Finding Shoe Stores in >100k Merchants: Using Spark to Group All Things

Solmaz Shahalizadeh (@solmaz_sh)

Shopify

Spark SUMMIT EAST

About me

Currently:

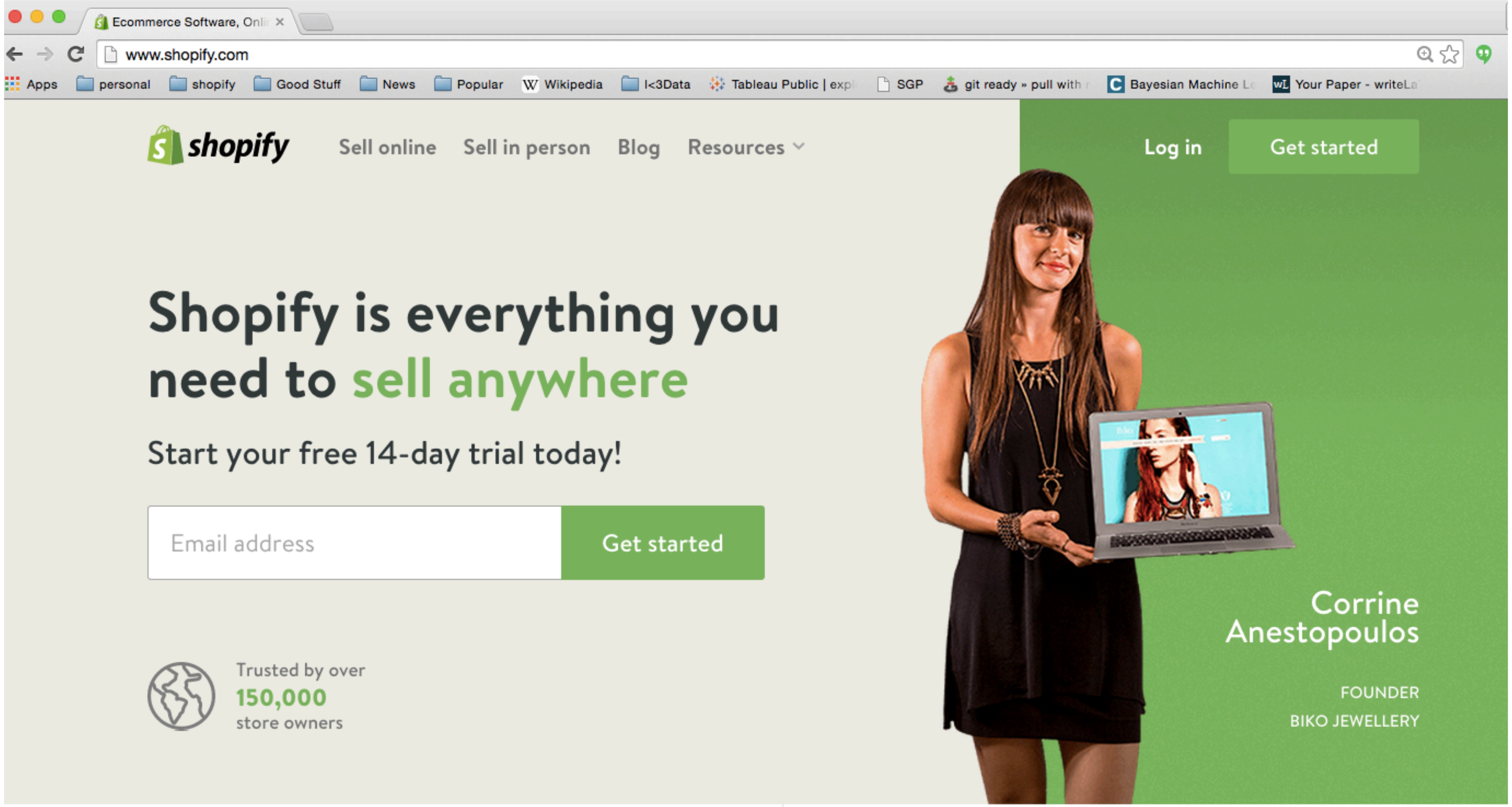
- Finance Data @Shopify

Previous Lives:

- Playing with data in Finance/ Bioinformatics/ Cancer Research

Will Talk about:

- Finding a needle in a haystack
- Trying all the wrong tools for getting insights out of data and course correction on the way
- Having fun during the process



Shopify is everything you need to sell anywhere

Start your free 14-day trial today!

Trusted by over 150,000 store owners

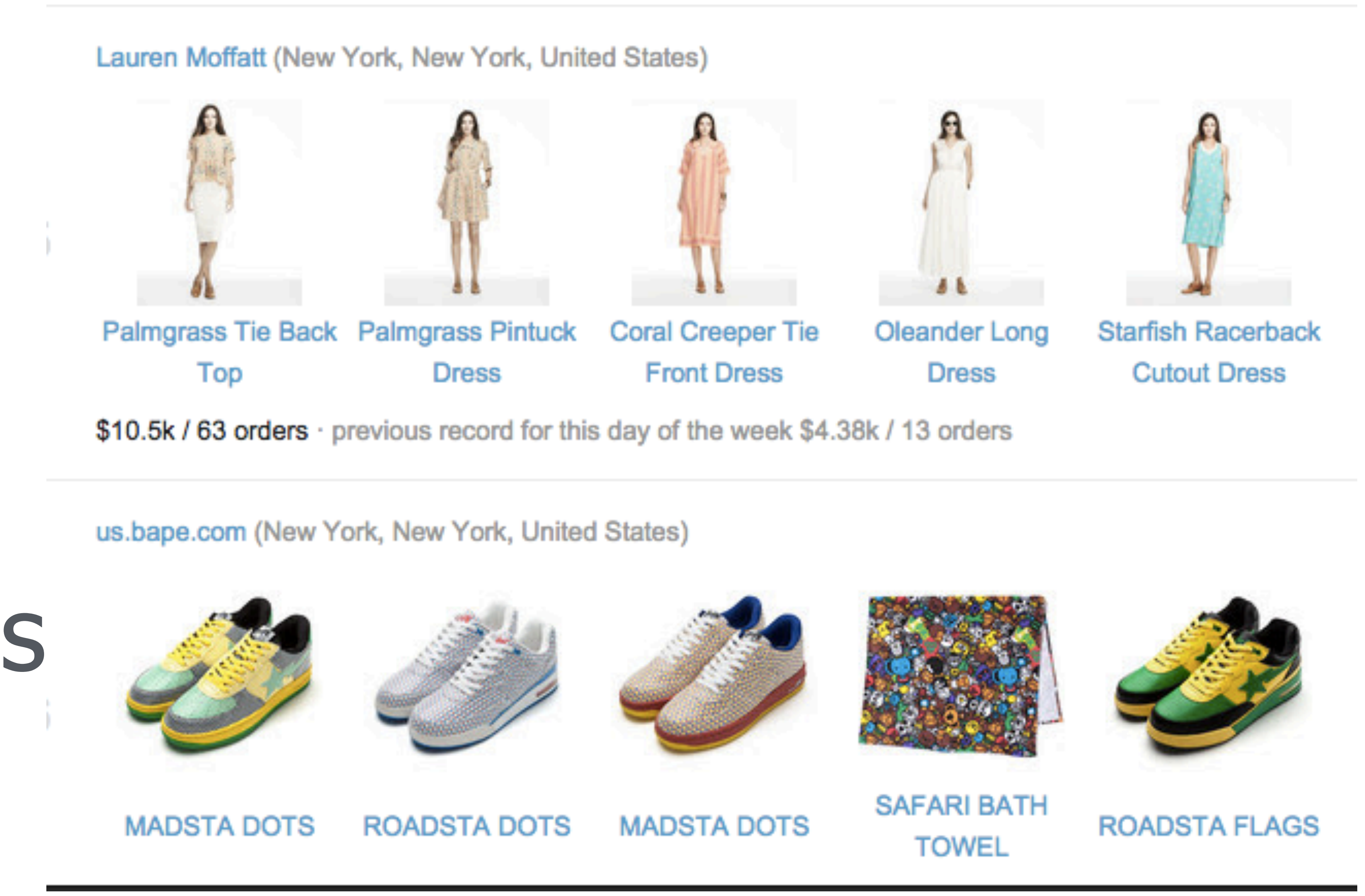
Corrine Anestopoulos

FOUNDER

BIKO JEWELLERY

Where are the shoes?

- Started ~ 1 year ago @Shopify
- Wanted desperately to buy ~~something~~ shoes from our merchants



Lauren Moffatt (New York, New York, United States)

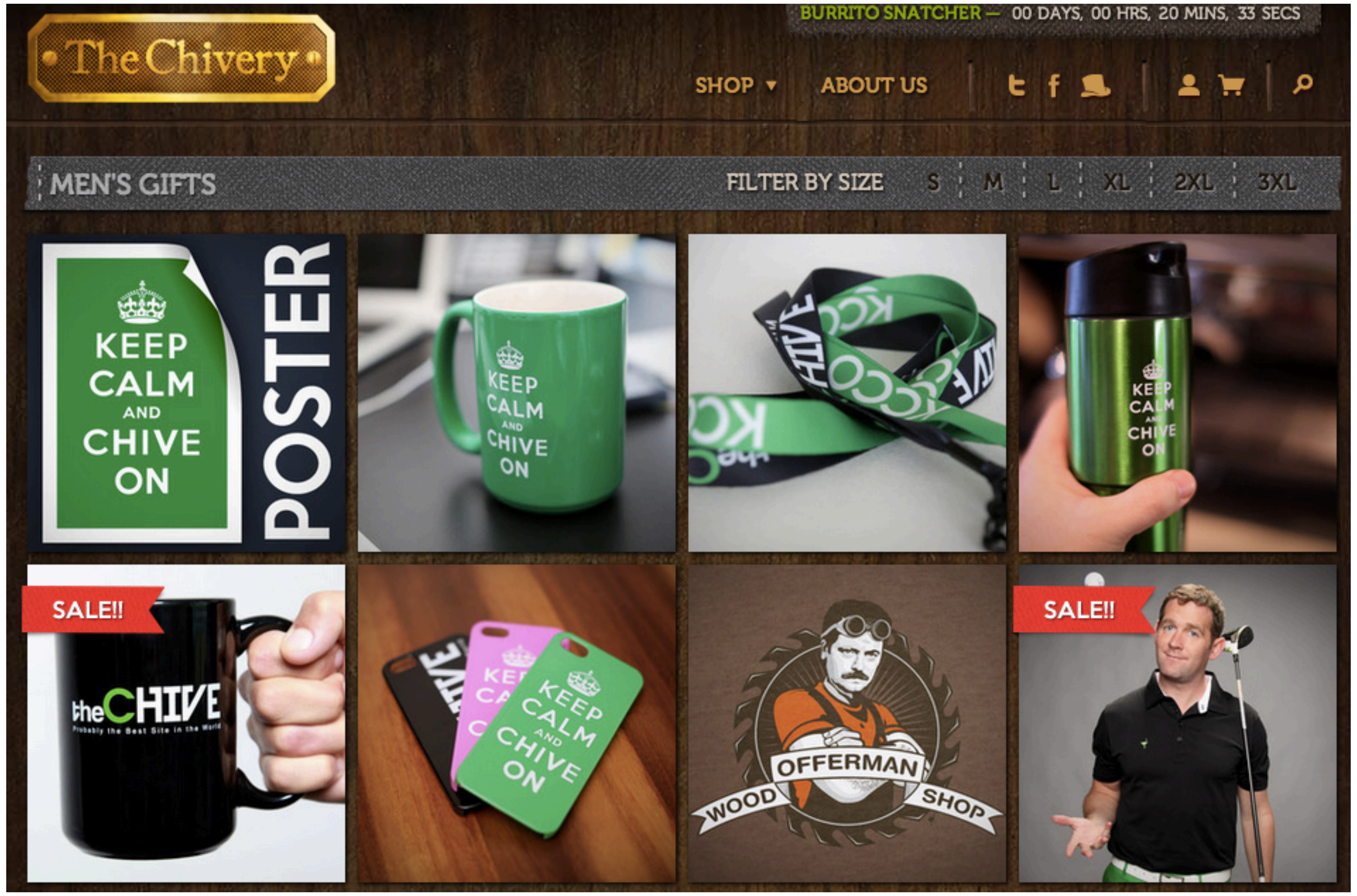
$10.5k / 63 orders · previous record for this day of the week $4.38k / 13 orders

us.bape.com (New York, New York, United States)

A bit about Shopify stores

Merchants can sell different

kinds of products in a single store

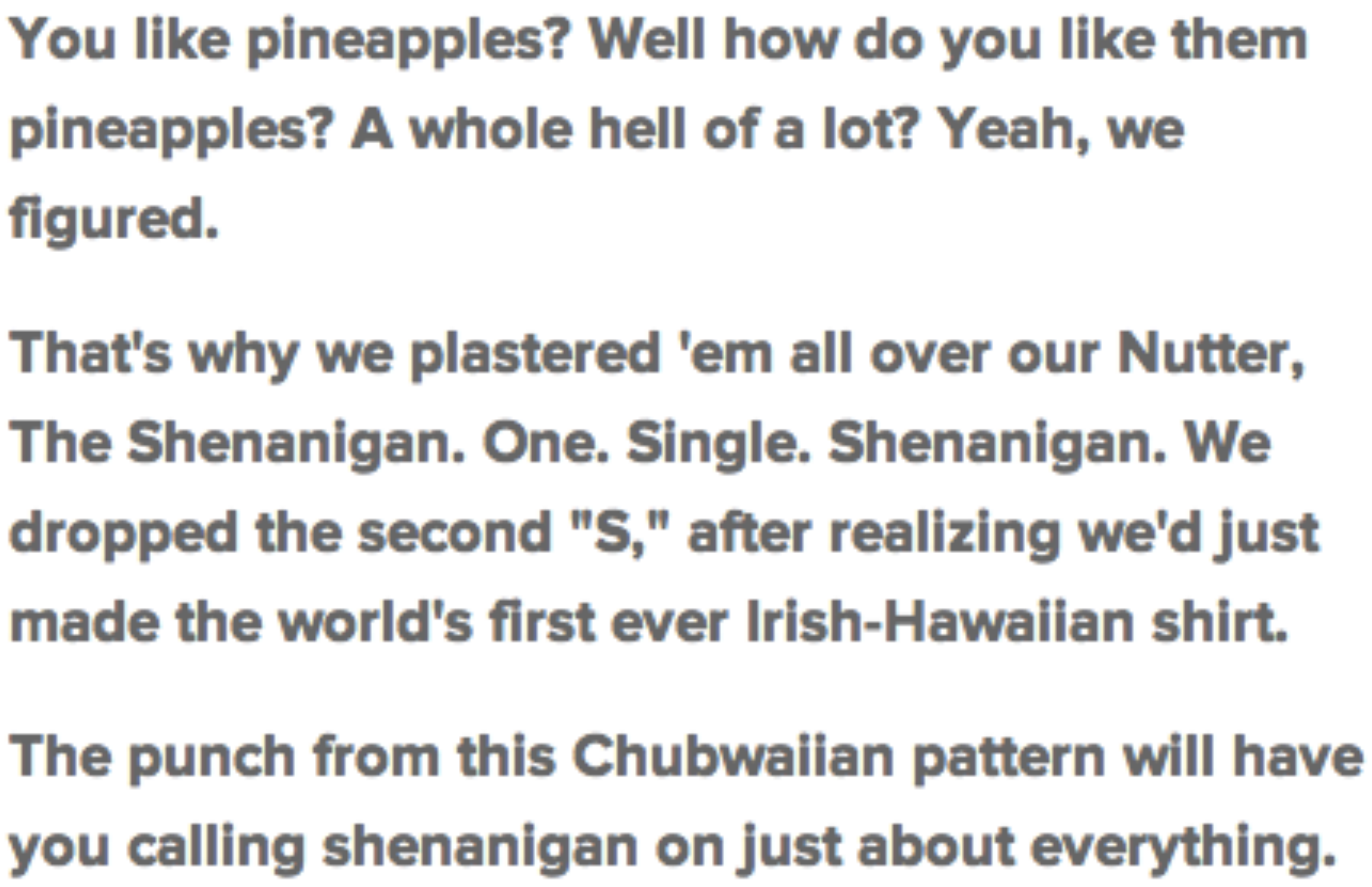


More than 60M products

We can give each person in Ottawa 67 products

A bit about Shopify stores

There is freedom of speech in describing a product



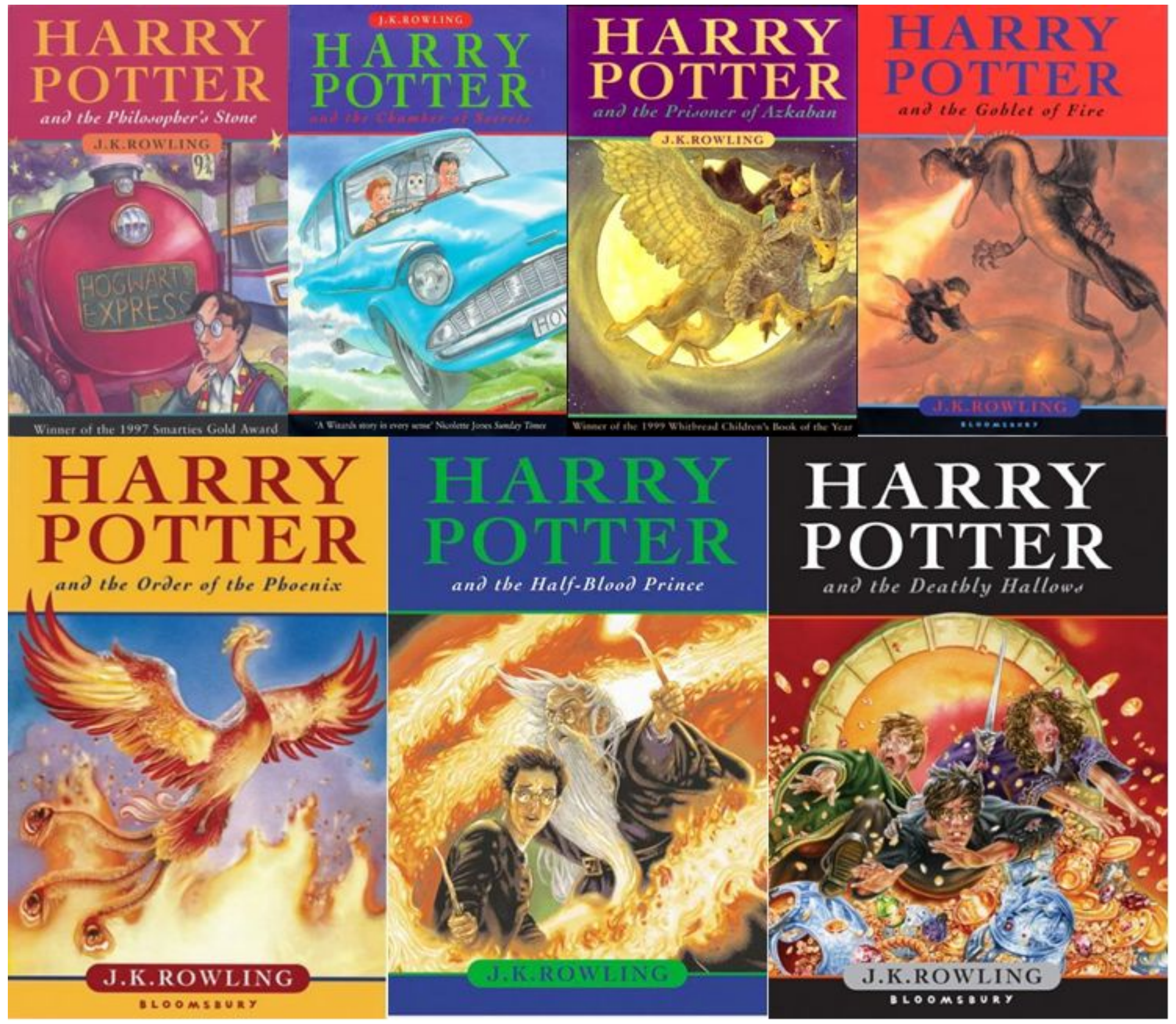
You like pineapples? Well how do you like them pineapples? A whole hell of a lot? Yeah, we figured.

That's why we plastered 'em all over our Nutter, The Shenanigan. One. Single. Shenanigan. We dropped the second "S," after realizing we'd just made the world's first ever Irish-Hawaiian shirt.

The punch from this Chubwaiian pattern will have you calling shenanigan on just about everything.

Too much text to process

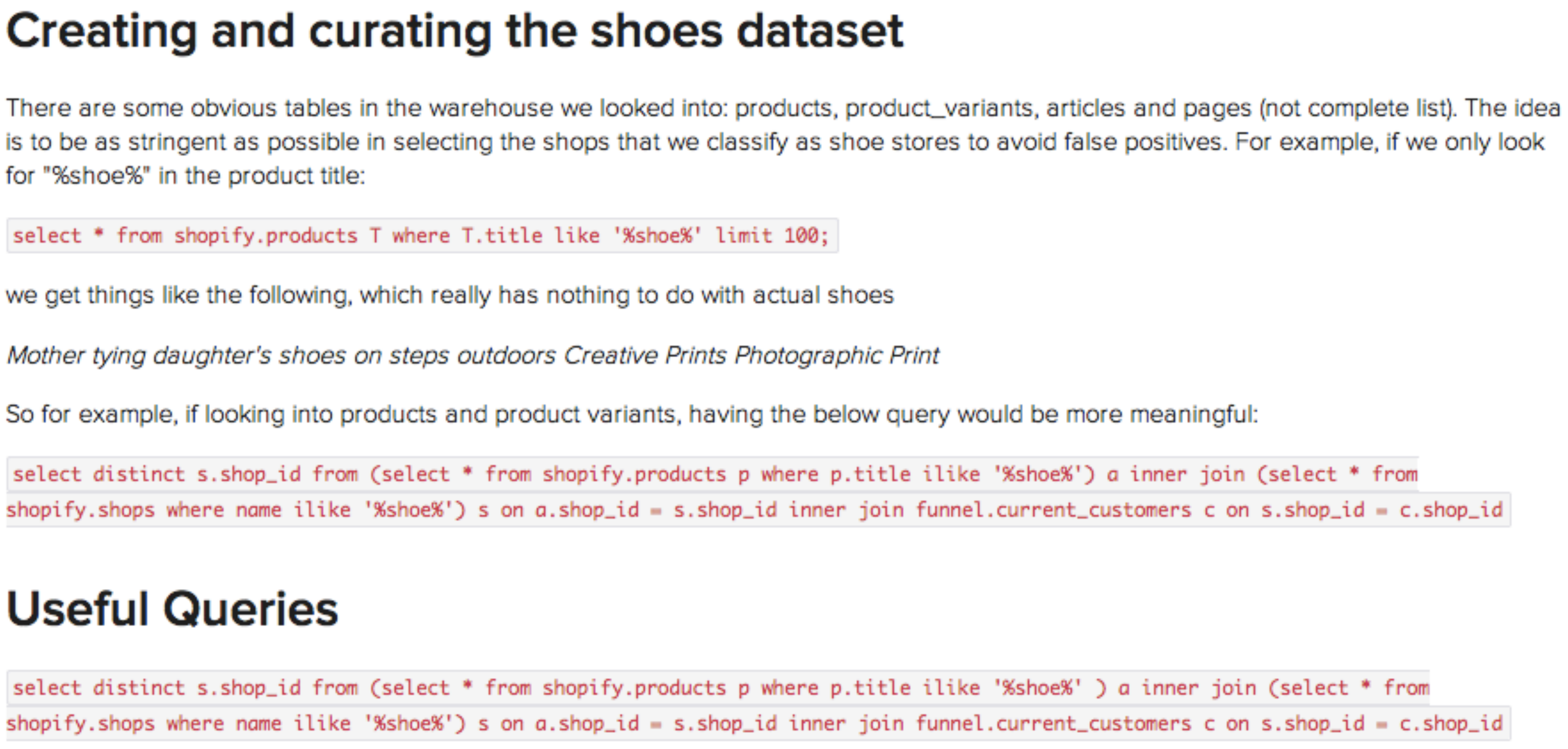
Product name, description, vendor, etc.


HARRY POTTER and the Philosopher's Stone
HARRY POTTER and the Chamber of Secrets
HARRY POTTER and the Prisoner of Azkaban
HARRY POTTER and the Goblet of Fire
HARRY POTTER and the Order of the Phoenix
HARRY POTTER and the Half-Blood Prince
HARRY POTTER and the Deathly Hallows
J.K.ROWLING

✖ 7,400

Shoe Mining

Its going to be a big win and so much FUN!!!



Creating and curating the shoes dataset

There are some obvious tables in the warehouse we looked into: products, product_variants, articles and pages (not complete list). The idea is to be as stringent as possible in selecting the shops that we classify as shoe stores to avoid false positives. For example, if we only look for "%shoe%" in the product title:

```
select * from shopify.products T where T.title like '%shoe%' limit 100;
```

we get things like the following, which really has nothing to do with actual shoes

Mother tying daughter's shoes on steps outdoors Creative Prints Photographic Print

So for example, if looking into products and product variants, having the below query would be more meaningful:

```
select distinct s.shop_id from (select * from shopify.products p where p.title ilike '%shoe%') a inner join (select * from
shopify.shops where name ilike '%shoe%') s on a.shop_id = s.shop_id inner join funnel.current_customers c on s.shop_id = c.shop_id
```

Useful Queries

```
select distinct s.shop_id from (select * from shopify.products p where p.title ilike '%shoe%' ) a inner join (select * from
shopify.shops where name ilike '%shoe%') s on a.shop_id = s.shop_id inner join funnel.current_customers c on s.shop_id = c.shop_id
```

It was a total success

It was a total ~~success~~

failure

Problems:

- No distributed data
- Not enough processing power
- Not very smart filters
- Not many examples of actual stores selling shoes

Shopify + Pinterest

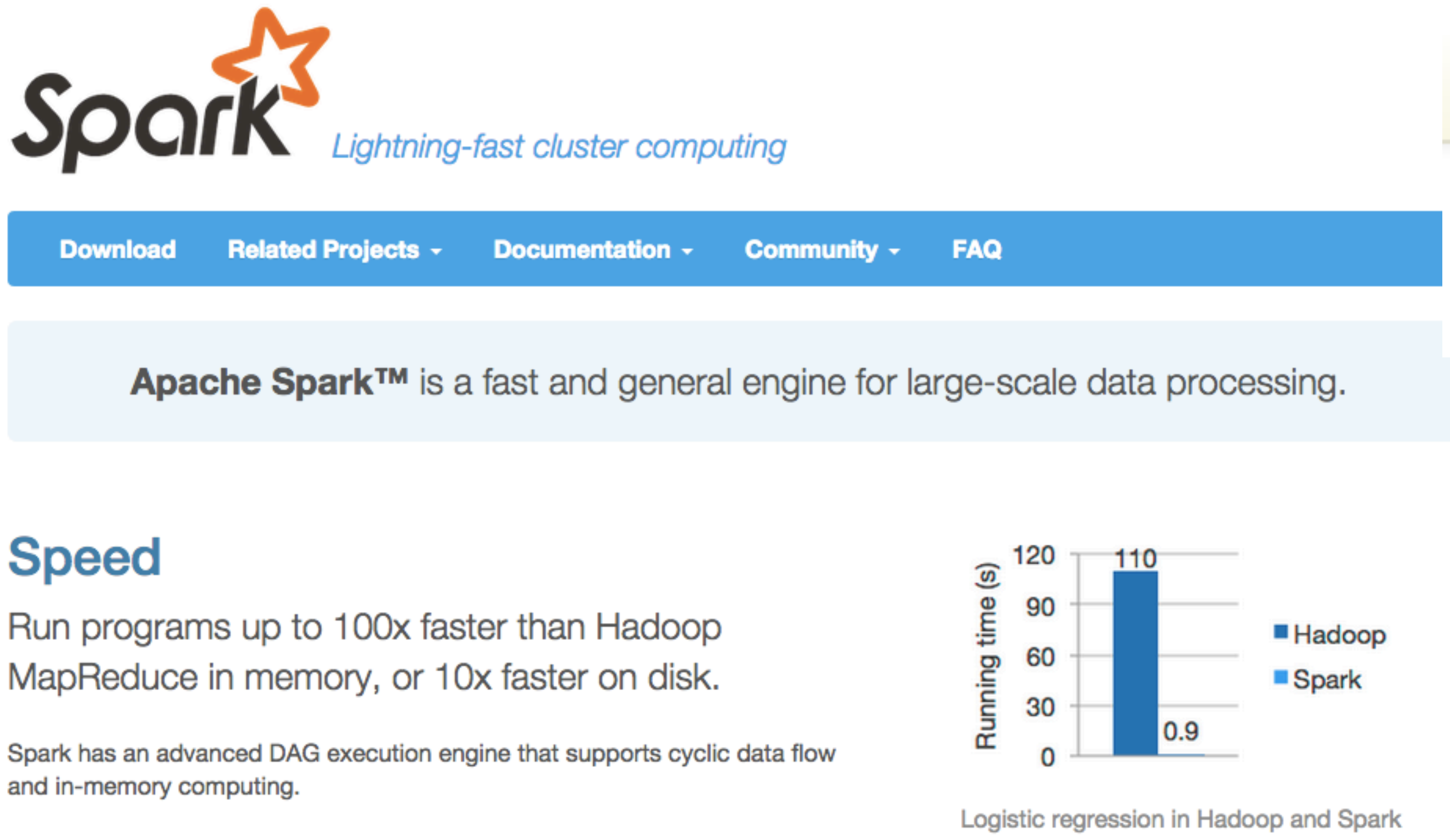
"Can you create a whitelist of eligible stores for a collaboration with Pinterest, stores not selling ammunition, adult material, cigarettes, etc.? It keeps timing out for me, but here is the SQL query that needs to be run."

- SELECT shops.domain FROM customers join shopify.products products on customers.shop_id = products.shop_id where (description not ilike '%ammunition%' and description not ilike '%cigarette%' and description not ilike '%bong%' and description not ilike '%ecstasy%' and description not ilike '%heroin%' and description not ilike '%opium%’ and description not ilike '%cocaine%' and description not ilike '%amphetamine%' and description not ilike '%mdma%' and description not ilike '%ghb%' and description not ilike '%ketamine%' and description not ilike '%pcp%' and description not ilike '%LSD%' and description not ilike '%steroid%' and description not ilike '%mescaline%' and description not ilike '%vaporizer%' and description not ilike '%hashish%' and description not ilike '%nicotine%' and description not ilike '%viagra%' and description not ilike '%cialis%' and description not ilike '%THC%' and description not

THEY SAID ITS JUST A SIMPLE SQL QUERY
imgflip.com

Distribute Code and Data

- Get data in distributed file system
- Use better-than-sql tools for analysis



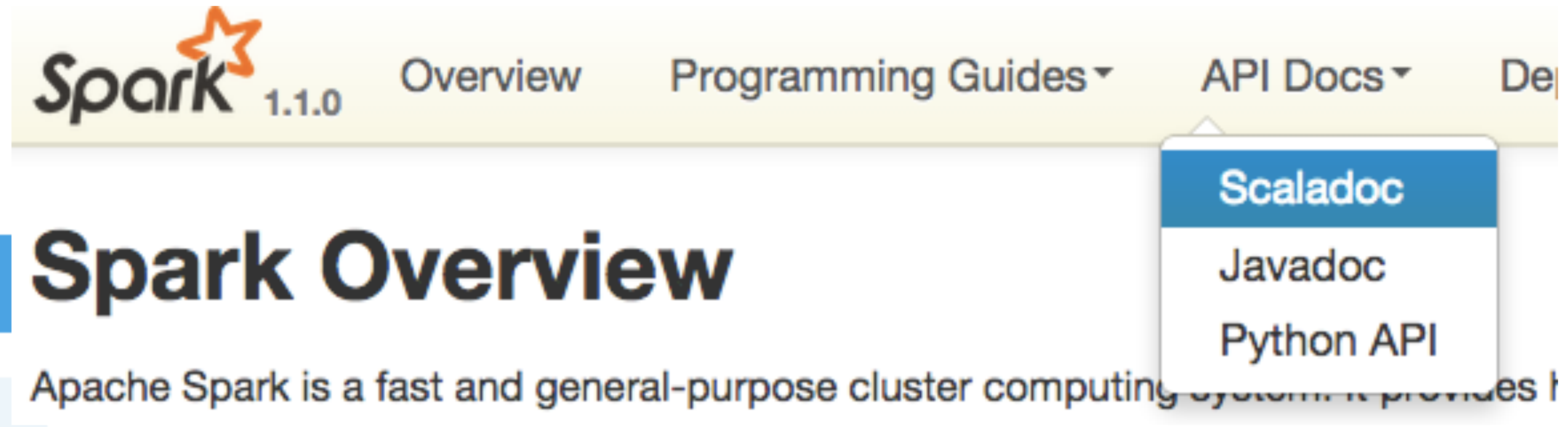
Apache Spark™ is a fast and general engine for large-scale data processing.

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.

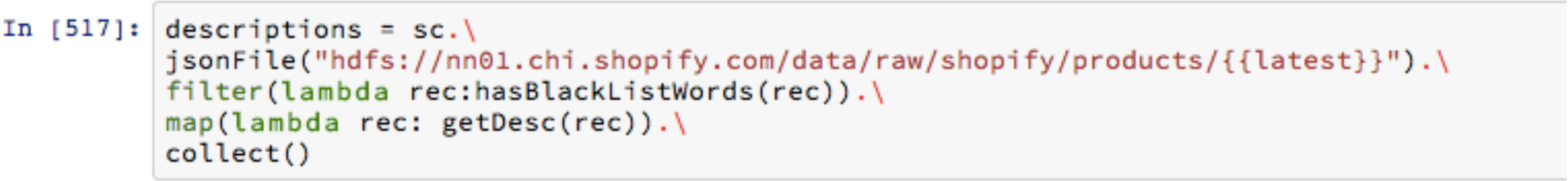
Logistic regression in Hadoop and Spark



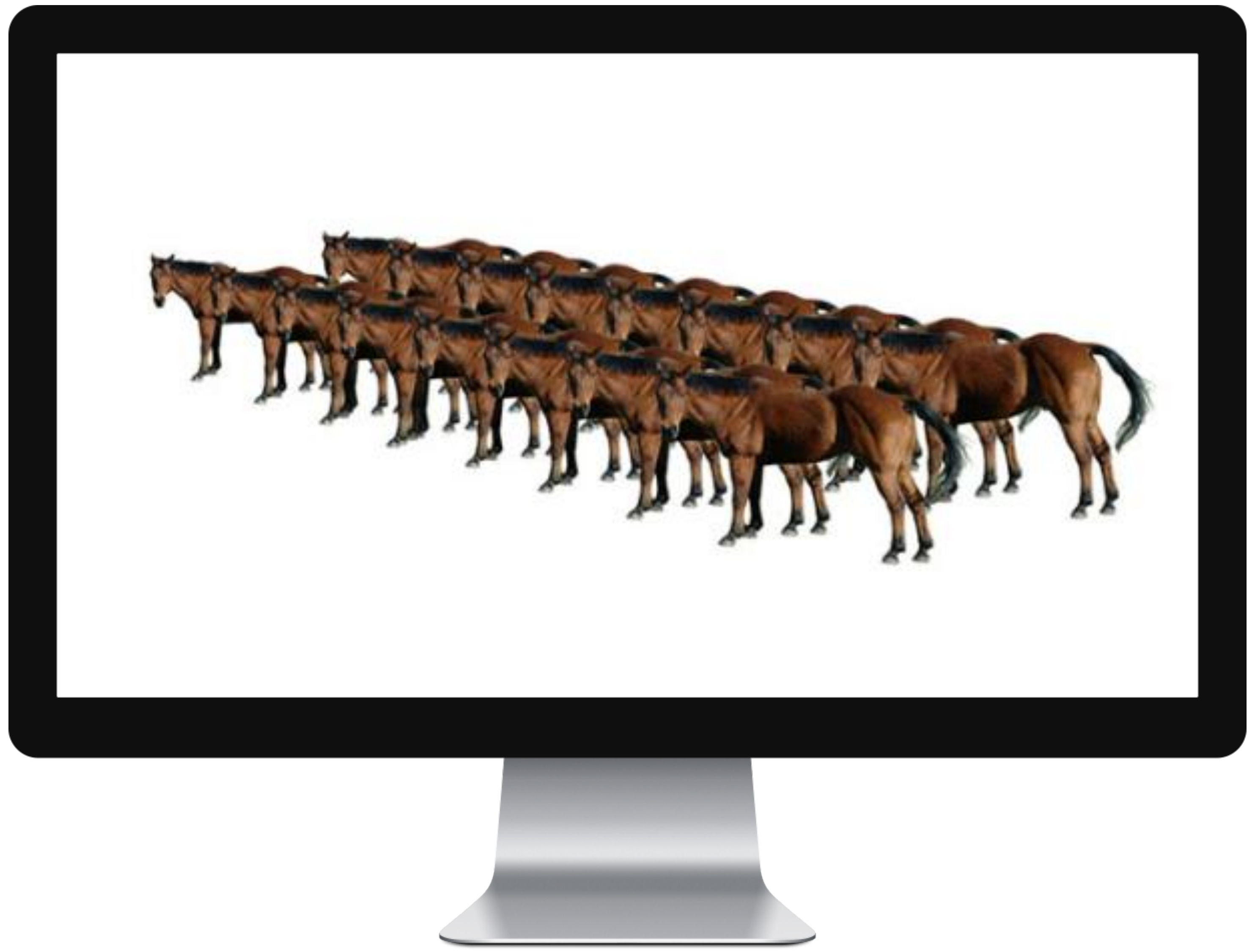
Spark Overview

Apache Spark is a fast and general-purpose cluster computing system. It provides h

Spark versus The World



```
In [517]: descriptions = sc.\
          jsonFile("hdfs://nn01.chi.shopify.com/data/raw/shopify/products/{{latest}}").\
          filter(lambda rec:hasBlackListWords(rec)).\
          map(lambda rec: getDesc(rec)).\
          collect()
```



Something was still missing

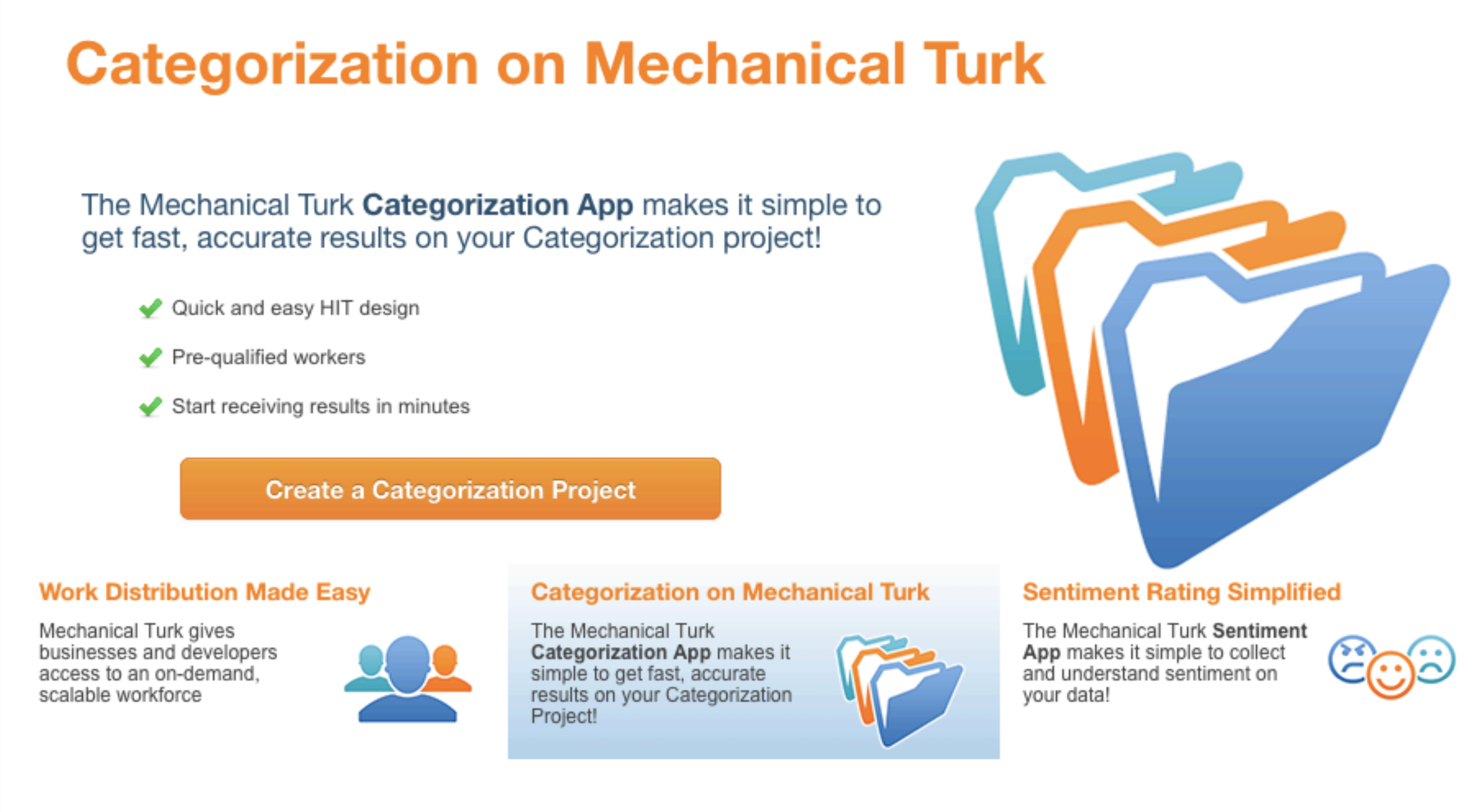
We had filtered around 30k merchants: too much!!

“The Buckshots, the world’s first ammunition for your thighs”

Mechanical Turk

The Amazon's Mechanical Turk (MTurk) is a crowdsourcing market place that enables individuals or businesses to co-ordinate the use of human intelligence to perform the tasks that computers are currently unable to do.

- classification
- sentiment analysis
- data cleaning



Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization project!

- ✓ Quick and easy HIT design
- ✓ Pre-qualified workers
- ✓ Start receiving results in minutes

Work Distribution Made Easy

Mechanical Turk gives businesses and developers access to an on-demand, scalable workforce

Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization Project!

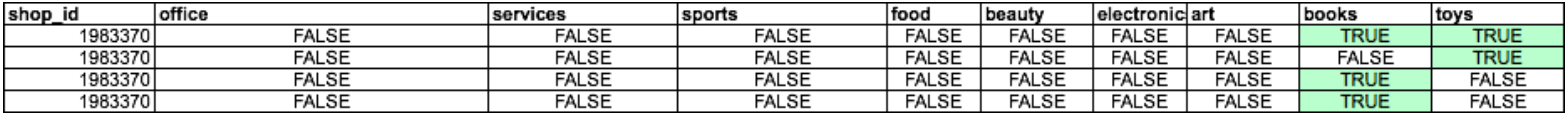
Sentiment Rating Simplified

The Mechanical Turk **Sentiment App** makes it simple to collect and understand sentiment on your data!

Lets try this!

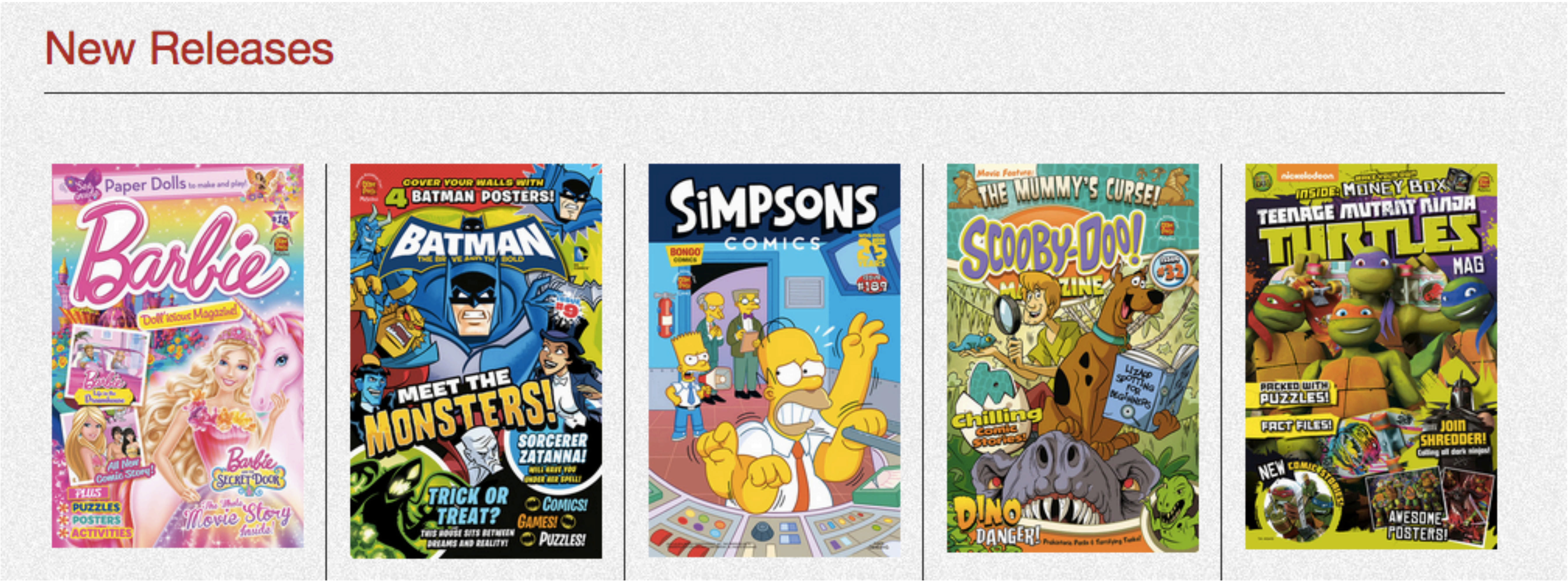
- Show the images of the 4 top selling products of each store to Turkers
- Allow for selection of multiple categories

Cleaning MTurk responses



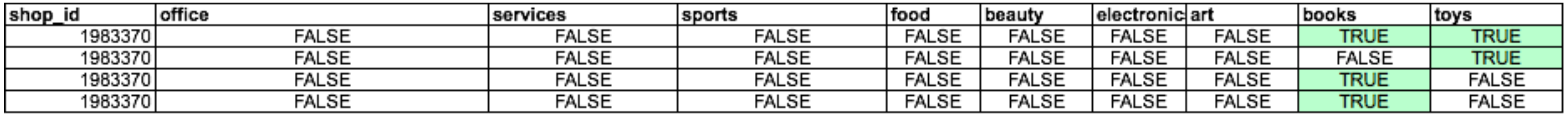
shop_id	office	services	sports	food	beauty	electronic	art	books	toys
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

- Otter Press http://www.otterpress.com.au/

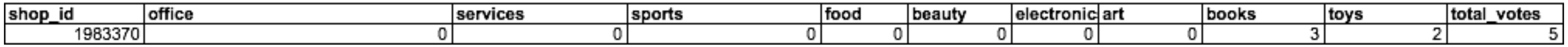


New Releases

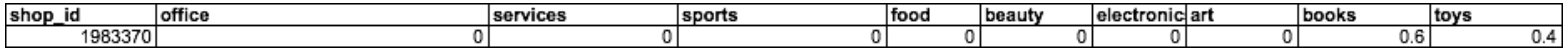
Summarizing responses



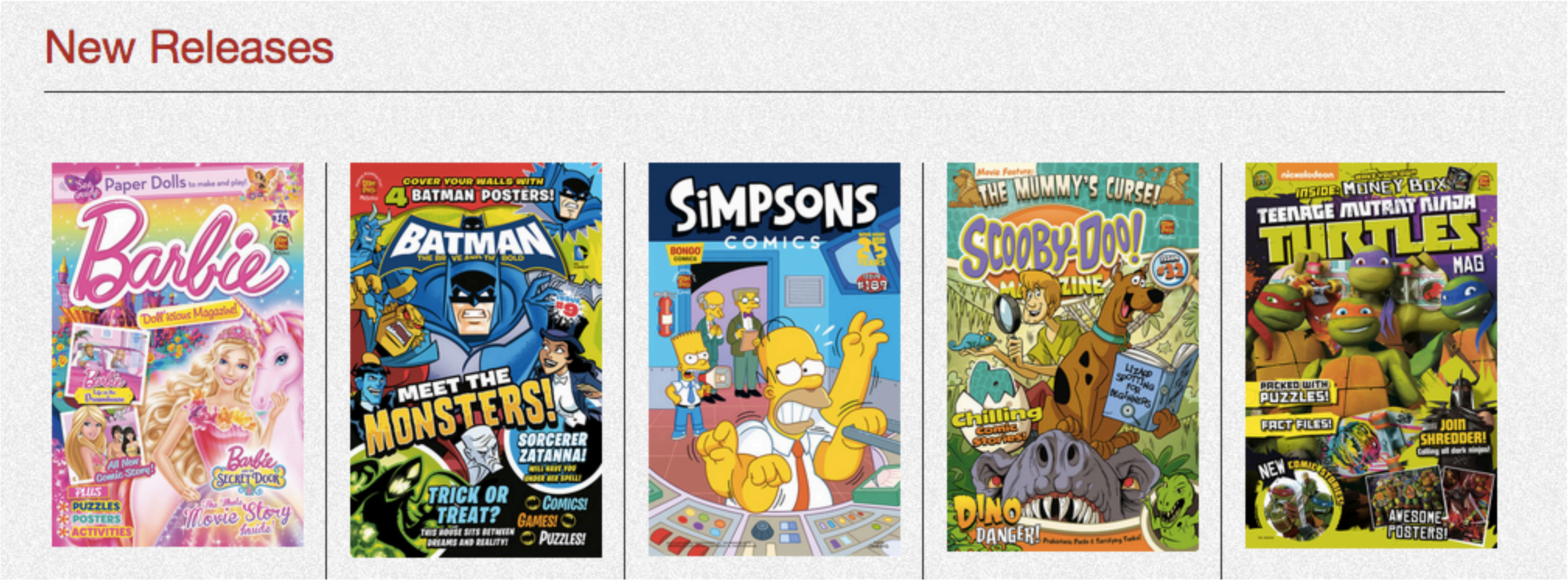
shop_id	office	services	sports	food	beauty	electronic	art	books	toys
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE



shop_id	office	services	sports	food	beauty	electronic	art	books	toys	total_votes
1983370	0	0	0	0	0	0	0	3	2	5

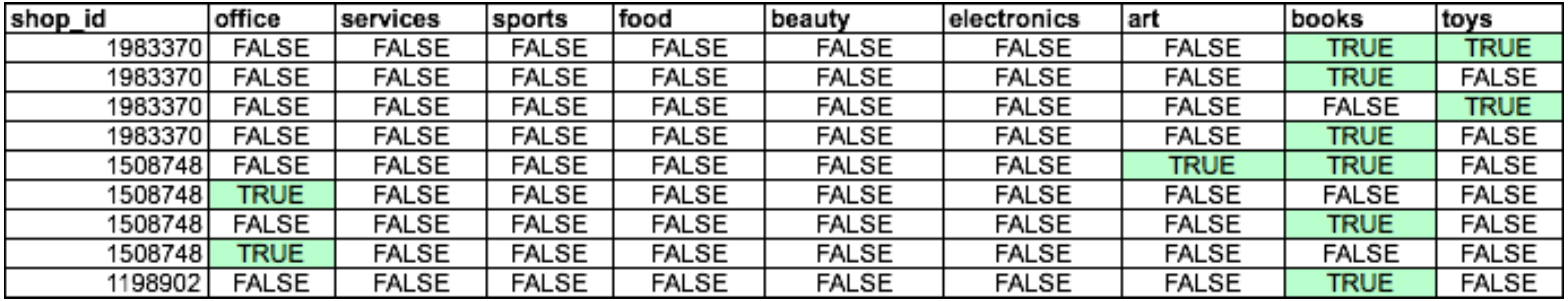


shop_id	office	services	sports	food	beauty	electronic	art	books	toys
1983370	0	0	0	0	0	0	0	0.6	0.4

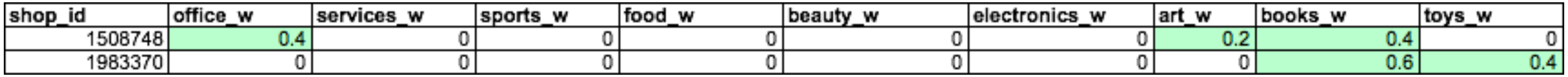


New Releases

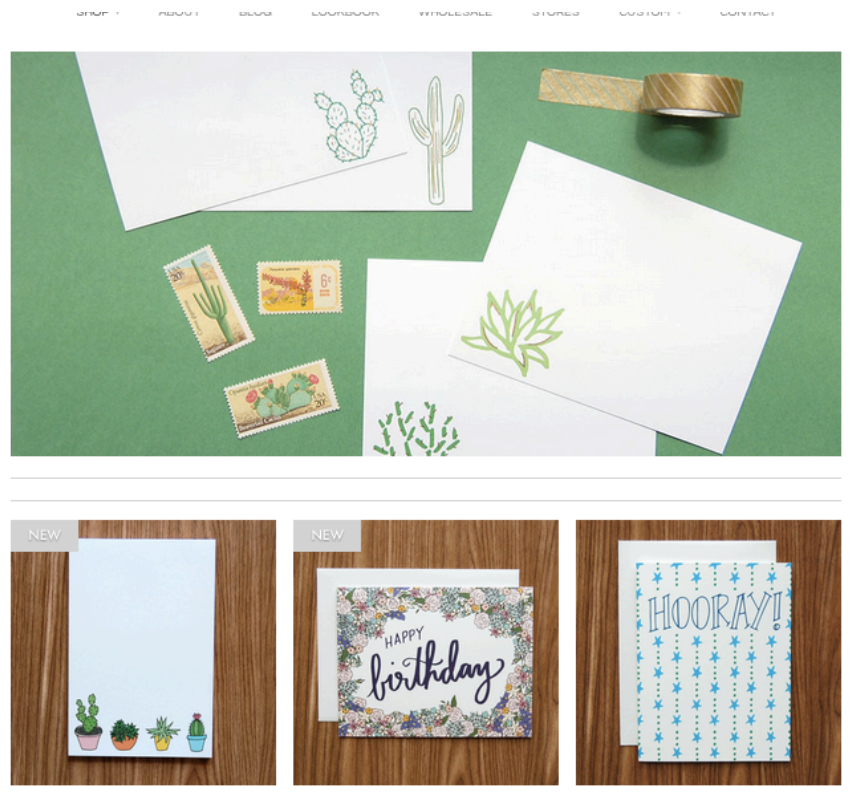
Summarizing responses

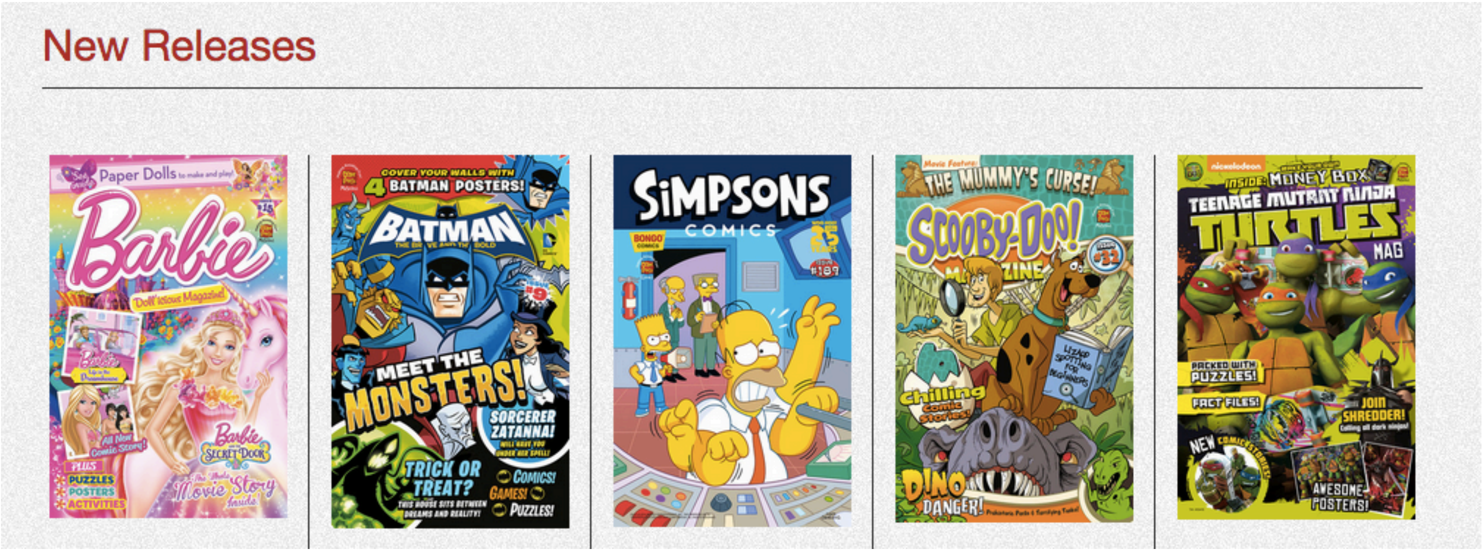


shop_id	office	services	sports	food	beauty	electronics	art	books	toys
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
1983370	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1508748	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
1508748	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1508748	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1508748	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1198902	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

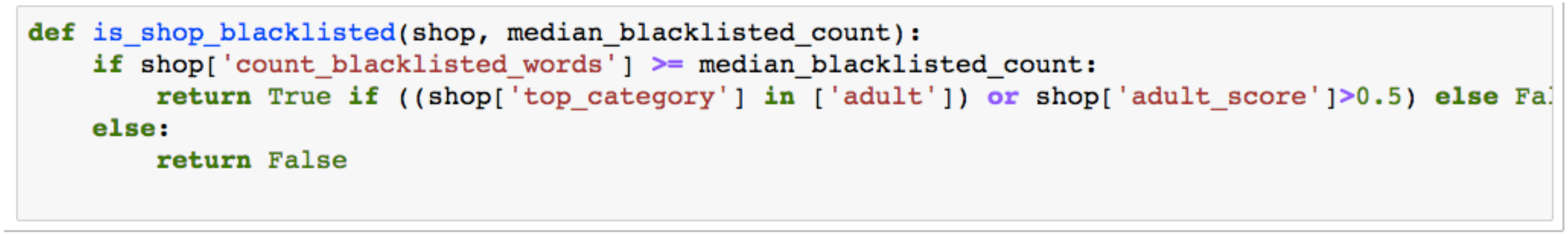


shop_id	office_w	services_w	sports_w	food_w	beauty_w	electronics_w	art_w	books_w	toys_w
1508748	0.4	0	0	0	0	0	0.2	0.4	0
1983370	0	0	0	0	0	0	0	0.6	0.4



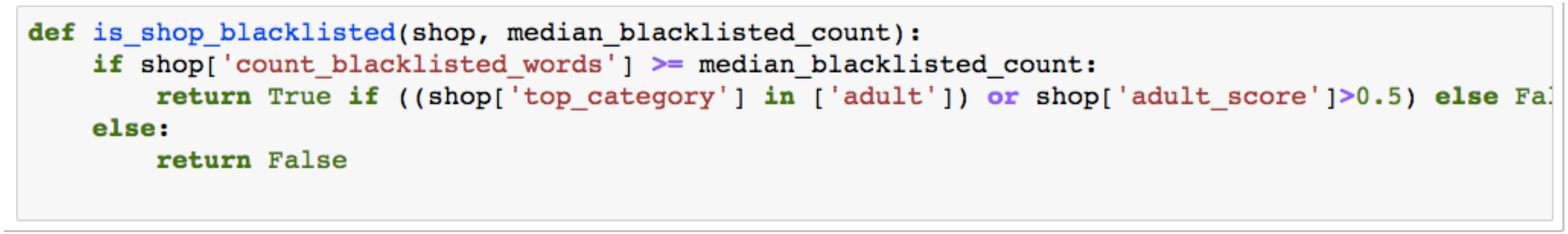


From 30k to 10k shops



```
def is_shop_blacklisted(shop, median_blacklisted_count):
    if shop['count_blacklisted_words'] >= median_blacklisted_count:
        return True if ((shop['top_category'] in ['adult']) or shop['adult_score']>0.5) else Fa
    else:
        return False
```

From 30k to 10k shops



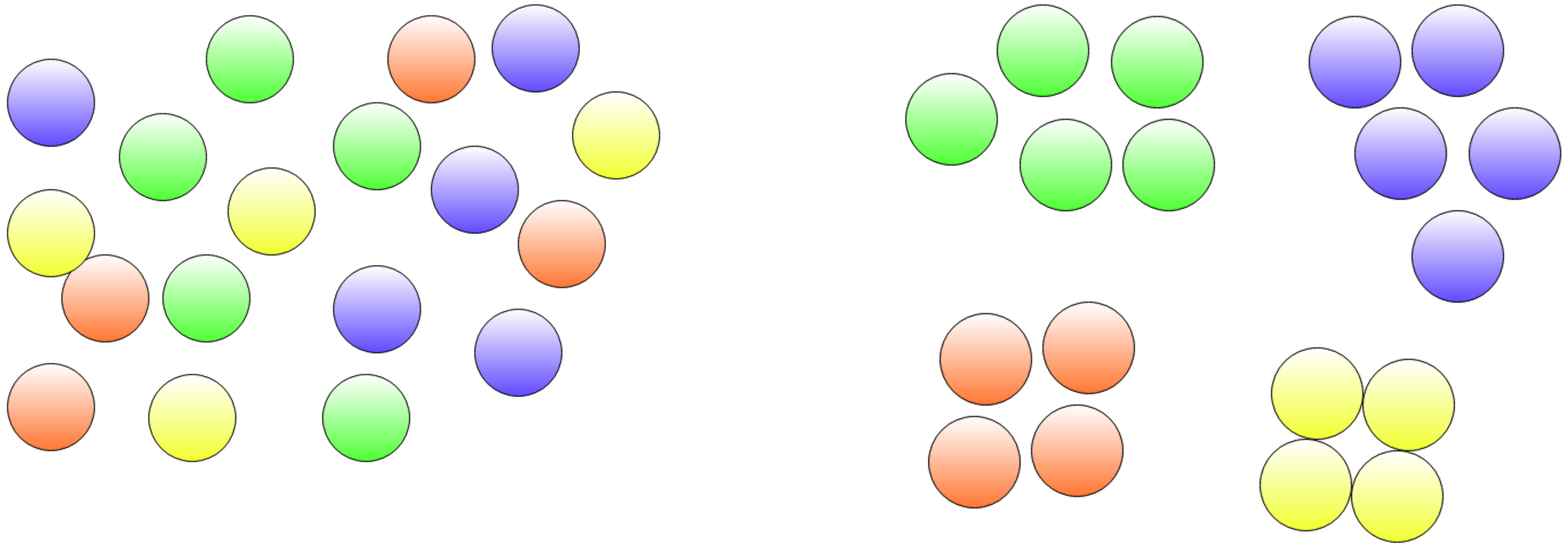
```python
def is_shop_blacklisted(shop, median_blacklisted_count):
    if shop['count_blacklisted_words'] >= median_blacklisted_count:
        return True if ((shop['top_category'] in ['adult']) or shop['adult_score']>0.5) else Fa
    else:
        return False
```

Pinterest Rich Pins Now Automatically Enabled for Shopify Merchants

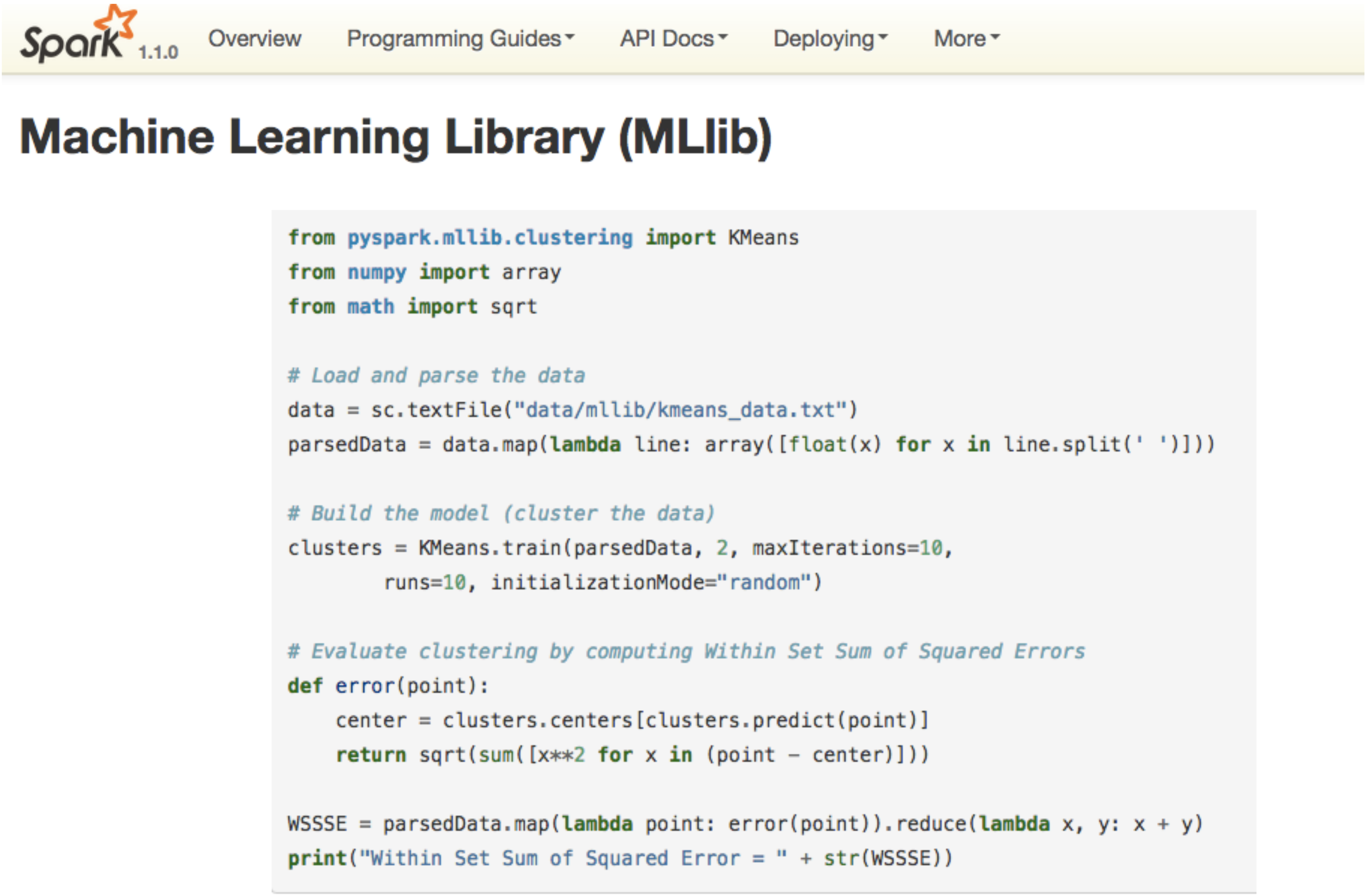
by Dayna Winter | Posted in Shopify Updates | July 17, 2014

Finding “Similar” stores

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some way or another) to each other than to those in other groups (clusters).

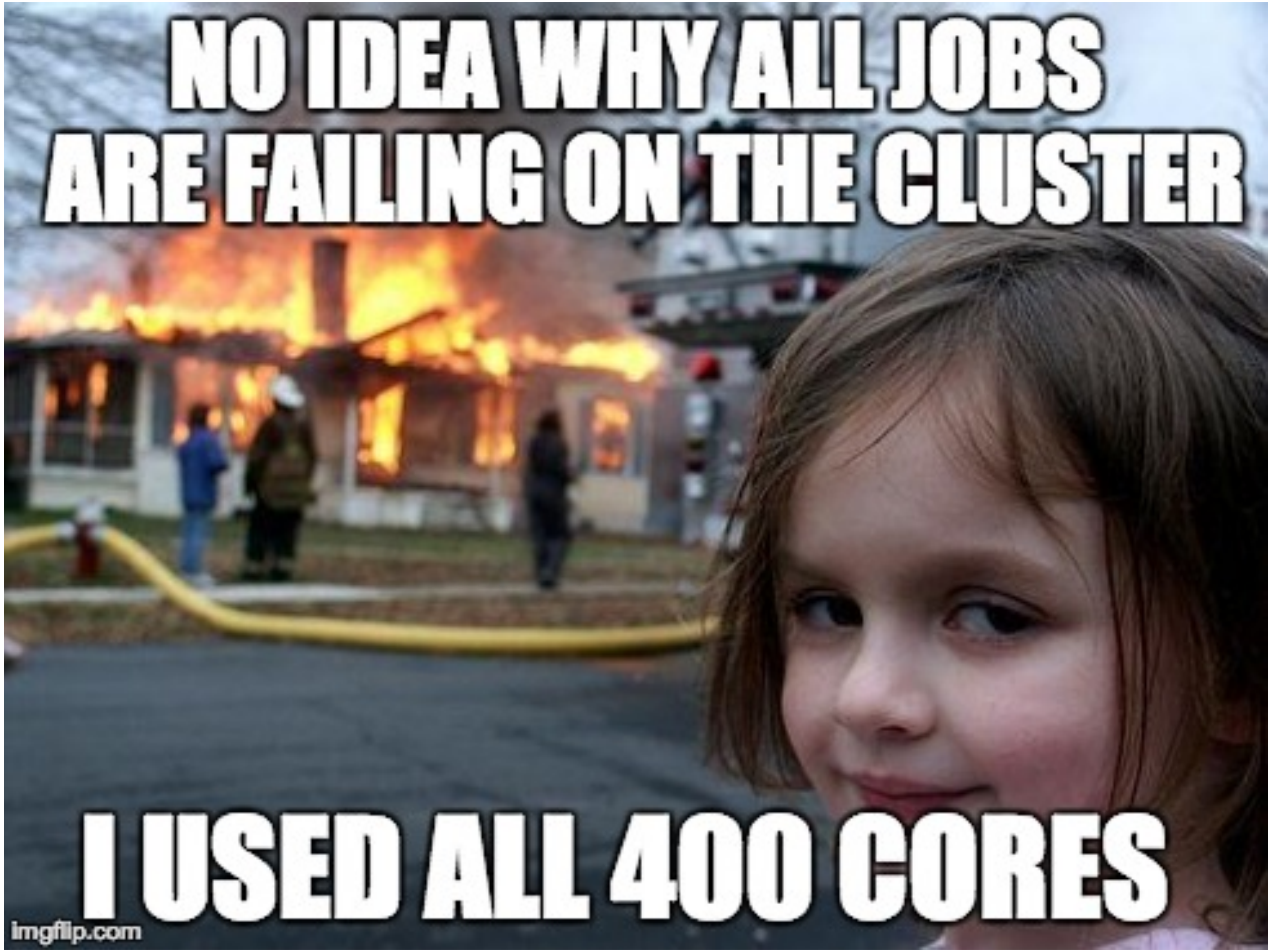


Clustering with Spark Mllib

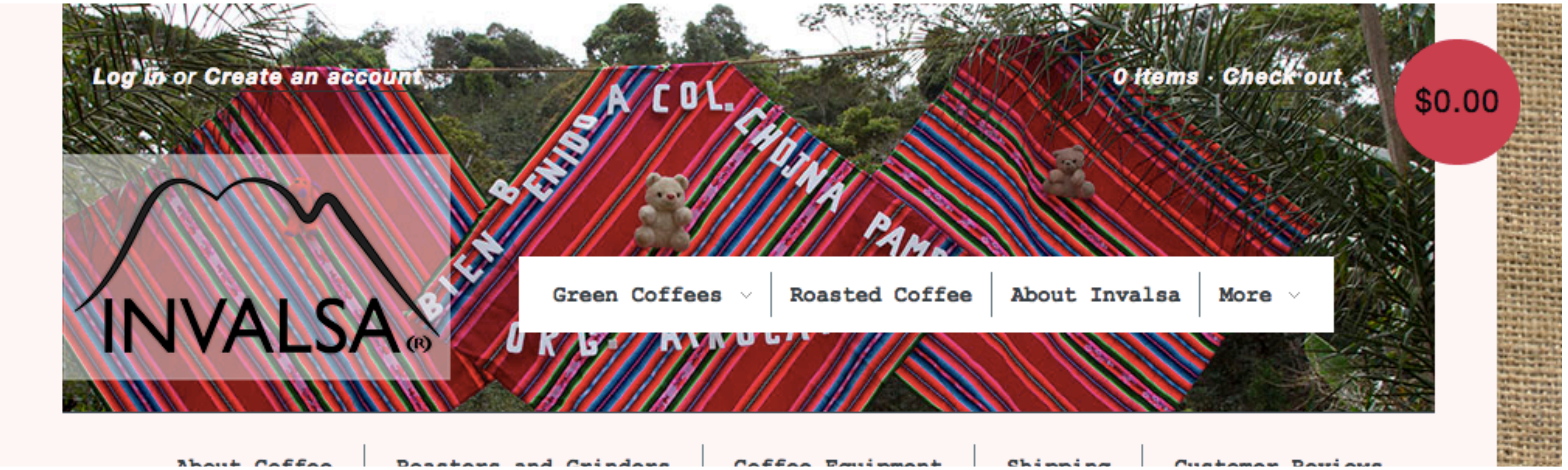


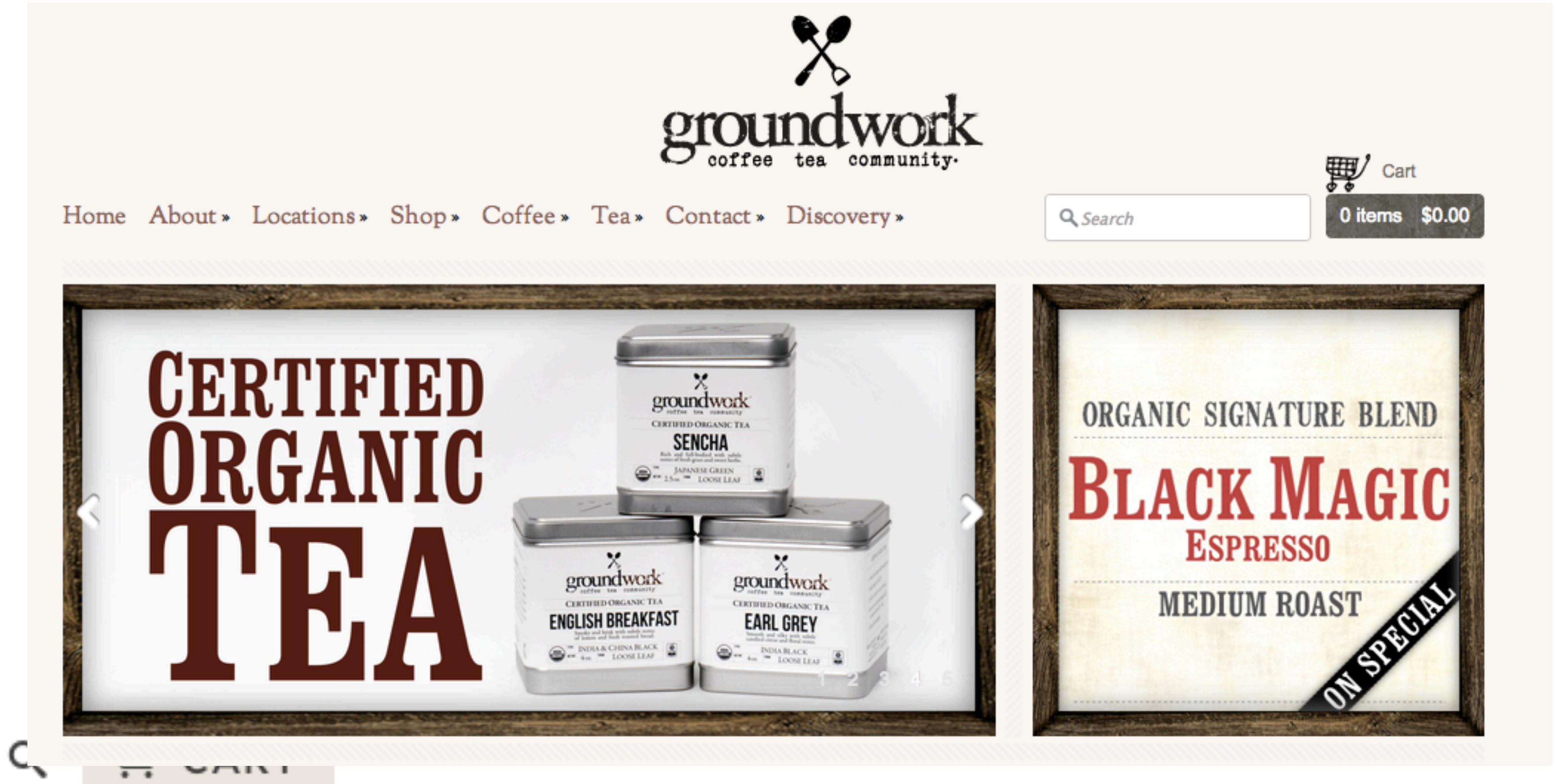
Machine Learning Library (MLlib)

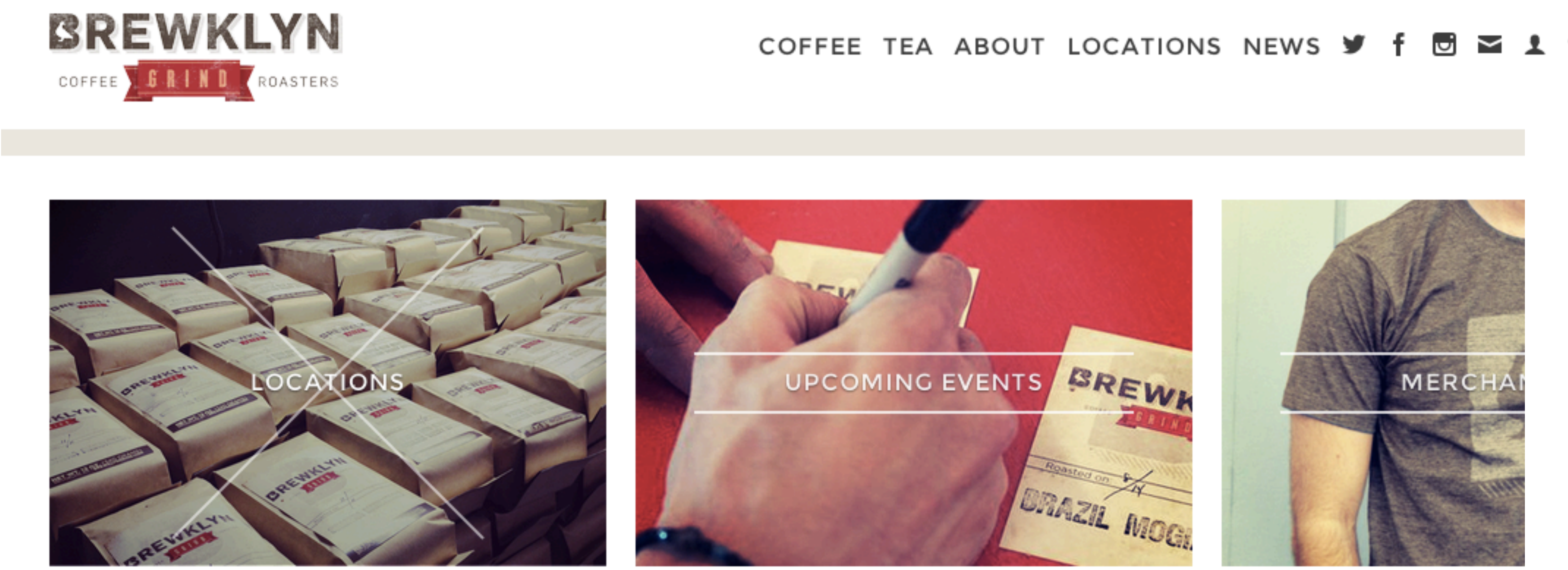
```python
from pyspark.mllib.clustering import KMeans
from numpy import array
from math import sqrt

# Load and parse the data
data = sc.textFile("data/mllib/kmeans_data.txt")
parsedData = data.map(lambda line: array([float(x) for x in line.split(' ')]))

# Build the model (cluster the data)
clusters = KMeans.train(parsedData, 2, maxIterations=10,
        runs=10, initializationMode="random")

# Evaluate clustering by computing Within Set Sum of Squared Errors
def error(point):
    center = clusters.centers[clusters.predict(point)]
    return sqrt(sum([x**2 for x in (point - center)]))

WSSSE = parsedData.map(lambda point: error(point)).reduce(lambda x, y: x + y)
print("Within Set Sum of Squared Error = " + str(WSSSE))
```

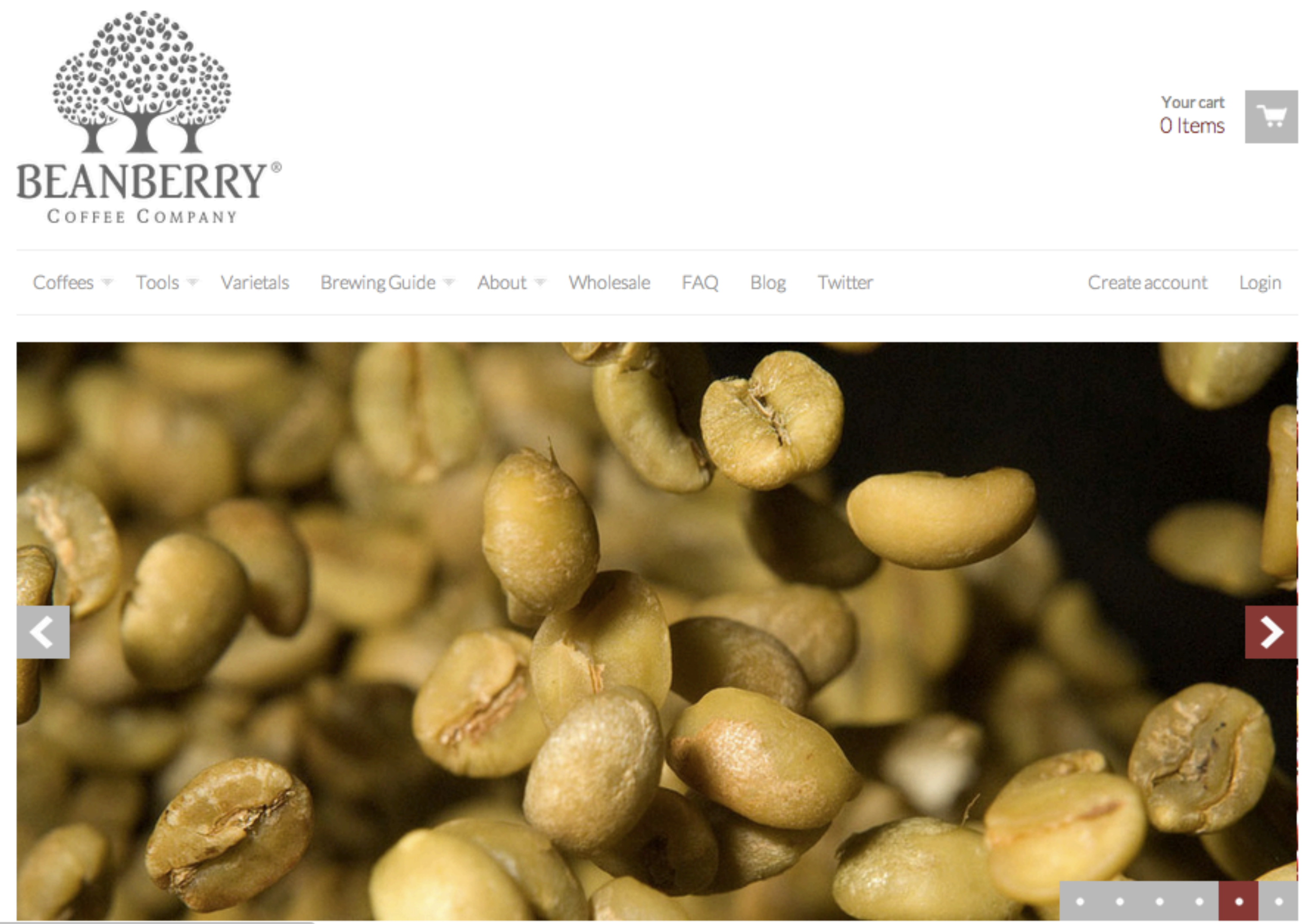

NO IDEA WHY ALL JOBS ARE FAILING ON THE CLUSTER
I USED ALL 400 CORES
imgflip.com

Some cool clusters

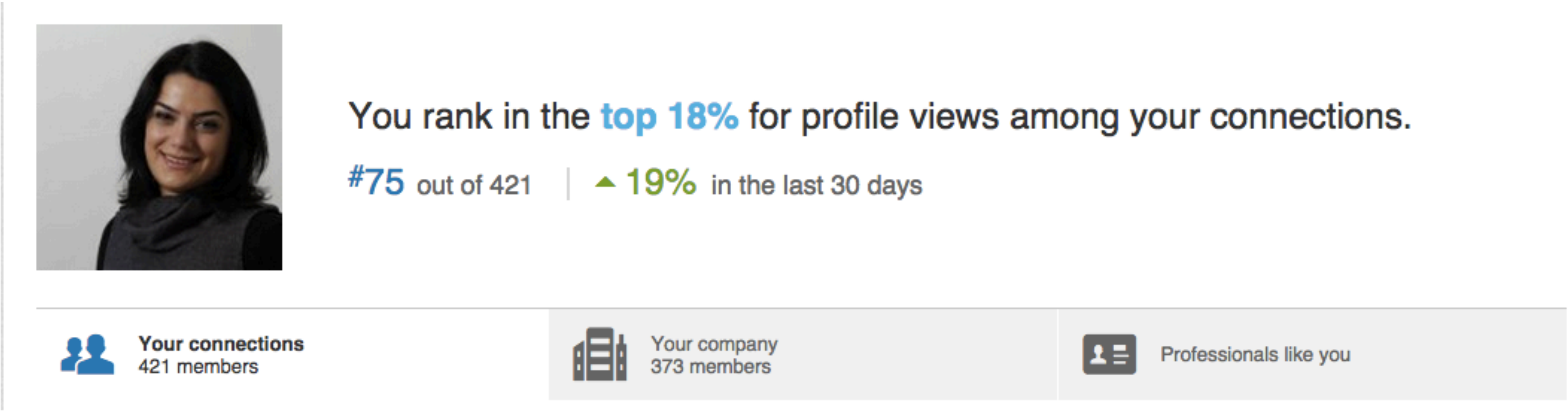








Comparison with Peers



You rank in the **top 18%** for profile views among your connections.

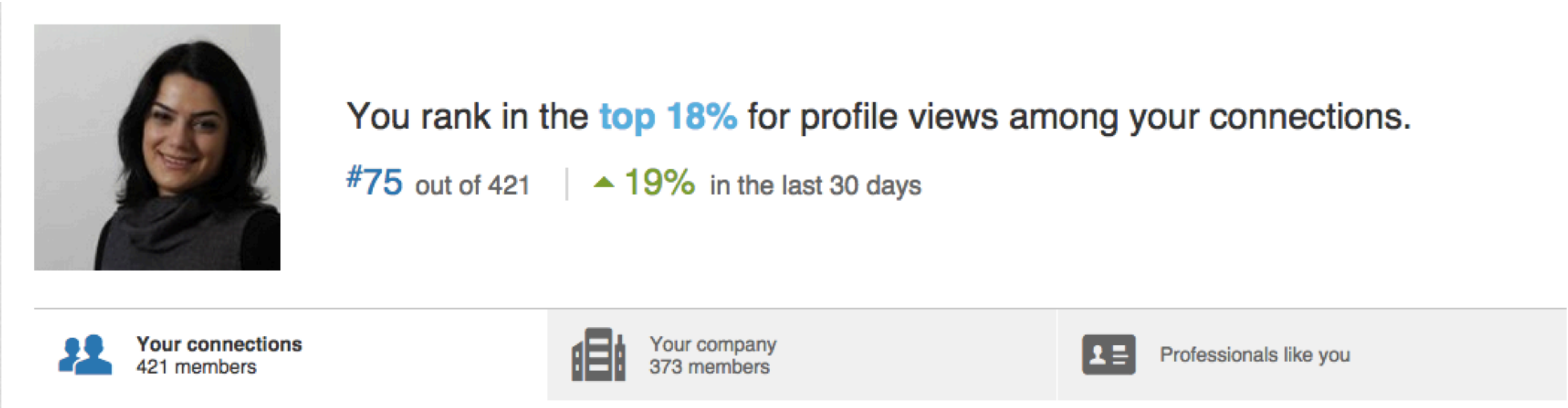
#75 out of 421 | ▲ 19% in the last 30 days

Your connections
421 members

Your company
373 members

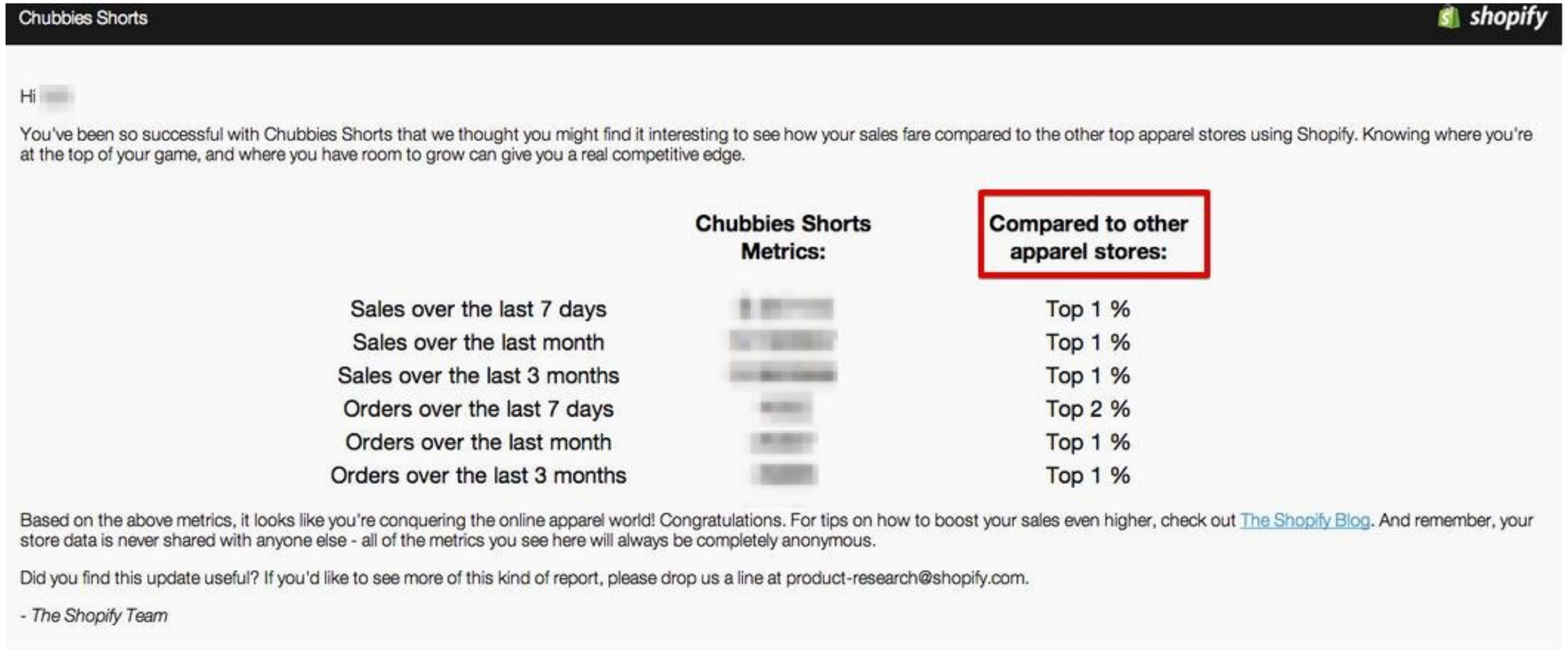
Professionals like you

Comparison with Peers



You rank in the **top 18%** for profile views among your connections.

#75 out of 421 | ▲ 19% in the last 30 days



Hi [redacted]

You've been so successful with Chubbies Shorts that we thought you might find it interesting to see how your sales fare compared to the other top apparel stores using Shopify. Knowing where you're at the top of your game, and where you have room to grow can give you a real competitive edge.

	Chubbies Shorts Metrics:	Compared to other apparel stores:
Sales over the last 7 days	[redacted]	Top 1 %
Sales over the last month	[redacted]	Top 1 %
Sales over the last 3 months	[redacted]	Top 1 %
Orders over the last 7 days	[redacted]	Top 2 %
Orders over the last month	[redacted]	Top 1 %
Orders over the last 3 months	[redacted]	Top 1 %

Based on the above metrics, it looks like you're conquering the online apparel world! Congratulations. For tips on how to boost your sales even higher, check out The Shopify Blog. And remember, your store data is never shared with anyone else - all of the metrics you see here will always be completely anonymous.

Did you find this update useful? If you'd like to see more of this kind of report, please drop us a line at product-research@shopify.com.

- The Shopify Team

Finally bought some shoes

Lessons Learned

- Need data: use mechanical Turks
- Need processing power: Spark is easy to get started
- Happy Hacking!

Questions?

Questions?

Thank you for listening

Feel free to ping me @solmaz_sh