CPI AUTOCODER

Branch of Consumer Prices

Nicole Shepler — Supervisory Economist

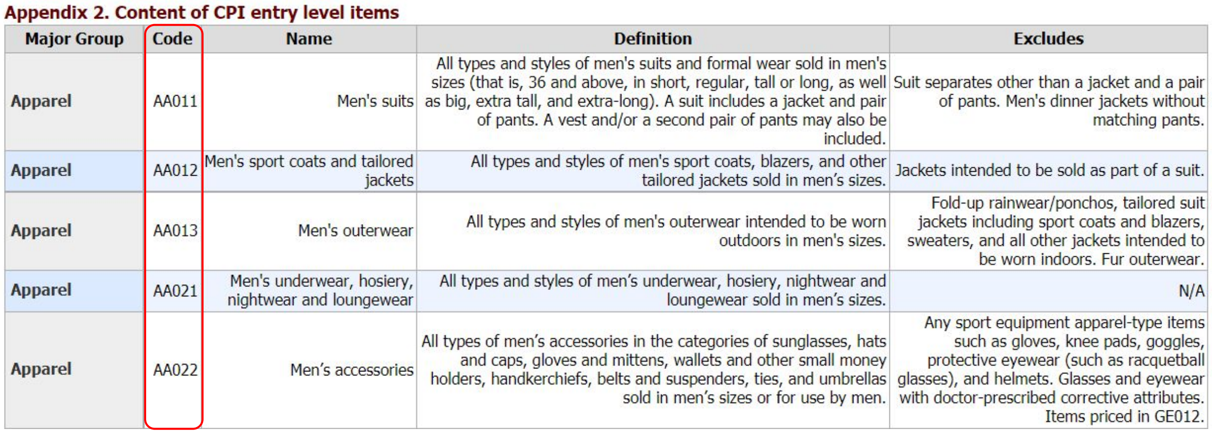
CIVIC
DIGITAL
FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

SYDNEY TRIEU
University of California, Berkeley
Data Science

BACKGROUND

- Use **alternative data sources** to produce the Consumer Price Index (CPI)
  - Includes third-party, corporate, and webscraped data
- Assign **entry level item (ELI) codes** to each item through **autocoding**



Appendix 2. Content of CPI entry level items

Major Group	Code	Name	Definition	Excludes
Apparel	AA011	Men's suits	All types and styles of men's suits and formal wear sold in men's sizes (that is, 36 and above, in short, regular, tall or long, as well as big, extra tall, and extra-long). A suit includes a jacket and pair of pants. A vest and/or a second pair of pants may also be included.	Suit separates other than a jacket and a pair of pants. Men's dinner jackets without matching pants.
Apparel	AA012	Men's sport coats and tailored jackets	All types and styles of men's sport coats, blazers, and other tailored jackets sold in men's sizes.	Jackets intended to be sold as part of a suit.
Apparel	AA013	Men's outerwear	All types and styles of men's outerwear intended to be worn outdoors in men's sizes.	Fold-up rainwear/ponchos, tailored suit jackets including sport coats and blazers, sweaters, and all other jackets intended to be worn indoors. Fur outerwear.
Apparel	AA021	Men's underwear, hosiery, nightwear and loungewear	All types and styles of men's underwear, hosiery, nightwear and loungewear sold in men's sizes.	N/A
Apparel	AA022	Men's accessories	All types of men's accessories in the categories of sunglasses, hats and caps, gloves and mittens, wallets and other small money holders, handkerchiefs, belts and suspenders, ties, and umbrellas sold in men's sizes or for use by men.	Any sport equipment apparel-type items such as gloves, knee pads, goggles, protective eyewear (such as racquetball glasses), and helmets. Glasses and eyewear with doctor-prescribed corrective attributes. Items priced in GE012.

https://www.bls.gov/cpi/additional-resources/entry-level-item-descriptions.htm

BACKGROUND

- Use **alternative data sources** to produce the Consumer Price Index (CPI)
    - Includes third-party, corporate, and webscraped data
- Assign **entry level item (ELI) codes** to each item through **autocoding**
- Previous work on **CorpX**, a department store's sales dataset containing predominantly **apparel** and **home furnishings**

OVERARCHING GOALS

1. Improve the existing autocoder

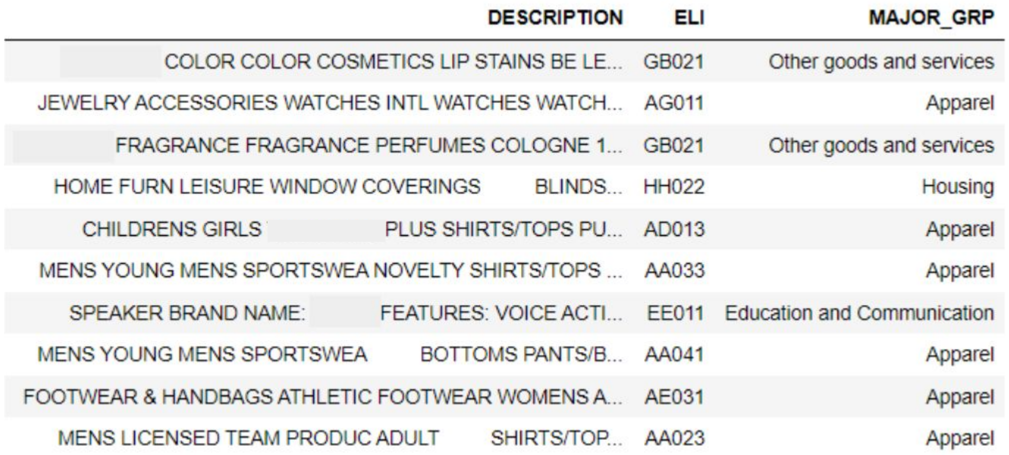
2. Generalize the autocoder

3. Provide insight into the model's decision-making

Build modeling infrastructure that reflects CPI standards

THE DATA

Structure



DESCRIPTION	ELI	MAJOR_GRP
[redacted] COLOR COLOR COSMETICS LIP STAINS BE LE...	GB021	Other goods and services
JEWELRY ACCESSORIES WATCHES INTL WATCHES WATCH...	AG011	Apparel
[redacted] FRAGRANCE FRAGRANCE PERFUMES COLOGNE 1...	GB021	Other goods and services
HOME FURN LEISURE WINDOW COVERINGS BLINDS...	HH022	Housing
CHILDRENS GIRLS [redacted] PLUS SHIRTS/TOPS PU...	AD013	Apparel
MENS YOUNG MENS SPORTSWEA NOVELTY SHIRTS/TOPS ...	AA033	Apparel
SPEAKER BRAND NAME: [redacted] FEATURES: VOICE ACTI...	EE011	Education and Communication
MENS YOUNG MENS SPORTSWEA BOTTOMS PANTS/B...	AA041	Apparel
FOOTWEAR & HANDBAGS ATHLETIC FOOTWEAR WOMENS A...	AE031	Apparel
MENS LICENSED TEAM PRODUC ADULT SHIRTS/TOP...	AA023	Apparel

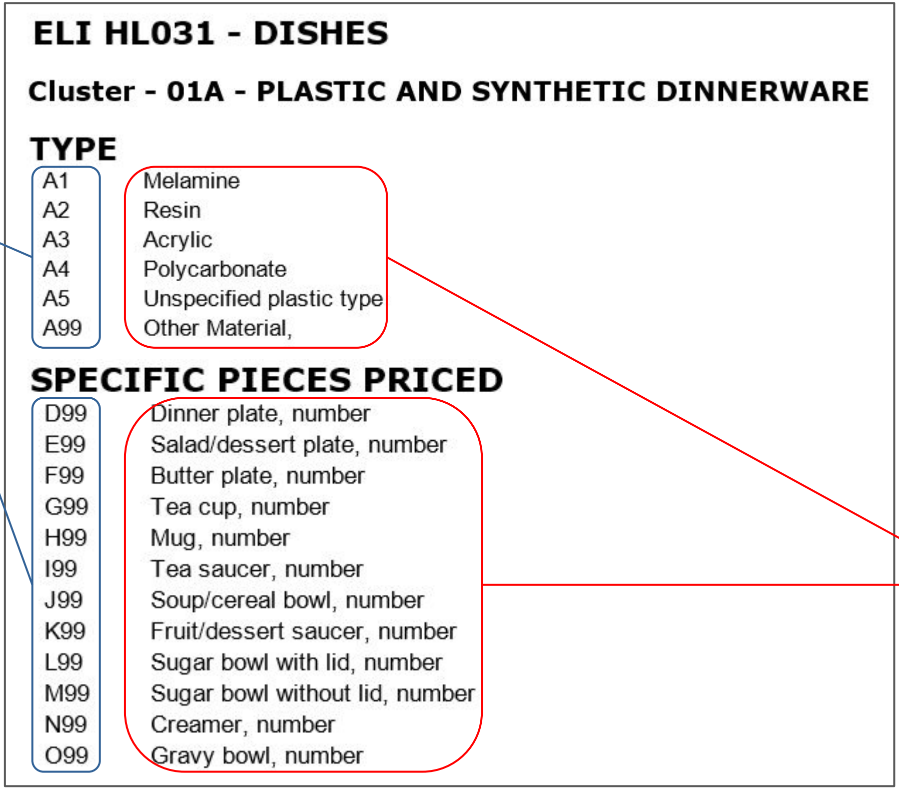
Sources

- CorpX (mostly **Apparel** and **Housing** items)
- Corp20 (mostly **Apparel** and **Housing**, but contains data from all major groups)
- ELI appendix + cluster names
- **Survey extracts**, in conjunction with survey checklists (or **specs**)

CIVIC DIGITAL FELLOWSHIP

EXAMPLE: SPECS EXCERPT

spec code



ELI HL031 - DISHES

Cluster - 01A - PLASTIC AND SYNTHETIC DINNERWARE

TYPE

A1	Melamine
A2	Resin
A3	Acrylic
A4	Polycarbonate
A5	Unspecified plastic type
A99	Other Material,

SPECIFIC PIECES PRICED

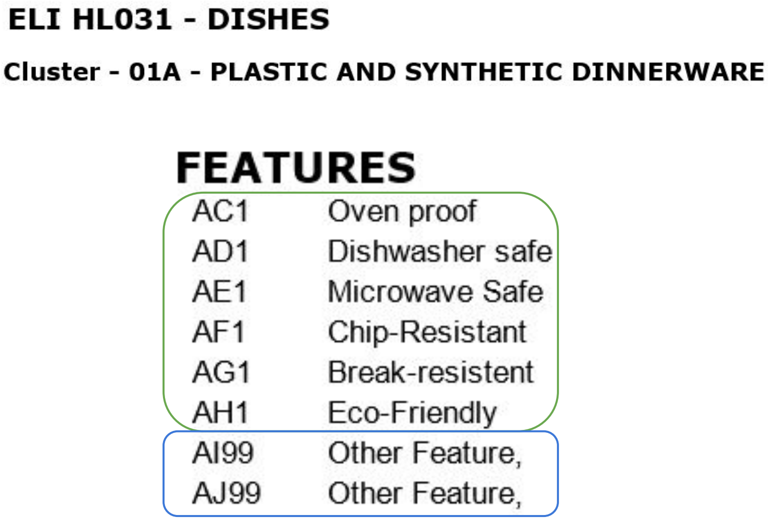
D99	Dinner plate, number
E99	Salad/dessert plate, number
F99	Butter plate, number
G99	Tea cup, number
H99	Mug, number
I99	Tea saucer, number
J99	Soup/cereal bowl, number
K99	Fruit/dessert saucer, number
L99	Sugar bowl with lid, number
M99	Sugar bowl without lid, number
N99	Creamer, number
O99	Gravy bowl, number

spec title

CIVIC
DIGITAL
FELLOWSHIP

EXAMPLE: EXTRACT DATA

Survey Checklist / Specs



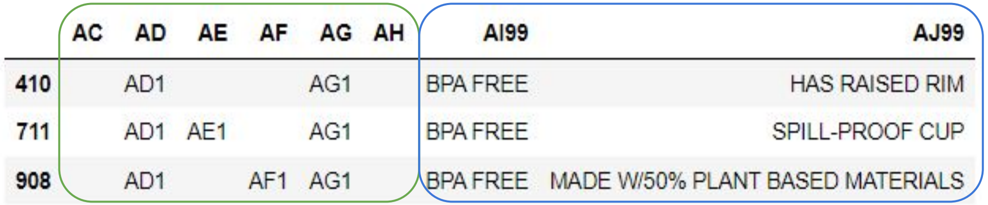
ELI HL031 - DISHES

Cluster - 01A - PLASTIC AND SYNTHETIC DINNERWARE

FEATURES

AC1	Oven proof
AD1	Dishwasher safe
AE1	Microwave Safe
AF1	Chip-Resistant
AG1	Break-resistent
AH1	Eco-Friendly
AI99	Other Feature,
AJ99	Other Feature,

Clipped Extract Sample



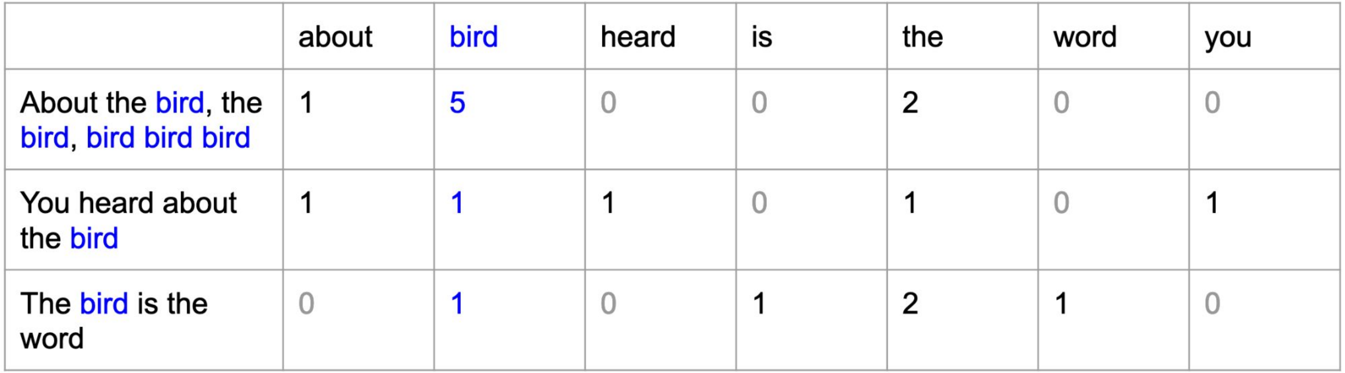
	AC	AD	AE	AF	AG	AH	AI99	AJ99
410		AD1			AG1		BPA FREE	HAS RAISED RIM
711		AD1	AE1		AG1		BPA FREE	SPILL-PROOF CUP
908		AD1		AF1	AG1		BPA FREE	MADE W/50% PLANT BASED MATERIALS

- Spec codes in **green** refer to spec titles
- Spec codes in **blue** refer to column values
- Output of row 410:

  “DISHWASHER SAFE BREAK-RESISTANT BPA FREE HAS RAISED RIM”

TEXT VECTORIZATION

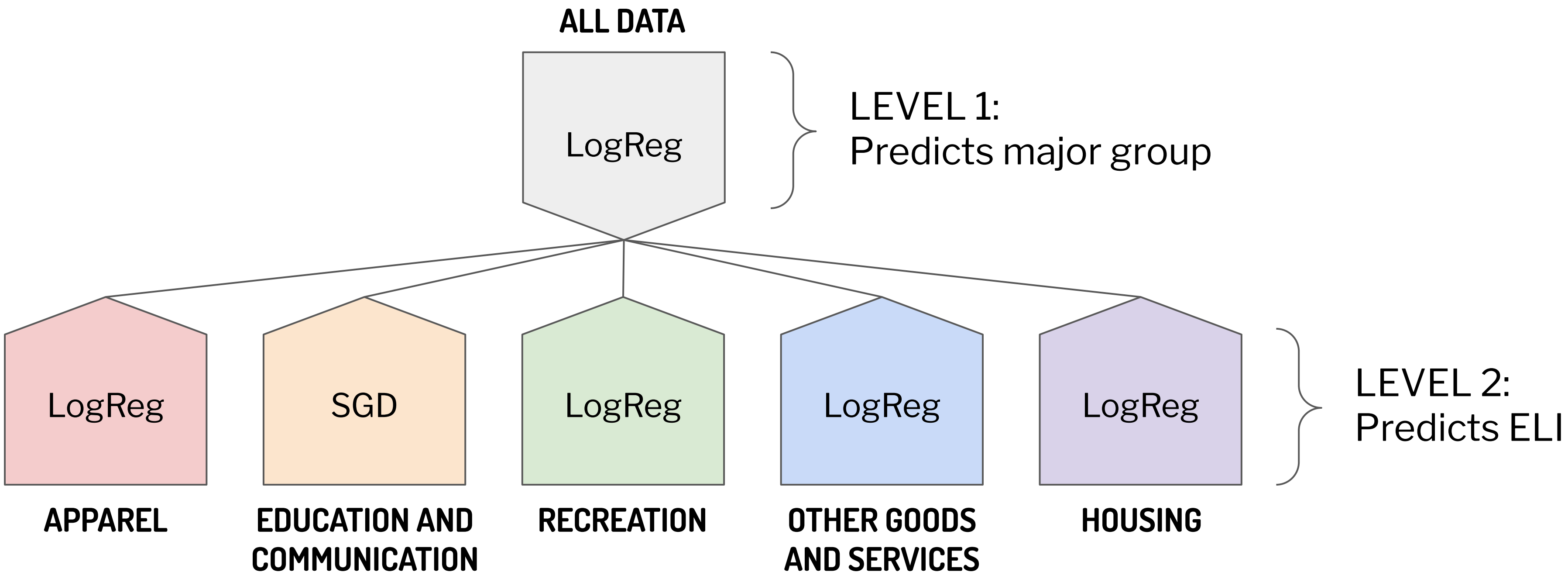
- Textual data is converted into a numeric form through **vectorization**
- **Bag-of-words** is simple yet effective, given constraints with training data



	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

https://openclassrooms.com/en/courses/6532301-introduction-to-natural-language-processing/6980811-apply-a-simple-bag-of-words-approach

HIERARCHICAL MODELS



CIVIC DIGITAL FELLOWSHIP

LogReg = logistic regression, *SGD* = linear classifier with stochastic gradient descent

FUTURE WORK

- **Feature selection** pipeline for extract data and alternative data sources
- Data cleaning and **quality assurance** for extract data
- Implementing more complex models for **text vectorization**
  - Requires diverse training data
  - Could **add noise to extract data** to avoid generalizing on the structure of descriptions
- Implementing more complex **classification models**

CLOSING THOUGHTS

& many thanks!!