CHAIN OF DEMAND

Lisa Kailai Han, Anwen Huang, Lexie Li, Joyce Moon, Dasson Tan

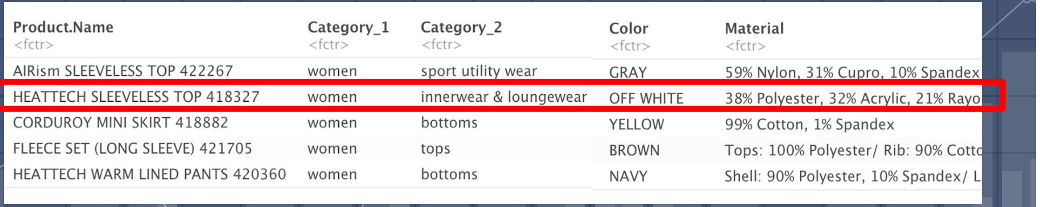
Department of Statistics & Data Science
Carnegie Mellon University

OBJECTIVE

Identify features that characterize an apparel’s online popularity and explore clustering relationships among items.

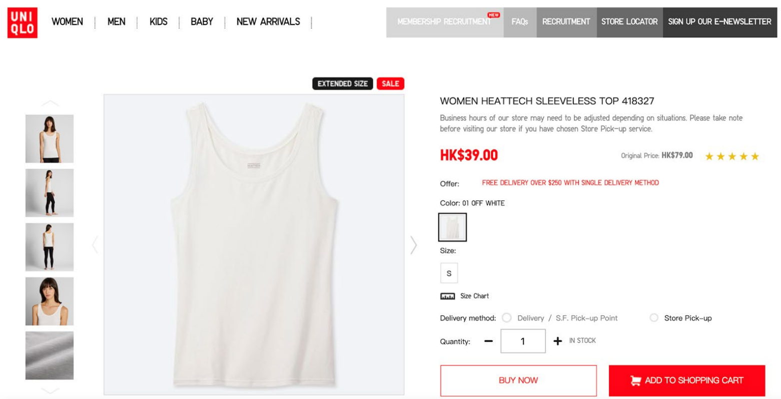
THE DATA SET

- Provided by Chain of Demand
- 2 spreadsheets from brands Uniqlo and Esprit containing item-level features
- 2382 Uniqlo items; 188 Esprit items



Product.Name <fctr>	Category_1 <fctr>	Category_2 <fctr>	Color <fctr>	Material <fctr>
AIRism SLEEVELESS TOP 422267	women	sport utility wear	GRAY	59% Nylon, 31% Cupro, 10% Spandex
HEATTECH SLEEVELESS TOP 418327	women	innerwear & loungewear	OFF WHITE	38% Polyester, 32% Acrylic, 21% Rayo
CORDUROY MINI SKIRT 418882	women	bottoms	YELLOW	99% Cotton, 1% Spandex
FLEECE SET (LONG SLEEVE) 421705	women	tops	BROWN	Tops: 100% Polyester/ Rib: 90% Cotto
HEATTECH WARM LINED PANTS 420360	women	bottoms	NAVY	Shell: 90% Polyester, 10% Spandex/ L

Dataset Preview



WOMEN HEATTECH SLEEVELESS TOP 418327

Business hours of our store may need to be adjusted depending on situations. Please take note before visiting our store if you have chosen Store Pick-up service.

HK$39.00 Original Price: HK$79.00 ★★★★★

Offer: FREE DELIVERY OVER $250 WITH SINGLE DELIVERY METHOD

Color: 01 OFF WHITE

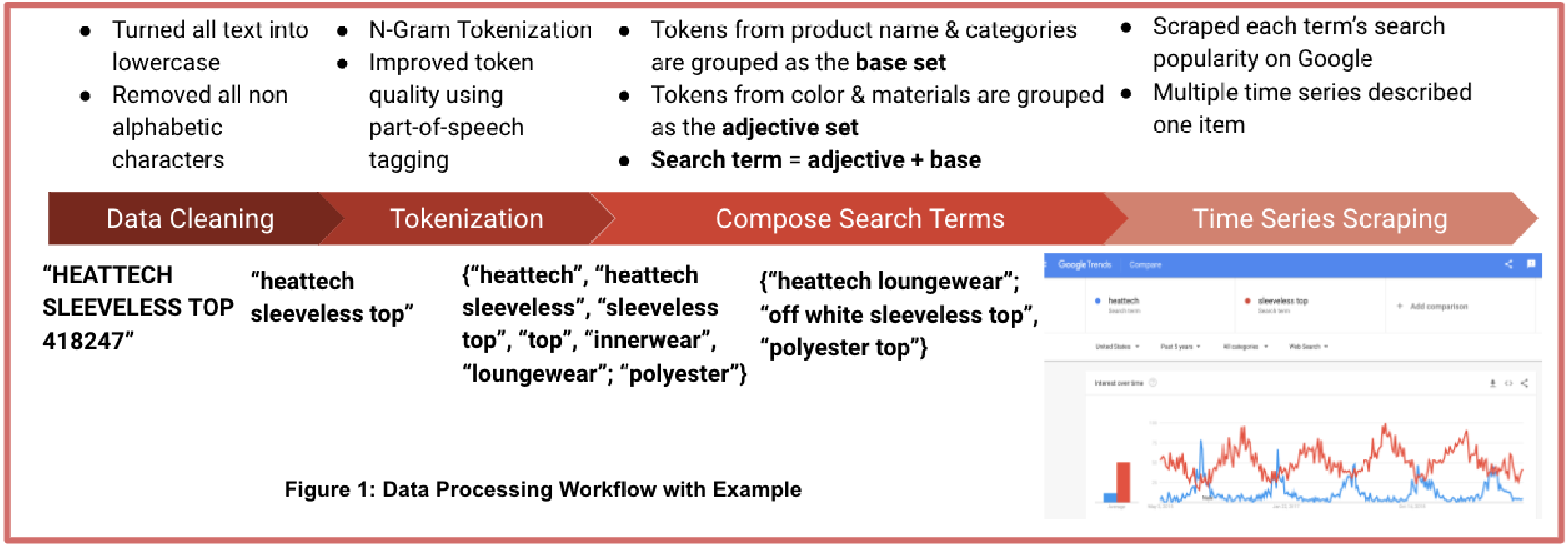
Size: S

Delivery method: ○ Delivery / S.F. Pick-up Point ○ Store Pick-up

Quantity: − 1 + IN STOCK

Example Item from Uniqlo Dataset

DATA PROCESSING STEPS



- Turned all text into lowercase
- Removed all non alphabetic characters

- N-Gram Tokenization
- Improved token quality using part-of-speech tagging

- Tokens from product name & categories are grouped as the **base set**
- Tokens from color & materials are grouped as the **adjective set**
- **Search term = adjective + base**

- Scraped each term's search popularity on Google
- Multiple time series described one item

Data Cleaning → Tokenization → Compose Search Terms → Time Series Scraping

"HEATTECH SLEEVELESS TOP 418247"

"heattech sleeveless top"

{"heattech", "heattech sleeveless", "sleeveless top", "top", "innerwear", "loungewear"; "polyester"}

{"heattech loungewear"; "off white sleeveless top", "polyester top"}

Figure 1: Data Processing Workflow with Example

Search Term Generation

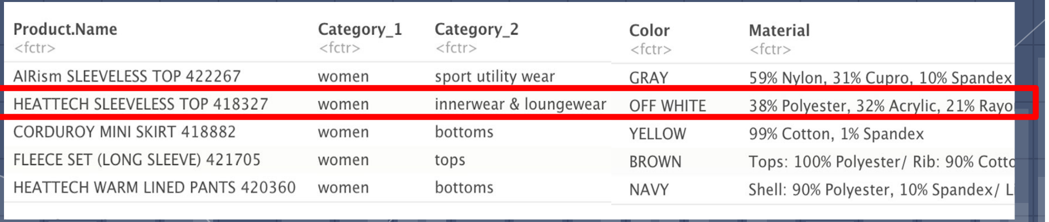
Time Series Generation

DATA PROCESSING: SEARCH TERM GENERATION

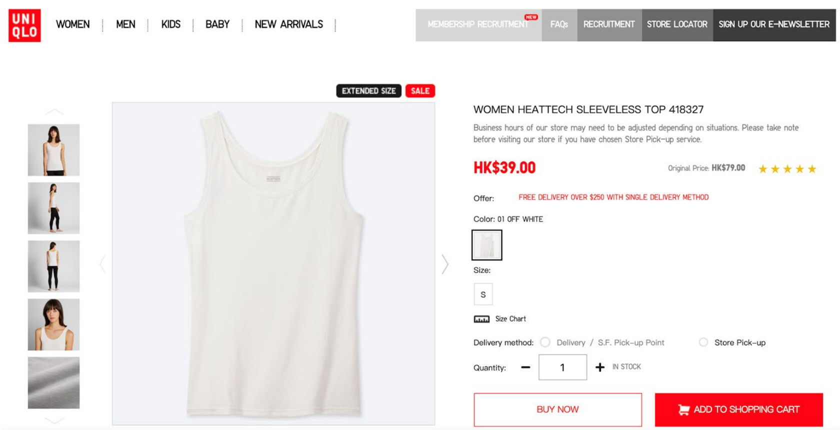
- **Goal**: For each item, automatically extract search terms similar to how people would search this item online

- Example: Heattech Sleeveless Top

{sleeveless top; heattech innerwear; off white loungewear; polyester sleeveless top…} are all possible online search terms people might choose



Product.Name <fctr>	Category_1 <fctr>	Category_2 <fctr>	Color <fctr>	Material <fctr>
AIRism SLEEVELESS TOP 422267	women	sport utility wear	GRAY	59% Nylon, 31% Cupro, 10% Spandex
HEATTECH SLEEVELESS TOP 418327	women	innerwear & loungewear	OFF WHITE	38% Polyester, 32% Acrylic, 21% Rayo
CORDUROY MINI SKIRT 418882	women	bottoms	YELLOW	99% Cotton, 1% Spandex
FLEECE SET (LONG SLEEVE) 421705	women	tops	BROWN	Tops: 100% Polyester/ Rib: 90% Cotto
HEATTECH WARM LINED PANTS 420360	women	bottoms	NAVY	Shell: 90% Polyester, 10% Spandex/ L



WOMEN HEATTECH SLEEVELESS TOP 418327

Business hours of our store may need to be adjusted depending on situations. Please take note before visiting our store if you have chosen Store Pick-up service.

HK$39.00 Original Price: HK$79.00 ★★★★★

Offer: FREE DELIVERY OVER $250 WITH SINGLE DELIVERY METHOD

Color: 01 OFF WHITE

Size: S

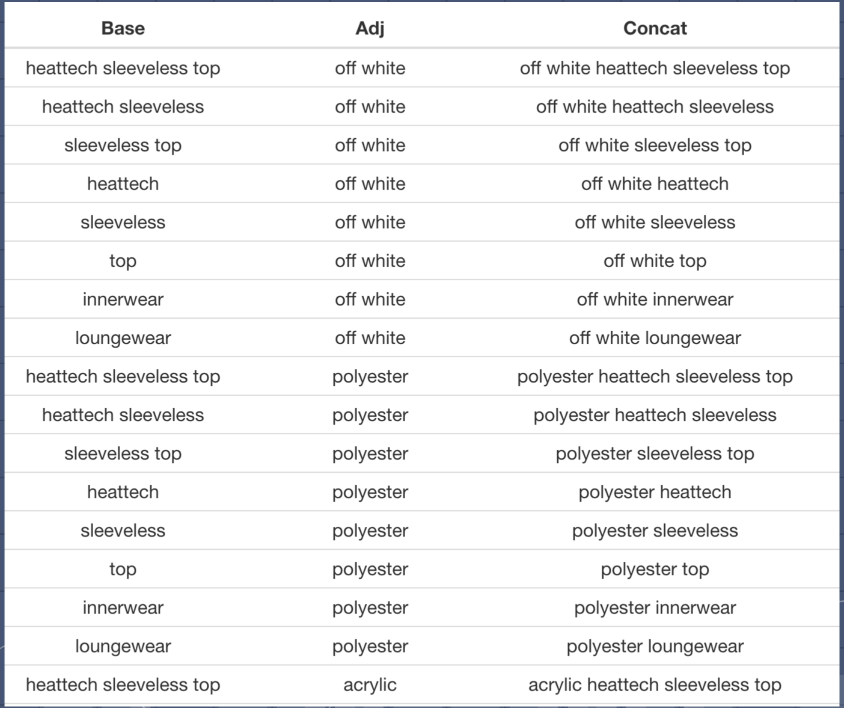
Size Chart

Delivery method: ○ Delivery / S.F. Pick-up Point ○ Store Pick-up

Quantity: − 1 + IN STOCK

Example Item from Uniqlo Dataset “heattech sleeveless top”

SEARCH TERM GENERATION: Before

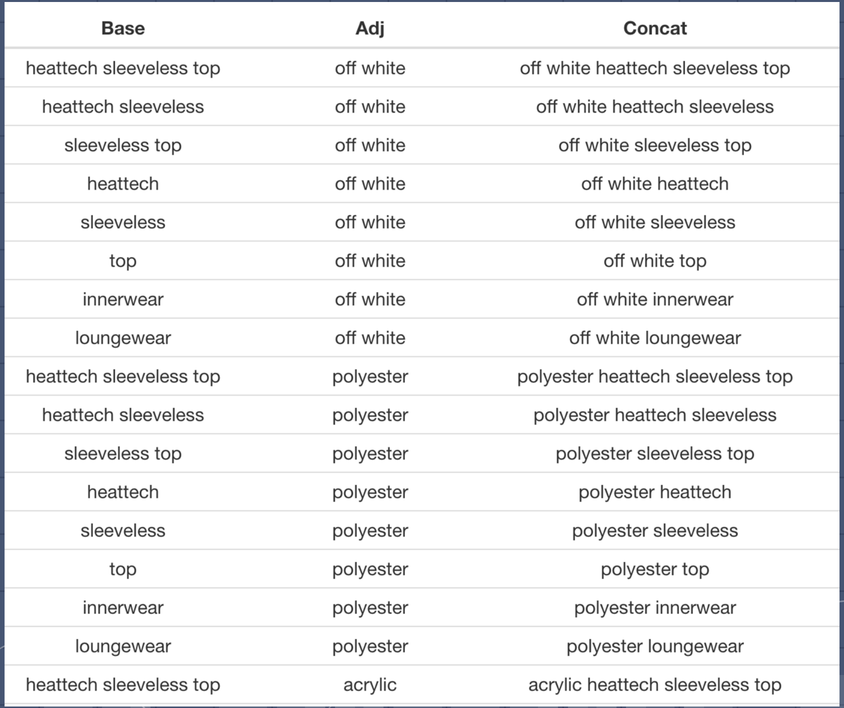


Base	Adj	Concat
heattech sleeveless top	off white	off white heattech sleeveless top
heattech sleeveless	off white	off white heattech sleeveless
sleeveless top	off white	off white sleeveless top
heattech	off white	off white heattech
sleeveless	off white	off white sleeveless
top	off white	off white top
innerwear	off white	off white innerwear
loungewear	off white	off white loungewear
heattech sleeveless top	polyester	polyester heattech sleeveless top
heattech sleeveless	polyester	polyester heattech sleeveless
sleeveless top	polyester	polyester sleeveless top
heattech	polyester	polyester heattech
sleeveless	polyester	polyester sleeveless
top	polyester	polyester top
innerwear	polyester	polyester innerwear
loungewear	polyester	polyester loungewear
heattech sleeveless top	acrylic	acrylic heattech sleeveless top

Example of Composed Search Terms for Item "heattech sleeveless top"

- Tokens generated by **N-gram tokenization** on product name, category, color, and materials
- Tokens grouped as bases/adjectives based on where they came from (product name: base; color: adjective)
- All combinations of **adjective + base**

SEARCH TERM GENERATION: Improved



Base	Adj	Concat
heattech sleeveless top	off white	off white heattech sleeveless top
heattech sleeveless	off white	off white heattech sleeveless
sleeveless top	off white	off white sleeveless top
heattech	off white	off white heattech
sleeveless	off white	off white sleeveless
top	off white	off white top
innerwear	off white	off white innerwear
loungewear	off white	off white loungewear
heattech sleeveless top	polyester	polyester heattech sleeveless top
heattech sleeveless	polyester	polyester heattech sleeveless
sleeveless top	polyester	polyester sleeveless top
heattech	polyester	polyester heattech
sleeveless	polyester	polyester sleeveless
top	polyester	polyester top
innerwear	polyester	polyester innerwear
loungewear	polyester	polyester loungewear
heattech sleeveless top	acrylic	acrylic heattech sleeveless top

Example of Composed Search Terms for Item
“heattech sleeveless top”

- Our previous method would generate too many tokens from product names
- Captured every adjacent two words
- Consider an alternative way to tokenize product names
- Motivated us to look into **part-of-speech tagging**

Part of Speech Tagging

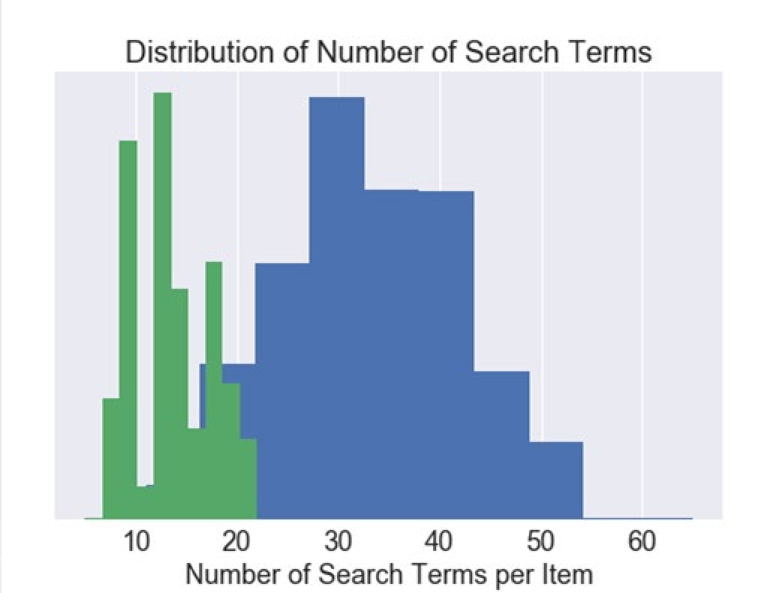
- A technique in computational linguistics
- ╟The process of marking up a word in a text as corresponding to a particular part of speech (nouns, verbs, adjectives, adverbs etc.)╢ (Wikipedia)
- There are trained models out there that take in a piece of text and return a string of tags associated with each term in the text
- We used the Natural Language Toolkit Module in Python (nltk)

Part of Speech Tagging

- How is this tool useful for us? (Our goal is to reduce the number of tokens)
  - Pass in full product names and tag each word in the name
  - Hard-code rules to break product name into two chunks (adj + base)
  - If rules fail, simply don’t break the name and use it as a base
- Examples on our texts:
  - oversized parka → [('oversized', 'JJ'), ('parka', 'NN')]
  - blocktech coat → [('blocktech', 'NN'), ('coat', 'NN')]
  - denim oversized jacket → [('denim', 'JJ'), ('oversized', 'JJ'), ('jacket', 'NN')]

Part of Speech Tagging

- We were able to substantially reduce the number of tokens, and hence reduce the number of search terms per item (search term = adj + base)



Distribution of Number of Search Terms Generated per Item

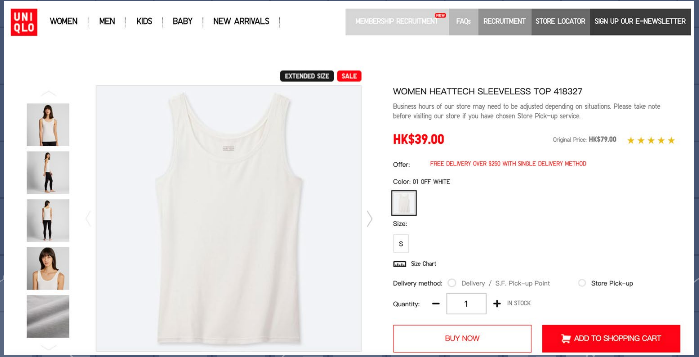
Blue: Previous Method
Green Current Method

Part of Speech Tagging: Issues

- Examples on our texts:
  - oversized parka → [('oversized', 'JJ'), ('parka', 'NN')] ☑
  - blocktech coat →[('blocktech', 'NN'), ('coat', 'NN')] ☑
  - denim oversized jacket →[('denim', 'JJ'), ('oversized', 'JJ'), ('jacket', 'NN')] ☑
  - jersey relaxed jacket →[('jersey', 'NN'), ('relaxed', 'VBD'), ('jacket', 'NN')] ✕
  - blocktech parka → [('blocktech', 'FW'), ('parka', 'FW')] ?
- Why this happened?
  - Trained models tag words based on definition and context
  - Our texts are very short, and are not complete sentences
  - Do not offer sufficient context
  - Even with sufficient context, many speech taggers are not expected to be perfect, especially in handling ambiguity

DATA PROCESSING: SEARCH TERM GENERATION
Summary

- Up to this point, each item is associated with a number of search terms
- What's next? For each search term, scrap its historical search popularity over time from Google Trends. (Time Series Generation)



WOMEN HEATTECH SLEEVELESS TOP 418327

Business hours of our store may need to be adjusted depending on situations. Please take note before visiting our store if you have chosen Store Pick-up service.

HK$39.00 Original Price: HK$79.00 ★★★★★

Offer: FREE DELIVERY OVER $250 WITH SINGLE DELIVERY METHOD

Color: 01 OFF WHITE

Size: S

Delivery method: ○ Delivery / S.F. Pick-up Point ○ Store Pick-up

Quantity: − 1 + IN STOCK

Item “heattech sleeveless top”

['top', 'heattech sleeveless top', 'heattech', 'off white loungewear', 'polyester top', 'off white innerwear', 'heattech sleeveless innerwear', 'heattech sleeveless heattech', 'polyester heattech', 'heattech sleeveless loungewear', 'off white heattech', 'polyester loungewear', 'innerwear ', ' loungewear', 'off white top', 'polyester innerwear', 'heattech sleeveless']

Search terms generated for this item

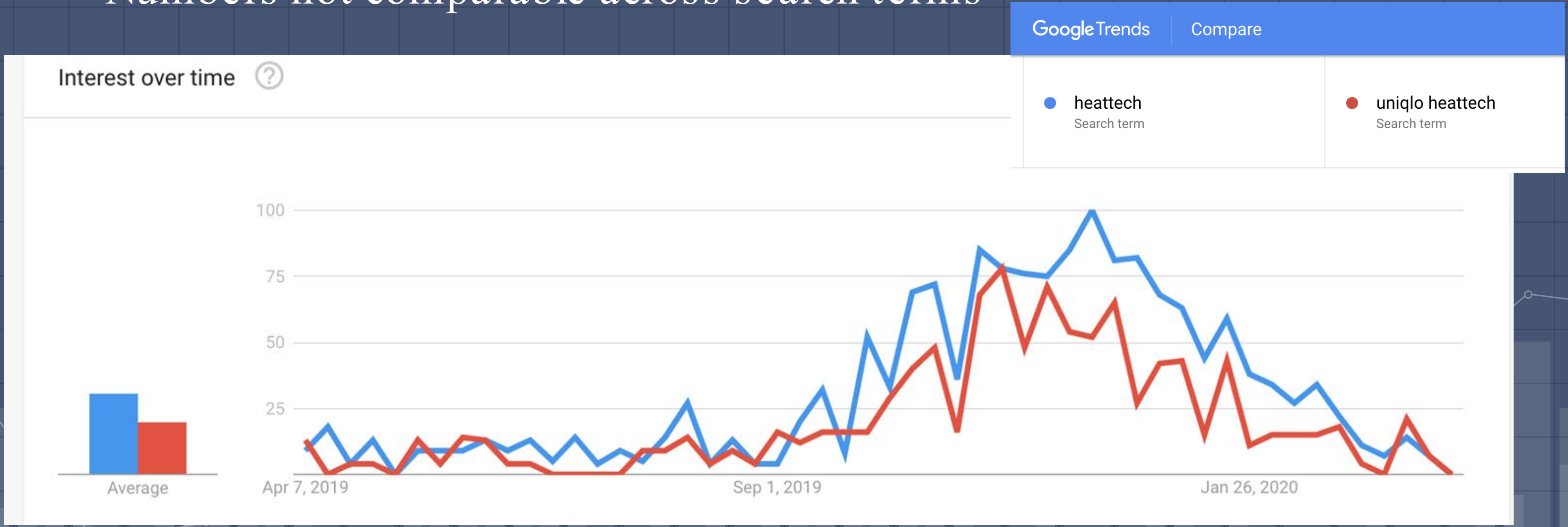
DATA PROCESSING: TIME SERIES GENERATION

For each item,

- Feed the list of search terms into Google Trends (used Python module called *pytrends*)
- Download historical interest over time for each term
- Maxed out on Google's rate limit for API calls
- Generated time series for 100 Uniqlo item examples
- Weekly timestamps from Jan. 2016 to Dec. 2019

DATA PROCESSING: TIME SERIES GENERATION

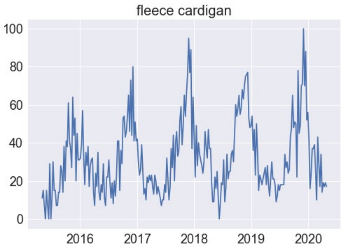
- Score from 0 to 100 is assigned to indicate each term’s current search frequency compared to historical high
- Numbers not comparable across search terms

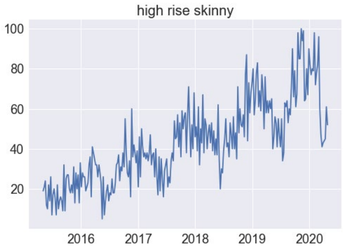


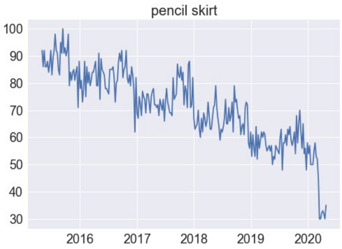
Example Time Series

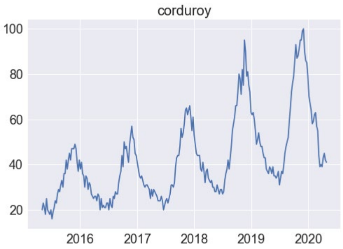
GROUPING ITEMS BY TIME SERIES

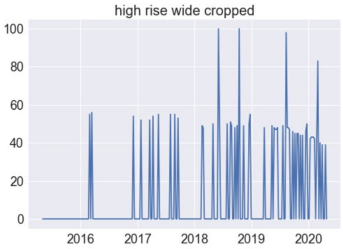
- Capture **cyclical, rising, dropping, or booming trends**
- Example search term time series







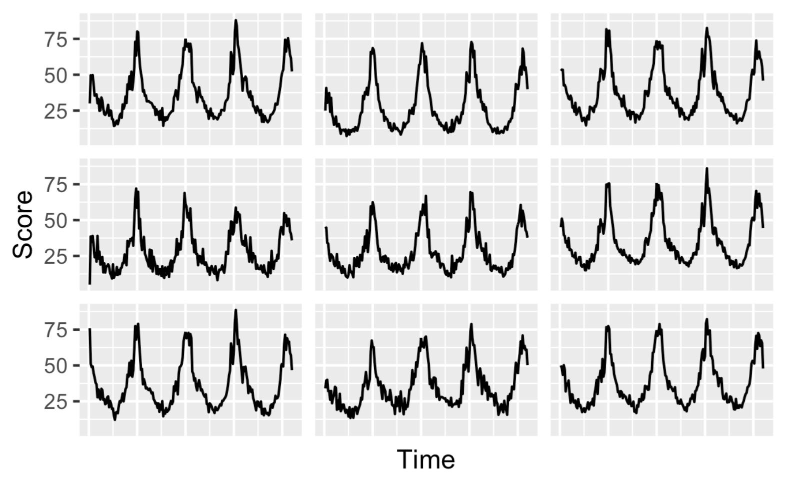




GROUPING ITEMS BY TIME SERIES

Aggregation:

- For each item, have several time series, one per term
- All terms' trend scores are averaged for each item

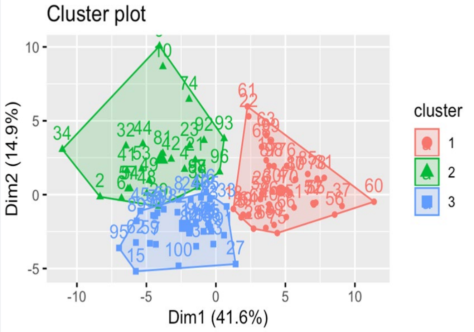


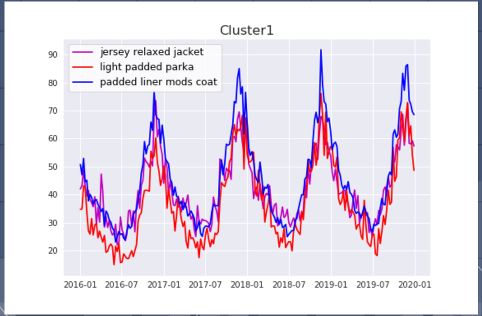
TIME SERIES FEATURE EXTRACTION

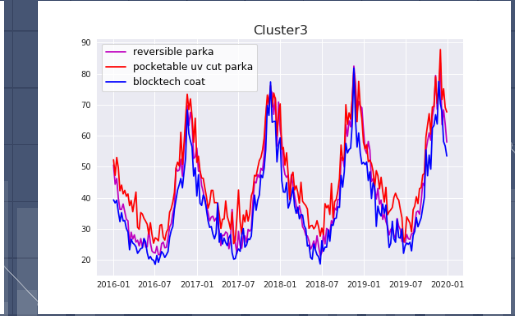
- Features extracted using TSFEL (Time Series Feature Extraction Library) package with three domains:
  - Spectral: describe the spectrum by representing the series with combinations of sinusoids
  - Statistical: describe the observed values |=marginal distribution
  - Temporal: describe the time features of the native time series
- 48 features like autocorrelation, entropy, percentile, and slope.

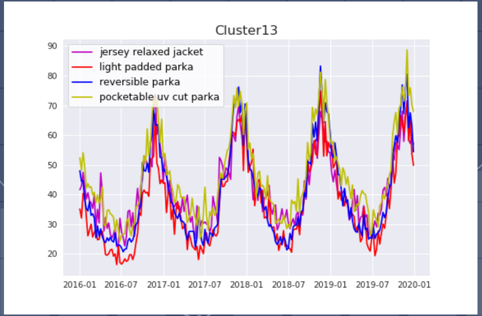
CLUSTERING TIME SERIES

- K-means Algorithm (Citation)
- Use K=3 clusters
- Cluster differences not very clear









EXTENSIONS AND NEXT STEPS

- Data points in sample too similar
  - API limitation on downloadable data
- Search term categories too heavily weighted
  - Many search terms contain overlapping keywords
  - Should put heavier weighting on item-unique words
- Figure out what features drive the clustering
  - Hard to find most "important" features based purely on sight alone

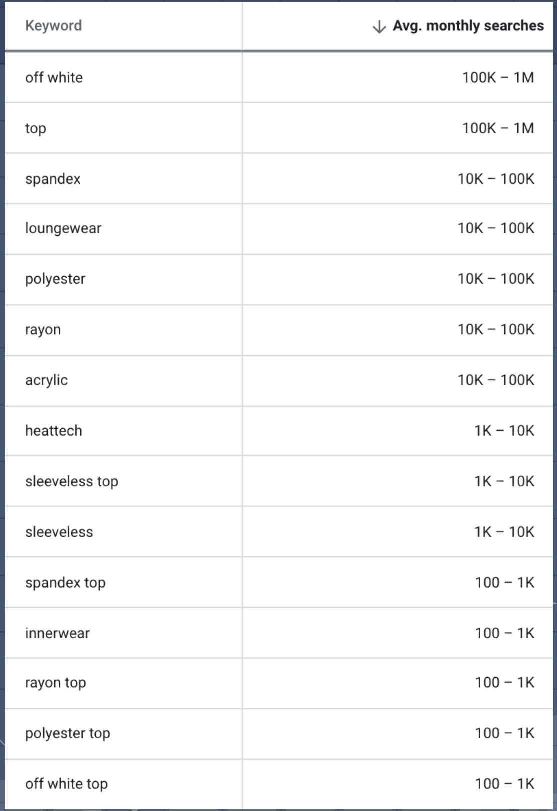
PART OF SPEECH TAGGING: FUTURE

- We believe there ╪s value in using this speech tagging tool
- Could be used to tag full sentences from product descriptions, social media
- Use cases:
  - Narrowing down key terms related to fashion trends
    - Adverbs may be less likely to be fashion concepts than adjectives and nouns
  - Sentiment analysis
    - Adjective-noun combinations may be more likely to be sentiment bearing (e.g. ‖╞brilliant acting╡ and ‖╞mediocre performance╡)

ALTERNATIVES WE TRIED

Google

- Monthly search volume from Google Ads
  - Monthly search volume’s range is too big (ex. 1K-10K,10K-100K)
  - Many keywords had very small search volume

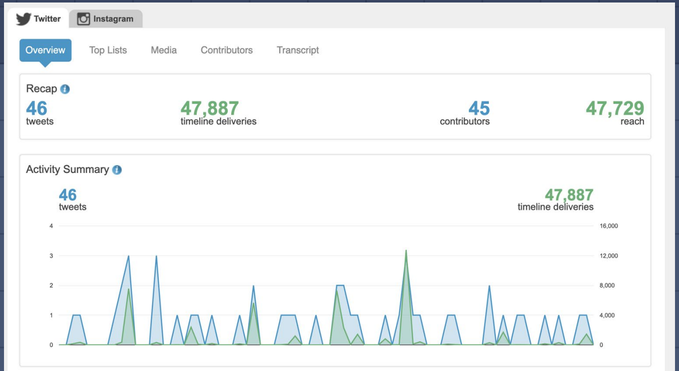


Keyword	↓ Avg. monthly searches
off white	100K – 1M
top	100K – 1M
spandex	10K – 100K
loungewear	10K – 100K
polyester	10K – 100K
rayon	10K – 100K
acrylic	10K – 100K
heattech	1K – 10K
sleeveless top	1K – 10K
sleeveless	1K – 10K
spandex top	100 – 1K
innerwear	100 – 1K
rayon top	100 – 1K
polyester top	100 – 1K
off white top	100 – 1K

ALTERNATIVES WE TRIED

Instagram/Twitter

- Attempted to input search terms into a keyword/hashtag tracking service
    - Tracking service not applicable for high volume of keywords
    - Limited historical data available



Recap

46 tweets

47,887 timeline deliveries

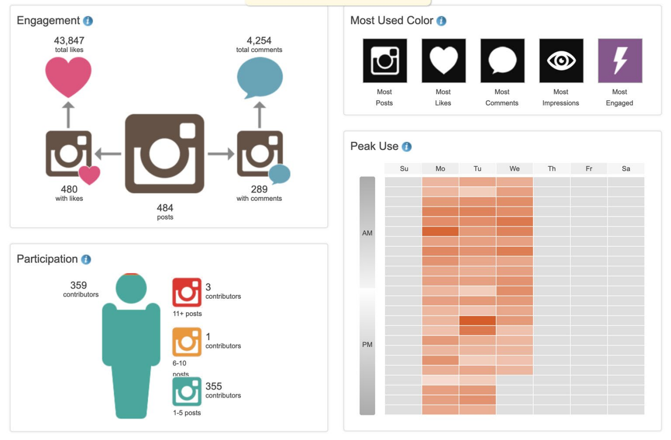
45 contributors

47,729 reach

Activity Summary

46 tweets

47,887 timeline deliveries



Engagement

43,847 total likes

4,254 total comments

480 with likes

484 posts

289 with comments

Most Used Color

Peak Use

Participation

359 contributors

3 contributors 11+ posts

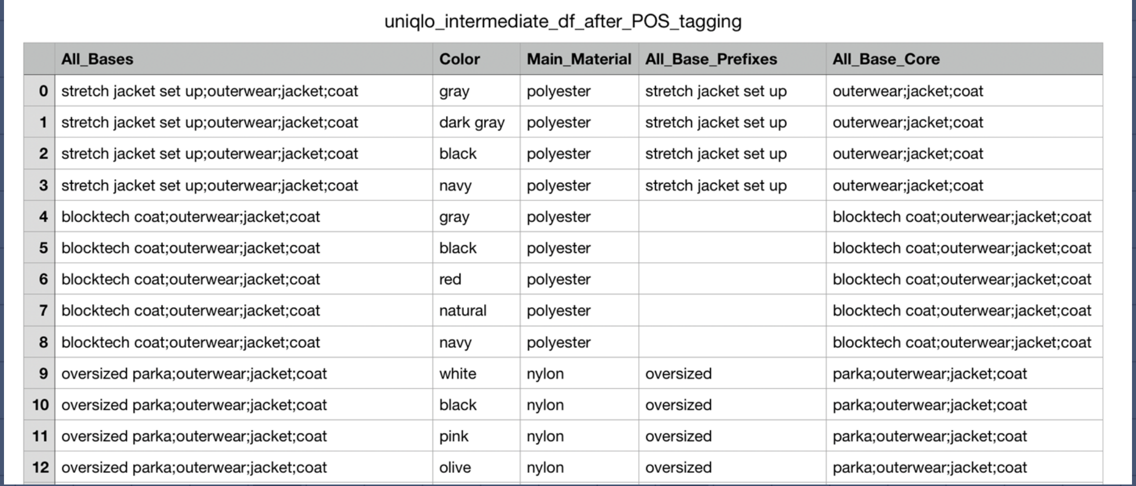
1 contributors 6-10 posts

355 contributors 1-5 posts

REFERENCES

- RWeka Library in R for tokenization
- Natural Language Toolkit library nltk for part of speech tagging
- Pytrends for scraping Google Trends data
- TSFEL (Time Series Feature Extraction Library) Package for time series feature extraction
- Base R K-means Algorithm for clustering
- Factoextra Package in R for visualization of clustering
- Data was provided by Chain of Demand, a firm which uses AI and big data to track and predict sales demand for fashion brands

QUESTIONS



uniqlo_intermediate_df_after_POS_tagging

	All_Bases	Color	Main_Material	All_Base_Prefixes	All_Base_Core
0	stretch jacket set up;outerwear;jacket;coat	gray	polyester	stretch jacket set up	outerwear;jacket;coat
1	stretch jacket set up;outerwear;jacket;coat	dark gray	polyester	stretch jacket set up	outerwear;jacket;coat
2	stretch jacket set up;outerwear;jacket;coat	black	polyester	stretch jacket set up	outerwear;jacket;coat
3	stretch jacket set up;outerwear;jacket;coat	navy	polyester	stretch jacket set up	outerwear;jacket;coat
4	blocktech coat;outerwear;jacket;coat	gray	polyester		blocktech coat;outerwear;jacket;coat
5	blocktech coat;outerwear;jacket;coat	black	polyester		blocktech coat;outerwear;jacket;coat
6	blocktech coat;outerwear;jacket;coat	red	polyester		blocktech coat;outerwear;jacket;coat
7	blocktech coat;outerwear;jacket;coat	natural	polyester		blocktech coat;outerwear;jacket;coat
8	blocktech coat;outerwear;jacket;coat	navy	polyester		blocktech coat;outerwear;jacket;coat
9	oversized parka;outerwear;jacket;coat	white	nylon	oversized	parka;outerwear;jacket;coat
10	oversized parka;outerwear;jacket;coat	black	nylon	oversized	parka;outerwear;jacket;coat
11	oversized parka;outerwear;jacket;coat	pink	nylon	oversized	parka;outerwear;jacket;coat
12	oversized parka;outerwear;jacket;coat	olive	nylon	oversized	parka;outerwear;jacket;coat

- The first 3 columns are derived from the given spreadsheet
  - **All_Bases**: product name and categories separated by semicolon. These are “bases” before tokenization
  - **Color**: copied from given spreadsheet. Different from the given spreadsheet, each row is an item with a specific color
  - **Main_Material** : first material from the list of materials in the given spreadsheet
- Column 4 and 5 are derived from **All_Bases**
  - Some tokens are used as prefixes (**All_Base_Prefixes**)
  - Some tokens remain bases, separated by semicolon (**All_Base_Core**)
  - This tokenization step is implemented via handcraft rules based on part-of-speech tagging.
- To get a concatenated term, take one token from one of the prefix columns (**Color**, **Main_Material** , **All_Base_Prefixes**) and append one token from the base column (**All_Base_Core**) (e.g. polyester outerwear in the first row)