

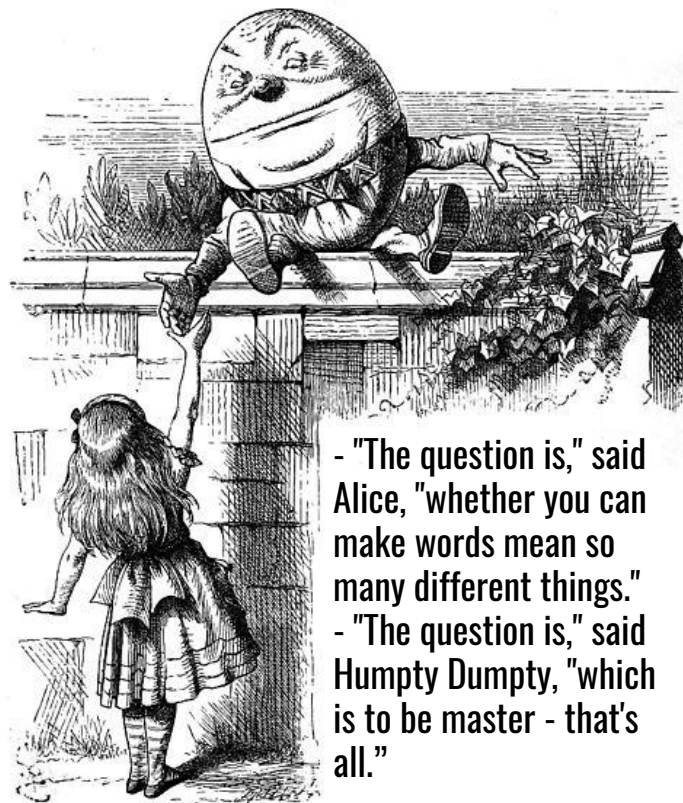
“Unnatural” Language processing

How NLP helps us map our product
catalog for recommendations,
search & product development

[Yizhar \(Izzy\) Toren](#)

Data Scientist @  **shopify**  **Oberlo**

WhyR 2018 Wrocław, 04.07.2018



- "The question is," said Alice, "whether you can make words mean so many different things."

- "The question is," said Humpty Dumpty, "which is to be master - that's all."

(Lewis Carroll)

Our catalog



Suppliers



Merchants



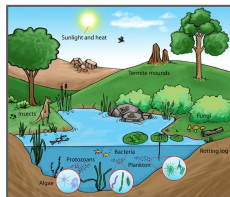
Merchants

Motivation

- Deliver “**good**” product recommendations to our merchants in product pages, search, dashboards, etc.
- Improve existing categories & identify gaps in the catalog
- Make this scalable & automatic

So... why not just recommend “best sellers”?

- **Cannibalisation:** Pushing merchants to a saturated markets can disrupt existing businesses (~~collab. filtering~~)
- **Privacy:** Recommending niche products based on sales is a sensitive topic
- **Healthy ecosystem:** We want to expose more of our catalog, not just a few “rockstars”



Let's focus on what each merchant does well (sell) and their interests (search, select, etc.)

Motivation (again)

- Deliver ~~“good”~~ **similar*** product recommendations to our merchants in product pages, etc.
- Improve existing categories & identify gaps in the catalog
- Make this scalable & automatic

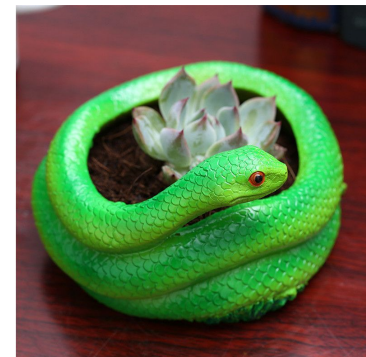
** Assumption: merchants that are able to sell / interested in a product will sell / show interest in “similar” products.*

Sources of information on our products

- Suppliers: category, product description, image upload
- Merchants: product description customisation, choice of images, search queries, etc.
- Performance Metrics: imports/orders/disputes/... as an indication of quality. Sensitive!

Wait, did you just say “we have images”?

Yes. We have multiple images for every product, so let's play...



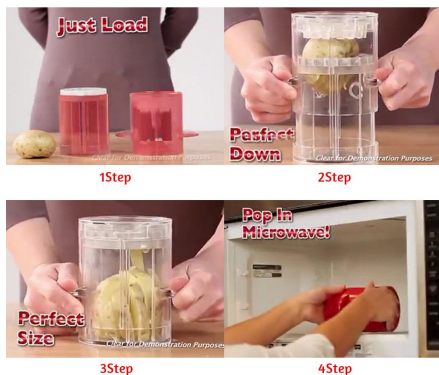
WHAT'S IN THIS IMAGE?

Wait, did you just say “we have images”?

Yes. We have multiple images for every product, so let's play...



Room Aromarizer
Not a mystery box



French fries slicer
Can you even see
the potato?



Phone cover
Not woman /
phone / blouse



Plant pot
Not snake /
cactus

I wish “someone” was trying to tell me what's in the picture...

Good news: Our suppliers are trying to!

Tags, Keywords & metadata

Curated closed list,
Indexed & organised

Keyword	Class
Phone cover	category
iPhone	brand
silicone	material
green	color
140x60mm	dimensions
smooth	texture

The titles we actually get

Not a sentence (context based NLP does
not work well), but also no uniformity

ID	Class
1234	Green silicon(e) iphone cover, 140x60mm smooth with good grip best quality 2018 collection excellent price
1235	...

Natural Language

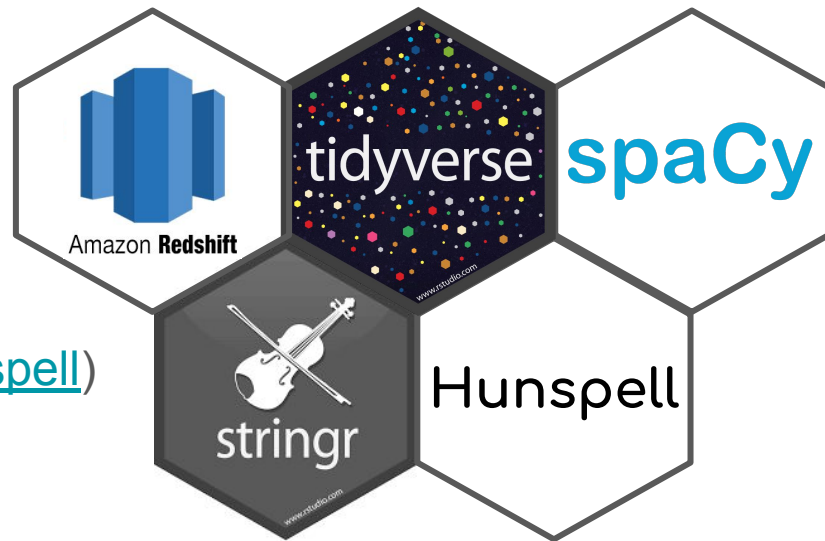
Context & depth to correct &
tag POS, extract topics...

This wonderful product!
It comes **in a beautiful**
green color. **The**
smooth silicon(e) surface
guarantees perfect grip.
This is the best protection
for your precious iPhone
from iCovers inc. 2018
collection!

Tags: Stopwords / Spelling errors / Spelling variants / Marketing phrases / Brand names / Nouns / Verbs / Adj / Adv / punctuation ...

Toolkit

- Data source: Product information from our DWH
(via Redshift JDBC)
- Scripting: R/tidyverse
- Spell correction: Hunspell
(<http://hunspell.github.io/>)
via R package
(<https://cran.r-project.org/package=hunspell>)
- POS tagging: spaCy (<https://spacy.io/>)
via R package
(<https://cran.r-project.org/package=spacyr>)



How standard NLP workflow (spaCy) went wrong

Tokenize

Spelling

Stop words

Lemmatization

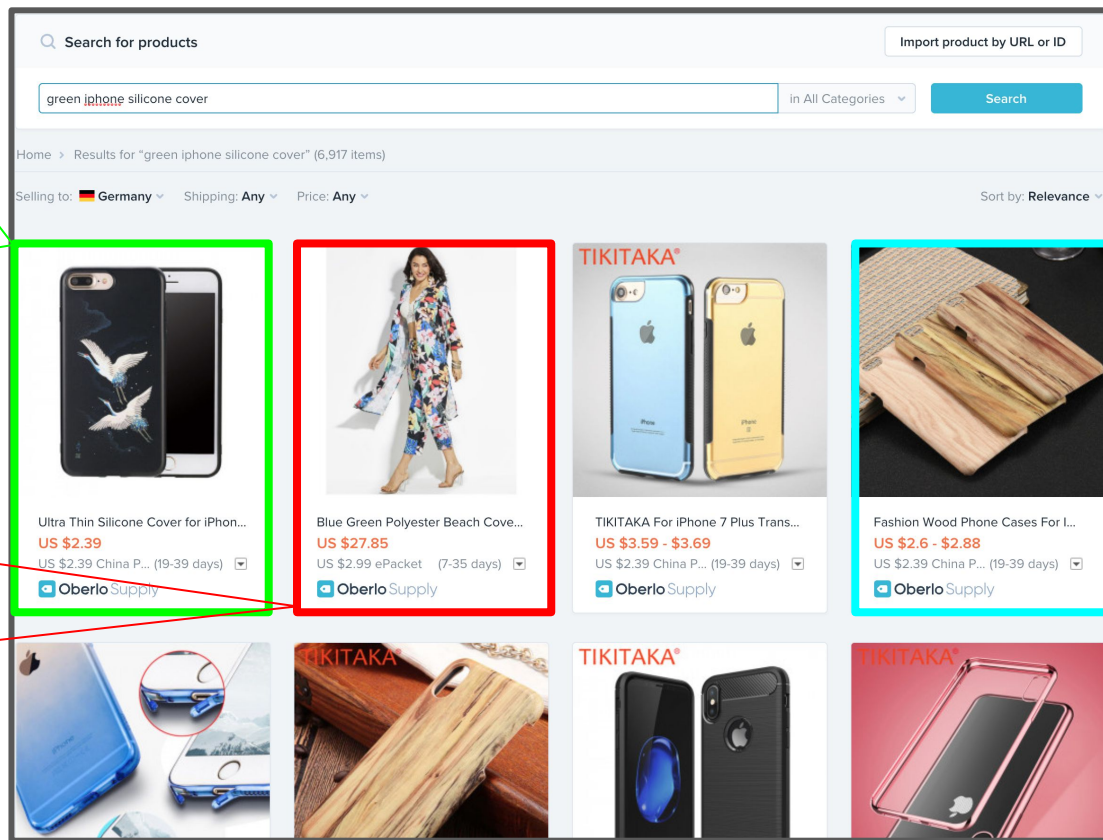
PoS Tagging

- Over-Tokenization: “19mm” \Rightarrow [19, mm] : [NUM, PROPN]
- Wrong lemmas: “sleeved” \Rightarrow “sleev” but “sleeve” \Rightarrow “sleeve”
- PoS tag inconsistency: “nine” is NOUN, NUM and stopword(?)

Our solution

- “Dumb” tokenization: we manually strip common punctuations and then split by whitespace (better safe than sorry)
- Single word NLP for lemmatization and tagging (for consistency)

How Elasticsearch went wrong



iPhone ✓
cover ✓
silicone ✓
green X

iPhone X
cover ✓
silicone X
green ✓

iPhone ✓
cover ✓
silicone X
green X

Let's apply some “world knowledge”

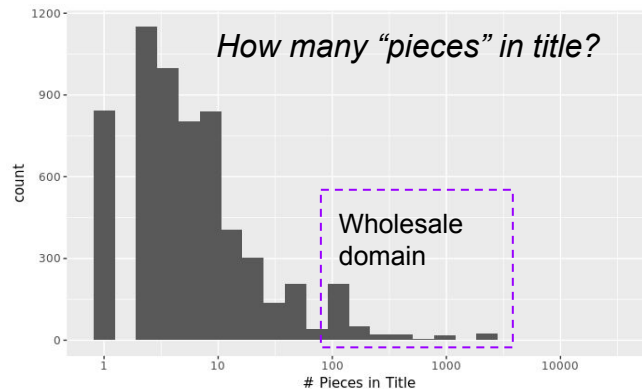


- Expressions:

- Remove marketing phrase:
“Drop shipping”, “Best price”, ...
- Extract metadata

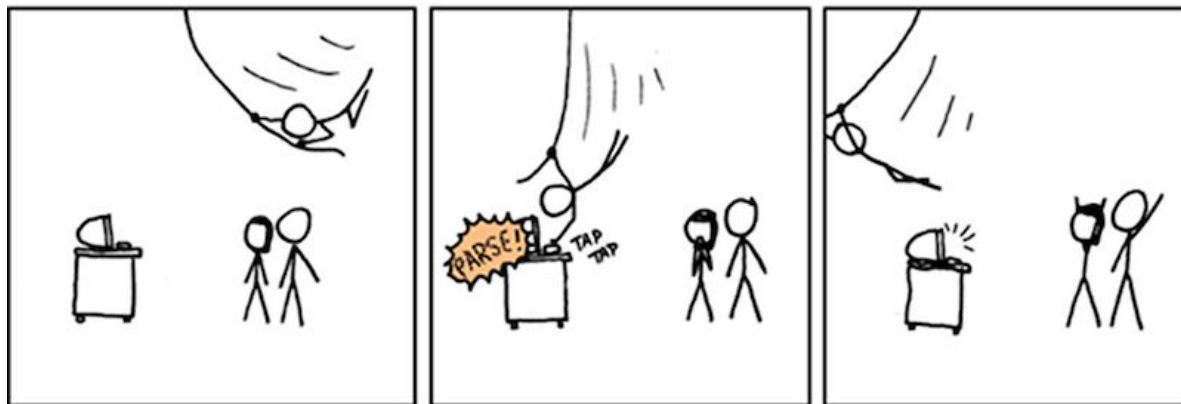
- Words:

- Industry specific: e.g. “top” is not a stopword
- Lexical: colors, materials, brands, device models, ...
- Structural: hashtags, capacity, volume, size, ...
- Synonyms / Antonyms
- ... (a long list of long lists)



And for the rest...

- We assume: *nouns* = object / *adjectives* = properties.
- Not having natural language in title means no context for PoS tagging models - They try very hard but are often inconsistent.
- We can make some simplifying assumptions (e.g. past tense verbs like “coloured” are adjectives, etc.) but not much more...



Grammar to the rescue!

Example

“BRAND-B Top quality Ellegant Blue mans shirt 2pack”



Example



“BRAND-B Top quality Ellegant Blue mans shirt 2pack”



Brand list	Keep?
Brand A	✓
Brand B	✗
...	

Phrase list
([Gg]reat [Tb]est)\s*[Qq]uality
[Ff]ree\s*[Sh]ipping\W
...

Regex list
([0-9]+)(\s*)([Pp]ack)
...

#	Match
1	2
2	<NA>
3	pack

Spell ✓

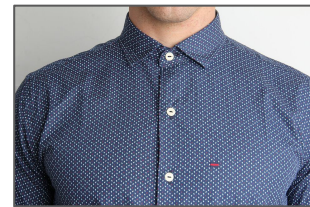
Manually	
From	To
Mans	men
Men's	men
...	

Spell ✗

Suggested
Elegant
Elephant
Elegance

Rules?

Example



“BRAND-B Top quality Ellegant Blue mans shirt 2pack”



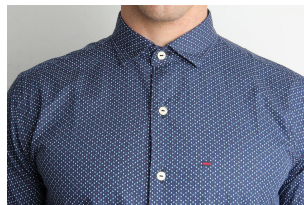
“Elegant Blue man shirt”



keyword	category
elegant	adj
shirt	noun
man	gender
blue	color



Example



“BRAND-B Top quality Ellegant Blue mans shirt 2pack”



“Elegant Blue men shirt”



keyword	category
elegant	adj
shirt	noun
man	gender
blue	color



keyword	cat
casual	adj
t-shirt	noun
man	gender
blue	color



keyword	cat
elegant	adj
shirt	noun
woman	gender
blue	color

Things we discovered:

- We can safely remove stop words, **except when we can't** (e.g. “top”)
- Spelling mistakes are useful!
 - No suggestion: **try to extract metadata** (model number, measurements, ...)
 - With suggestion: **curate brand names**
- We **can't** remove/fix spelling automatically because of new terms (e.g. “splitter”)
- **When in doubt, avoid stemming**

Has spelling suggestion

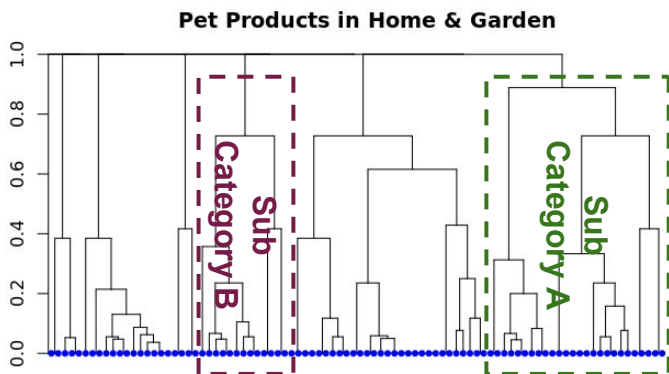
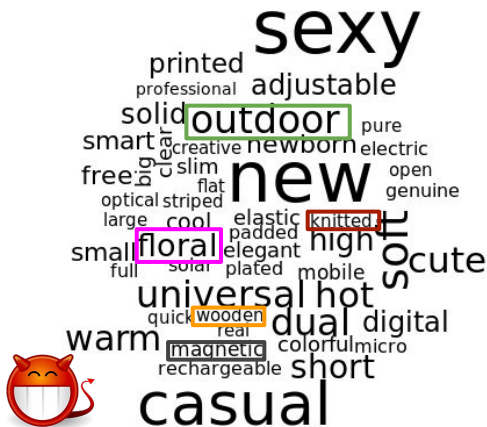


No spelling suggestion

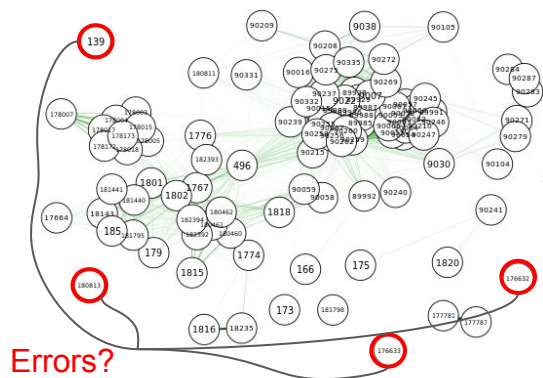


Implementation

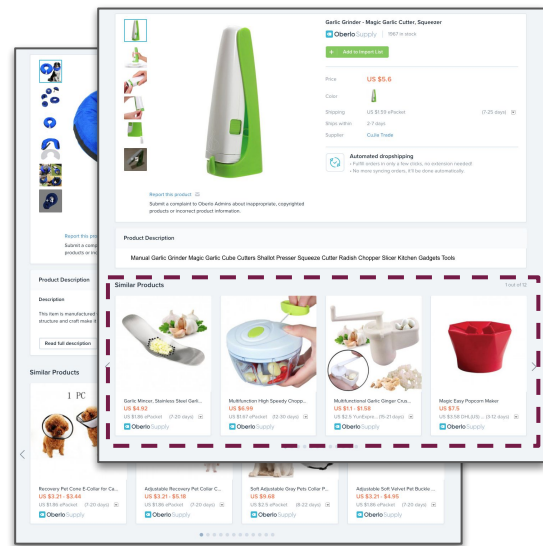
- Browsing: Similar products (local), Sub-sub-... categories, Catalog maintenance, etc.
- Personalized \ context based recommendations
- Feed this to ElasticSearch
- Ideas for product collections, etc...



Pet Products in Home & Garden




Errors?



In Conclusion

- Data provenance is critical, but always validate your assumptions **by looking at your data!**
- “Out of the box” tools are awesome, but their default models are often trained on very “specific” datasets. Tune or replace the components that don’t work for you,
& use errors / issues you discover more about your data.
- When presented with the right type of aggregated data product (list, node-graph, word-cloud 🐱), a human-in-the-loop can solve at a glance a lot of problems that are very hard to solve algorithmically.
Use your colleagues!

So.... Why ?

- Because it's an awesome tool to **look at your data!**
- Because it's easy to integrate different data sources and frameworks super quickly (thanks CRAN) and make them play nicely together (DF as a first class citizen)
- Because you can do things in a tidy way, so it's an awesome tool for prototyping
- Because notebooks are awesome (We  reproducibility)



Oberlo



shopify

We're Hiring!

https://jobs.lever.co/shopify?lever-via=_eJyY6KbYT