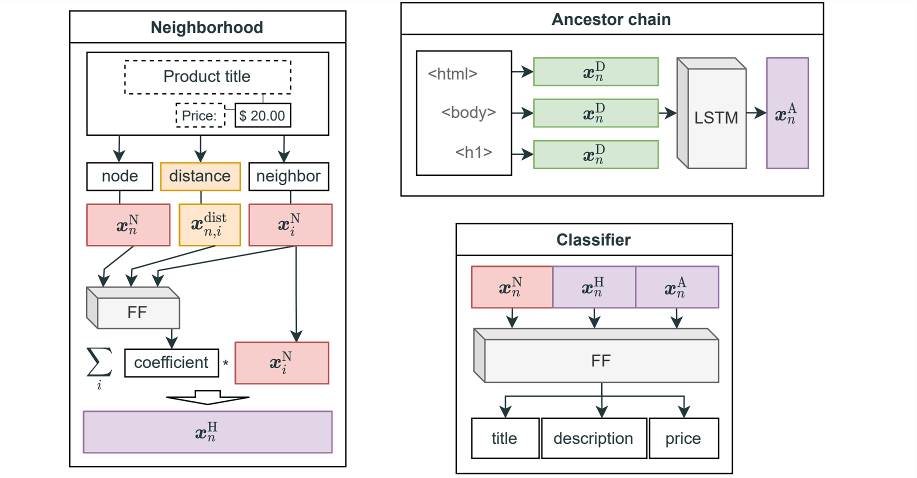
Q. Hao et al. “From One Tree to a Forest: a Unified Solution for Structured Web Data Extraction”. 2011

Features



Text
<span>$ 20.00</span>
$
20
.
00
embedding
LSTM
$\boldsymbol{w}_n$
DOM
<span class="price">...</span>
span
class="price"
embed
token, embed, LSTM
$\boldsymbol{x}_n^{\mathrm{D}}$
Node-level
$\boldsymbol{w}_n$
$\boldsymbol{x}_n^{\mathrm{D}}$
$\boldsymbol{x}_n^{\mathrm{V}}$
$\boldsymbol{x}_n^{\mathrm{N}}$
Visual
Product title
Price:
$ 20.00
style
position
size, weight
1-hot
log
min-max
$\boldsymbol{x}_n^{\mathrm{V}}$

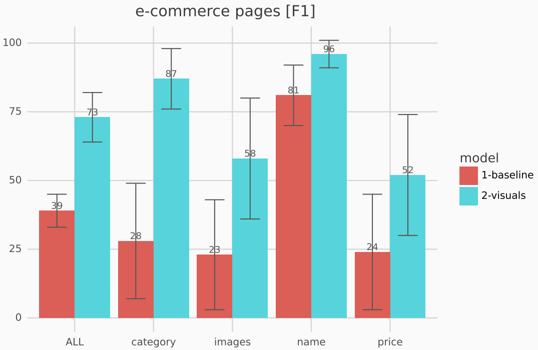
Model


Neighborhood
Product title
Price:
$ 20.00
node
distance
neighbor
$\boldsymbol{x}_n^{\mathrm{N}}$
$\boldsymbol{x}_{n,i}^{\mathrm{dist}}$
$\boldsymbol{x}_i^{\mathrm{N}}$
FF
$\sum_i$
coefficient
*
$\boldsymbol{x}_i^{\mathrm{N}}$
$\boldsymbol{x}_n^{\mathrm{H}}$
Ancestor chain
<html>
<body>
<h1>
$\boldsymbol{x}_n^{\mathrm{D}}$
$\boldsymbol{x}_n^{\mathrm{D}}$
$\boldsymbol{x}_n^{\mathrm{D}}$
LSTM
$\boldsymbol{x}_n^{\mathrm{A}}$
Classifier
$\boldsymbol{x}_n^{\mathrm{N}}$
$\boldsymbol{x}_n^{\mathrm{H}}$
$\boldsymbol{x}_n^{\mathrm{A}}$
FF
title
description
price

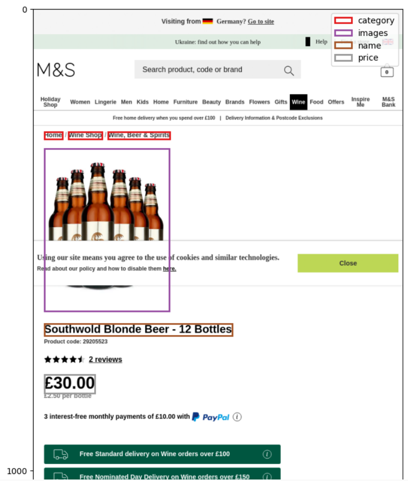
Baseline (2021)

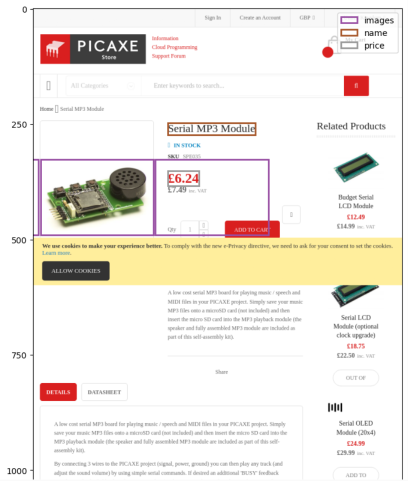
- No visuals
- Old dataset
- Non-separated test set

Results


e-commerce pages [F1]
100
75
50
25
0
39
73
28
87
23
58
81
96
24
52
ALL
category
images
name
price
model
1-baseline
2-visuals

Demo


0
1000
category
images
name
price
Visiting from Germany? Go to site
Ukraine: find out how you can help
Help
M&S
Search product, code or brand
Holiday Shop
Women
Lingerie
Men
Kids
Home
Furniture
Beauty
Brands
Flowers
Gifts
Wine
Food
Offers
Inspire Me
M&S Bank
Free home delivery when you spend over £100 | Delivery Information & Postcode Exclusions
Home / Wine Shop / Wine, Beer & Spirits
Using our site means you agree to the use of cookies and similar technologies.
Read about our policy and how to disable them here.
Close
Southwold Blonde Beer - 12 Bottles
Product code: 29205523
2 reviews
£30.00
£2.50 per bottle
3 interest-free monthly payments of £10.00 with PayPal
Free Standard delivery on Wine orders over £100
Free Nominated Day Delivery on Wine orders over £150


0
250
500
750
1000
images
name
price
Sign In
Create an Account
GBP
PICAXE Store
Information
Cloud Programming
Support Forum
All Categories
Enter keywords to search...
Home Serial MP3 Module
Serial MP3 Module
IN STOCK
SKU SPE035
£6.24
£7.49 inc. VAT
Qty 1
ADD TO CART
Related Products
Budget Serial LCD Module
£12.49
£14.99 inc. VAT
We use cookies to make your experience better. To comply with the new e-Privacy directive, we need to ask for your consent to set the cookies. Learn more.
ALLOW COOKIES
A low cost serial MP3 board for playing music / speech and MIDI files in your PICAXE project. Simply save your music MP3 files onto a microSD card (not included) and then insert the micro SD card into the MP3 playback module (the speaker and fully assembled MP3 module are included as part of this self-assembly kit).
Serial LCD Module (optional clock upgrade)
£18.75
£22.50 inc. VAT
Share
OUT OF
DETAILS
DATASHEET
Serial OLED Module (20x4)
£24.99
£29.99 inc. VAT
ADD TO
By connecting 3 wires to the PICAXE project (signal, power, ground) you can then play any track (and adjust the sound volume) by using simple serial commands. If desired an additional 'BUSY' feedback

Generated by live demo at `bit.ly/awedemo`

Summary

- Node.js extractor of visual features

Summary

- Node.js extractor of visual features
- Python deep learning model

Summary

- Node.js extractor of visual features
- Python deep learning model
- Evaluated on a modern dataset

Summary

- Node.js extractor of visual features
- Python deep learning model
- Evaluated on a modern dataset
- Docker image for reproducible training

Summary

- Node.js extractor of visual features
- Python deep learning model
- Evaluated on a modern dataset
- Docker image for reproducible training
- Demo app for live inference

Thank you for your attention

bit.ly/awedemo

github.com/jjonescz/awe