AUTOCODING OCCUPATION DATA

Occupational Employment Statistics (BLS-OES)

James Walker — Branch Chief

CIVIC DIGITAL FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

VINESH KANNAN
Illinois Institute of Technology
Computer Science (‘19)

SOC AUTOCODING (Standard Occupation Classification System)

CHALLENGE

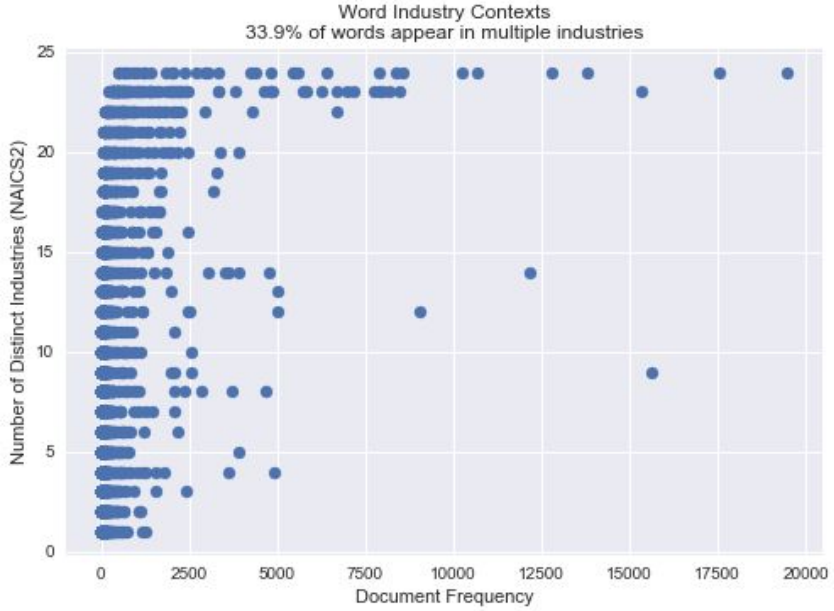
- 850 imbalanced occupations, 10 million records, 600,000 term vocabulary
- Current model ~62% accurate

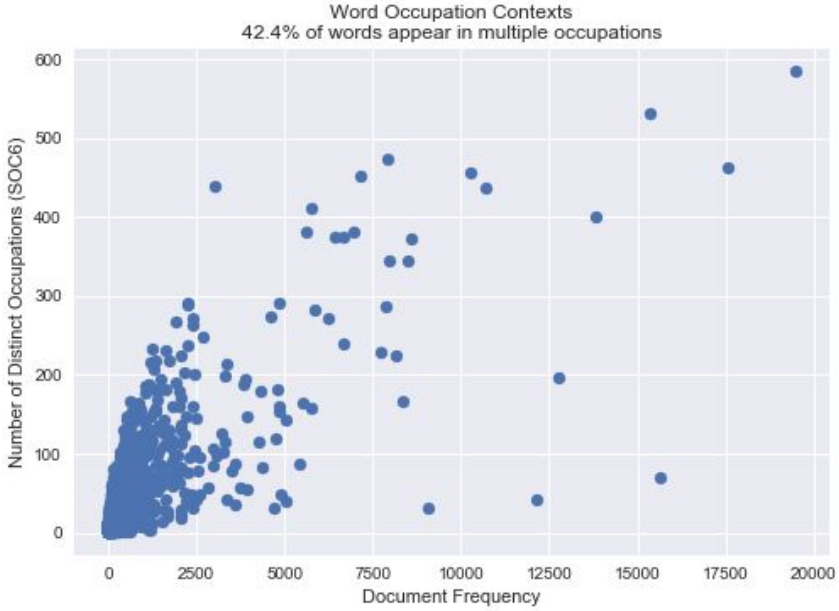
CONTRIBUTIONS

- Contextual Models (+ Open Source Module)
- Automatic discovery of 800,000+ unclassifiable records
- Addressing feedback loops

CIVIC DIGITAL FELLOWSHIP

JOB TITLES LACK CONTEXT





CIVIC DIGITAL FELLOWSHIP

Figure 1. Words that appear in multiple industries (left) and occupation groups (right).

CONTEXTUAL MODELING

- Autocoder is biased towards common jobs
- Classify rare jobs with word-industry interactions (`analyst_x_news`)
  - Computationally efficient: 23k actual interactions (10 mins) vs 600k possible (2 days)

	occupation	code	jobtitle	industry	count
0	Accountants and Auditors	13-2011	accountant	finance	7
1	Broadcast News Analysts	27-3021	news analyst	news	2
2	Computer Systems Analysts	15-1121	computer analyst	tech	2
3	Financial Analysts	13-2051	financial analyst	finance	2
4	Reporters and Correspondents	27-3022	reporter	news	7
5	Software Developers, Applications	15-1132	developer	tech	7

	correct	occupation	pred	proba	jobtitle	industry
0	False	Computer Systems Analysts	Software Developers, Applications	0.325774	analyst	tech
1	False	Broadcast News Analysts	Reporters and Correspondents	0.325774	analyst	news
2	False	Financial Analysts	Accountants and Auditors	0.325774	analyst	finance

	correct	occupation	pred	proba	jobtitle	industry
0	True	Computer Systems Analysts	Computer Systems Analysts	0.399234	analyst	tech
1	True	Broadcast News Analysts	Broadcast News Analysts	0.399234	analyst	news
2	True	Financial Analysts	Financial Analysts	0.399234	analyst	finance

CIVIC DIGITAL FELLOWSHIP

Table 1. Fictional example: interactions determine type of analyst based on industry.

MODEL RESULTS (73 candidate models)

	name	matrices	hyper	dedupe	partition	v_hacc	v_acc	v_f1_macro	v_prec_macro	v_rec_macro
14	M14	{'E', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H1	True	O4	0.71	0.629	0.441	0.569	0.407
33	M33	{'E', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H0	False	O4	0.71	0.629	0.423	0.553	0.39
15	M15	{'E', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H0	False		0.707	0.626	0.394	0.534	0.367
13	M13	{'E', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H0	True	O4	0.707	0.625	0.44	0.571	0.405
26	M26	{'E', 'WxN2', 'N2', 'M2E', 'N4', 'C', 'W', 'N6'}	H0	False		0.711	0.625	0.399	0.547	0.368
28	M28	{'E', 'M2', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H1	False		0.71	0.625	0.398	0.548	0.367
27	M27	{'E', 'M2', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H0	False		0.709	0.625	0.396	0.543	0.366
21	M21	{'E', 'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H1	False		0.707	0.625	0.394	0.534	0.366
46	M46	{'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}		False		0.703	0.622	0.389	0.526	0.362
8	M08	{'WxN2', 'N2', 'N4', 'C', 'W', 'N6'}	H0	False		0.703	0.62	0.387	0.522	0.361
60	PROD	{'E', 'W', 'N6', 'C'}		False		0.699	0.617	0.38	0.511	0.355

code	matrix
C	char grams
CxN2	char grams x NAICS-2
CxN4	char grams x NAICS-4
E	EIN
EN	narrative word grams
ENxN2	narrative word grams x NAICS-2
M2	major code
M2E	expanded major codes
M2L	leaked major code
M2LE	leaked expanded major codes
N2	NAICS-2
N4	NAICS-4
N6	NAICS-6
W	word grams
WxN2	single words x NAICS-2
WxN4	single words x NAICS-4

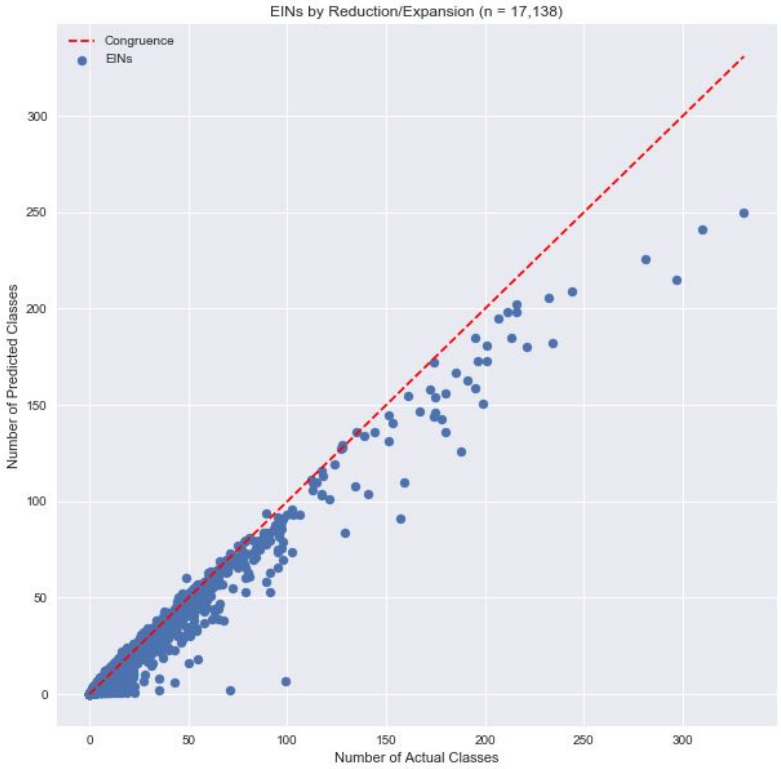
CIVIC DIGITAL FELLOWSHIP

Table 2. Comparison of production model (PROD) with top ten models.

OCCUPATION REDUCTION

Figure 2. Cases where the autocoder misrepresents diversity of occupations employed at firms.

Helped automatically find 800,000+ cases (12% of training data) of unclassifiable data submitted by employers.



CIVIC DIGITAL FELLOWSHIP