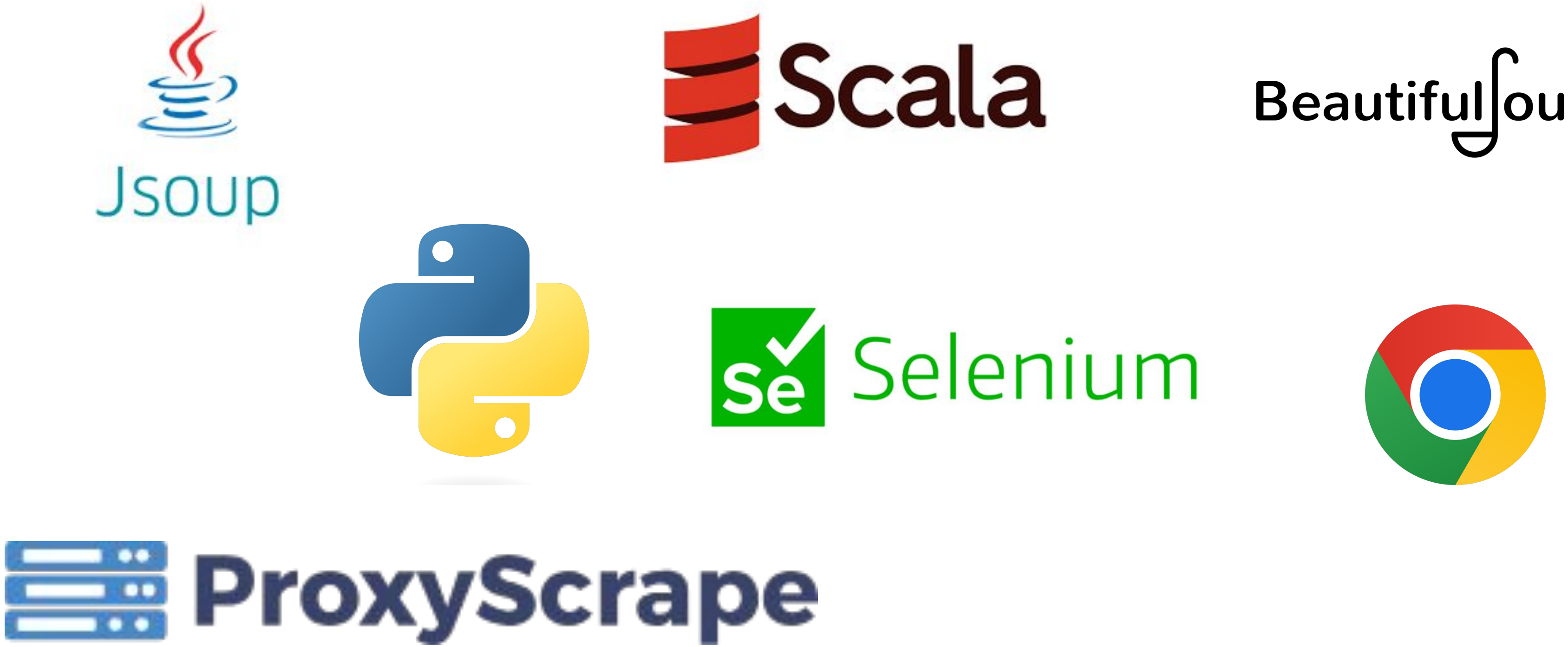
# FunPar Final Project
# Parallel Web Scraping

Porter Fredrickson

# The Basic Task

- Scrape lots of data from the web which would normally take long in a sequential fashion
- Analyze the data in some useful way
- Find a way to speedup inherently slow sequential code

# Technologies used

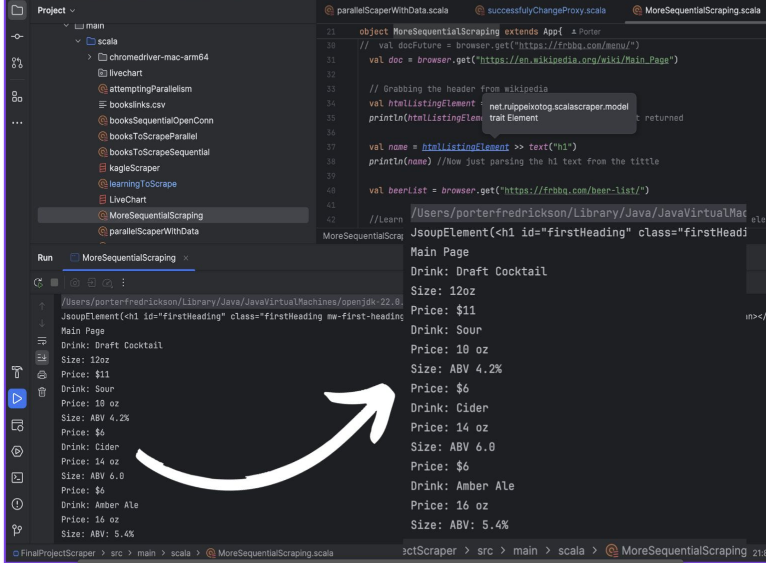


# Phase 1: Figuring out how to Scrape

- My only experience “web scraping” was from assignment 3 in this class
- First thing I needed to do was learn exactly what web scraping is and how to do it
- Followed various scraping with Scala + JSoup tutorials

# Went on scraping journey

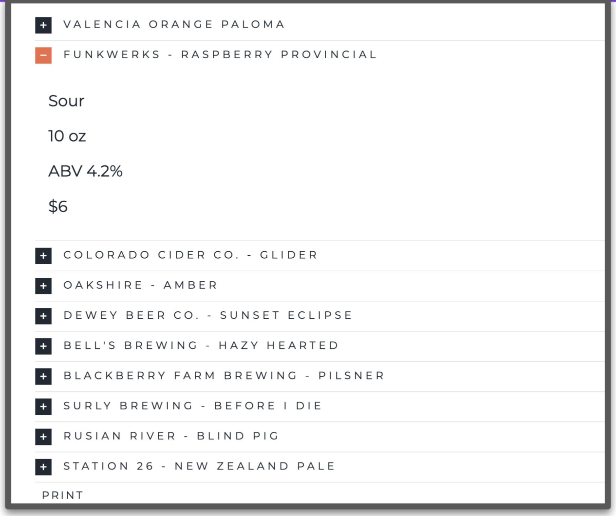
- After learning the basics with Jsoup and Scala I began testing my skills
- Connected to the website of my previous employer
- Scraped menu items
- Tried scraping other more secure websites and ran into my first blocks



```scala
object MoreSequentialScraping extends App{
//  val docFuture = browser.get("https://frbbq.com/menu/")
  val doc = browser.get("https://en.wikipedia.org/wiki/Main_Page")

  // Grabbing the header from wikipedia
  val htmlListingElement =
  println(htmlListingEleme                              t returned

  val name = htmlListingElement >> text("h1")
  println(name) //Now just parsing the h1 text from the tittle

  val beerList = browser.get("https://frbbq.com/beer-list/")

  //Learn
```

```
/Users/porterfredrickson/Library/Java/JavaVirtualMachines/openjdk-22.0.
JsoupElement(<h1 id="firstHeading" class="firstHeading mw-first-heading
Main Page
Drink: Draft Cocktail
Size: 12oz
Price: $11
Drink: Sour
Price: 10 oz
Size: ABV 4.2%
Price: $6
Drink: Cider
Price: 14 oz
Size: ABV 6.0
Price: $6
Drink: Amber Ale
Price: 16 oz
Size: ABV: 5.4%
```

```
/Users/porterfredrickson/Library/Java/JavaVirtualMac
JsoupElement(<h1 id="firstHeading" class="firstHeadi
Main Page
Drink: Draft Cocktail
Size: 12oz
Price: $11
Drink: Sour
Price: 10 oz
Size: ABV 4.2%
Price: $6
Drink: Cider
Price: 14 oz
Size: ABV 6.0
Price: $6
Drink: Amber Ale
Price: 16 oz
Size: ABV: 5.4%
```



+ VALENCIA ORANGE PALOMA

− FUNKWERKS - RASPBERRY PROVINCIAL

Sour

10 oz

ABV 4.2%

$6

+ COLORADO CIDER CO. - GLIDER

+ OAKSHIRE - AMBER

+ DEWEY BEER CO. - SUNSET ECLIPSE

+ BELL'S BREWING - HAZY HEARTED

+ BLACKBERRY FARM BREWING - PILSNER

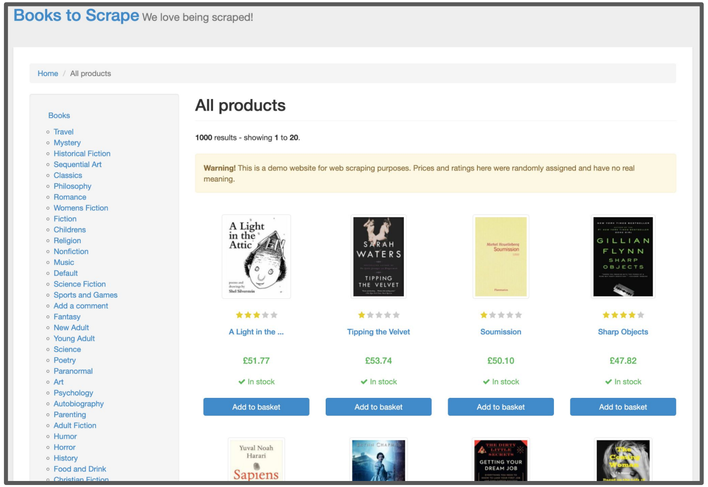
+ SURLY BREWING - BEFORE I DIE

+ RUSIAN RIVER - BLIND PIG

+ STATION 26 - NEW ZEALAND PALE

PRINT

# Books to scrape



**Books to Scrape** We love being scraped!

Home / All products

Books

- Travel
- Mystery
- Historical Fiction
- Sequential Art
- Classics
- Philosophy
- Romance
- Womens Fiction
- Fiction
- Childrens
- Religion
- Nonfiction
- Music
- Default
- Science Fiction
- Sports and Games
- Add a comment
- Fantasy
- New Adult
- Young Adult
- Science
- Poetry
- Paranormal
- Art
- Psychology
- Autobiography
- Parenting
- Adult Fiction
- Humor
- Horror
- History
- Food and Drink
- Christian Fiction

## All products

**1000** results - showing **1** to **20**.

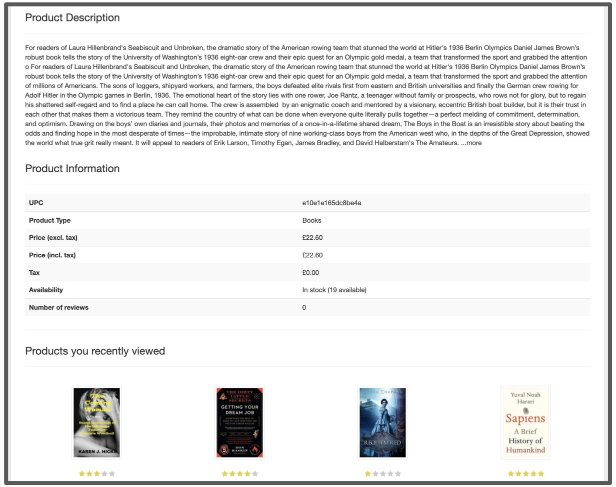
**Warning!** This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.

★★★☆☆ A Light in the ... £51.77 ✓ In stock [Add to basket]

★☆☆☆☆ Tipping the Velvet £53.74 ✓ In stock [Add to basket]

★☆☆☆☆ Soumission £50.10 ✓ In stock [Add to basket]

★★★★☆ Sharp Objects £47.82 ✓ In stock [Add to basket]



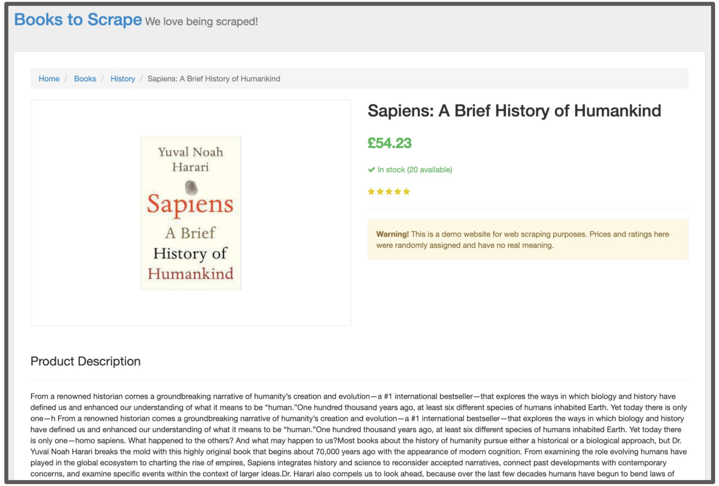
## Product Description

For readers of Laura Hillenbrand's Seabiscuit and Unbroken, the dramatic story of the American rowing team that stunned the world at Hitler's 1936 Berlin Olympics Daniel James Brown's robust book tells the story of the University of Washington's 1936 eight-oar crew and their epic quest for an Olympic gold medal, a team that transformed the sport and grabbed the attention o For readers of Laura Hillenbrand's Seabiscuit and Unbroken, the dramatic story of the American rowing team that stunned the world at Hitler's 1936 Berlin Olympics Daniel James Brown's robust book tells the story of the University of Washington's 1936 eight-oar crew and their epic quest for an Olympic gold medal, a team that transformed the sport and grabbed the attention of millions of Americans. The sons of loggers, shipyard workers, and farmers, the boys defeated elite rivals first from eastern and British universities and finally the German crew rowing for Adolf Hitler in the Olympic games in Berlin, 1936. The emotional heart of the story lies with one rower, Joe Rantz, a teenager without family or prospects, who rows not for glory, but to regain his shattered self-regard and to find a place he can call home. The crew is assembled by an enigmatic coach and mentored by a visionary, eccentric British boat builder, but it is their trust in each other that makes them a victorious team. They remind the country of what can be done when everyone quite literally pulls together—a perfect melding of commitment, determination, and optimism. Drawing on the boys' own diaries and journals, their photos and memories of a once-in-a-lifetime shared dream, The Boys in the Boat is an irresistible story about beating the odds and finding hope in the most desperate of times—the improbable, intimate story of nine working-class boys from the American west who, in the depths of the Great Depression, showed the world what true grit really meant. It will appeal to readers of Erik Larson, Timothy Egan, James Bradley, and David Halberstam's The Amateurs. ...more

## Product Information

| | |
|---|---|
| **UPC** | e10e1e165dc8be4a |
| **Product Type** | Books |
| **Price (excl. tax)** | £22.60 |
| **Price (incl. tax)** | £22.60 |
| **Tax** | £0.00 |
| **Availability** | In stock (19 available) |
| **Number of reviews** | 0 |

## Products you recently viewed



**Books to Scrape** We love being scraped!

Home / Books / History / Sapiens: A Brief History of Humankind

## Sapiens: A Brief History of Humankind

£54.23

✓ In stock (20 available)

★★★★★

**Warning!** This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.

## Product Description

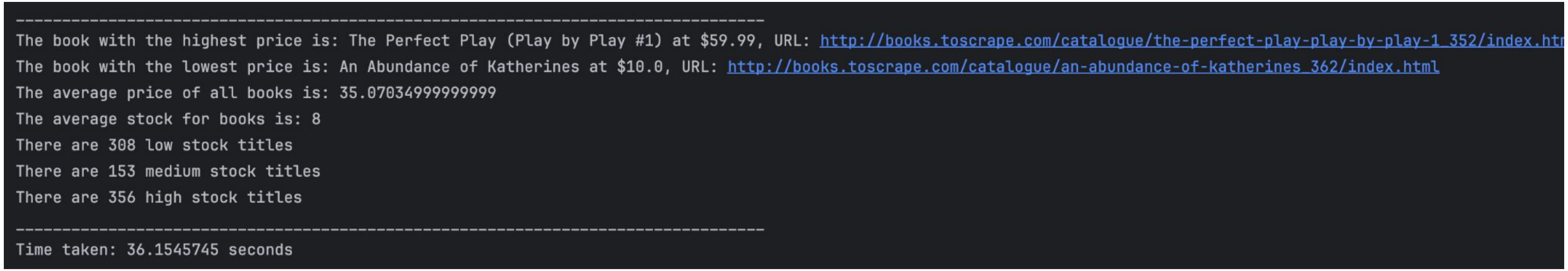
From a renowned historian comes a groundbreaking narrative of humanity's creation and evolution—a #1 international bestseller—that explores the ways in which biology and history have defined us and enhanced our understanding of what it means to be "human."One hundred thousand years ago, at least six different species of humans inhabited Earth. Yet today there is only one—h From a renowned historian comes a groundbreaking narrative of humanity's creation and evolution—a #1 international bestseller—that explores the ways in which biology and history have defined us and enhanced our understanding of what it means to be "human."One hundred thousand years ago, at least six different species of humans inhabited Earth. Yet today there is only one—homo sapiens. What happened to the others? And what may happen to us?Most books about the history of humanity pursue either a historical or a biological approach, but Dr. Yuval Noah Harari breaks the mold with this highly original book that begins about 70,000 years ago with the appearance of modern cognition. From examining the role evolving humans have played in the global ecosystem to charting the rise of empires, Sapiens integrates history and science to reconsider accepted narratives, connect past developments with contemporary concerns, and examine specific events within the context of larger ideas.Dr. Harari also compels us to look ahead, because over the last few decades humans have begun to bend laws of

# Books to Scrape Cont.

- Takes around ~315 seconds to scrape 1000’s pages sequentially.
- No aggregate calculations, only storing price and title
- With Open session: ~307 seconds
- Minimal 8 second speedup
- Obviously slow

# Books to Scrape Parallel

- Scraped the same 1000 book pages
- Obtained useful data beyond title such as price, stock,
- Calculated aggregate statistics
  - Max book price = $59.99 and Min book price = $10.00
  - Average Stock = 8 books
  - Low stock count = 308 titles
  - Medium stock count = 153 titles
  - High stock count = 356 titles
- All said in done the parallel solution which did considerable amounts of work more than the sequential solution ran in **36 seconds**



```
---------------------------------------------------------------------------------
The book with the highest price is: The Perfect Play (Play by Play #1) at $59.99, URL: http://books.toscrape.com/catalogue/the-perfect-play-play-by-play-1_352/index.htm
The book with the lowest price is: An Abundance of Katherines at $10.0, URL: http://books.toscrape.com/catalogue/an-abundance-of-katherines_362/index.html
The average price of all books is: 35.07034999999999
The average stock for books is: 8
There are 308 low stock titles
There are 153 medium stock titles
There are 356 high stock titles

---------------------------------------------------------------------------------
Time taken: 36.1545745 seconds
```

# Proxies

- On a mission to compare my scraped data with Amazon’s data
- Successfully searched book titles on amazon and obtained their price
- However, when scaling to 1000 books/requests I got blocked
- Introduced timeouts - partially successful
- Discovered the concept of using proxies (also recommended by Ajarn)
- Used a free proxy API
- Checked proxy validity in parallel
- Wrote valid proxies out to a csv file
- Checking over 1200 proxies takes 193.66 seconds
- A sequential solution should take about 75 minutes!

# Demo

# Difficulties

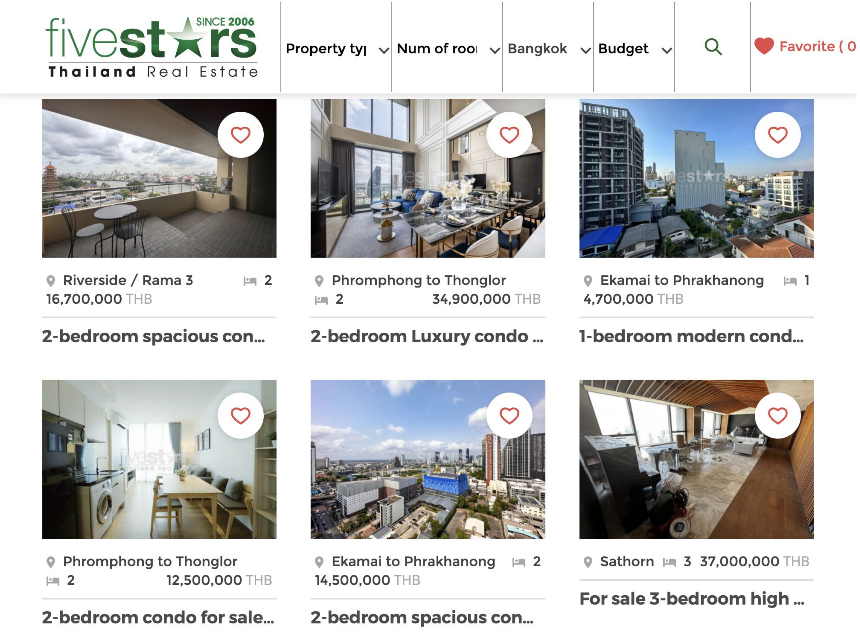
- Numerous difficulties and learning points
- Scala isn’t the greatest to scrape in
- Obtaining large amounts of content from web pages can be very hard because it’s dynamically loaded.
- High traffic sites are very difficult to scrape (Amazon)
- One solution to being blocked is proxies
- There are free proxies and paid for proxies
- Free proxies are for the most part useless
    - Most don't work
    - For those that do, major sites already block the free proxies

# Using Selenium  Python

- Determined to find more success and learn more about web scraping
- Selenium Webdriver is a software used for browser automation
- Works well with python
- Capable of loading dynamic content
- Used to scrape fivestars-thailand realty website
- Has realty listings for multiple areas in Thailand

fivestars SINCE 2006

Thailand Real Estate



Riverside / Rama 3 2
16,700,000 THB
2-bedroom spacious con...

Phromphong to Thonglor 2
34,900,000 THB
2-bedroom Luxury condo ...

Ekamai to Phrakhanong 1
4,700,000 THB
1-bedroom modern cond...

Phromphong to Thonglor 2
12,500,000 THB
2-bedroom condo for sale...

Ekamai to Phrakhanong 2
14,500,000 THB
2-bedroom spacious con...

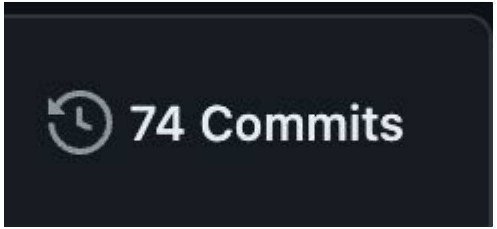
Sathorn 3 37,000,000 THB
For sale 3-bedroom high ...

# Another Demo

# Ultimate outcome

- Produced many different parallel web scraping solutions for this final project
- What better way to test them than to run everything at once
- Scraping 1000 website from books to scrape, all the current properties in bangkok, and all of the current properties in Phuket
- Doing all of this sequentially takes a gruelling ~370 seconds
- Doing all of this in parallel with Scala (futures) takes 54 seconds
- Shows a massive 6.85X speedup and a huge success

# How I was successful


74 Commits

Definition of Success in Proposal:

“The first component of success for this project is developing a web scraper that connects to a website (or list of websites) and obtains target data.”

“The second component has a wider definition of success. This concurrent solution can be successful in several ways. First, the concurrent solution could allow me to scrape multiple sites at once at a similar rate to the first solution. Second, the concurrent solution could be successful at scraping the same data as the first solution, but much faster. Finally, the third solution could be successful in scraping so fast that it actually gets blocked by the website for making too many requests.”

“Finally, the most important goals of this project are to learn about web scraping since I currently have no scaping knowledge and to learn about creating my own concurrent speedups.”