

Natural Language Processing to Help Develop a Frame of Energy Suppliers

Meghan Martin, Cindy Good, Michelle Amsbary (Westat), Francisco Cifuentes (U.S. Energy Information Administration)

Westat @ AAPOR 2022 — Take Survey Research to New Heights

The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat.

Study Background

Residential Energy Consumption Survey (RECS)

- > Household Survey
 - 19,000 households
- > Energy Supplier Survey (ESS)
 - Case = Household + Energy Source 30,000 cases
 - Respondent = Energy Supplier 3,000 suppliers



Assign each CASE to a SUPPLIER

Data Sources

Residential Energy Consumption Survey (RECS)

- > Household Survey
- > Energy Supplier Survey (ESS)

Reference list of energy suppliers from prior cycles

- Self-administered web/paper
- Supplier name, account number: open text fields

Matching Challenge



Match supplier names from HH survey to suppliers on reference list



Reference List

WASHINGTON GAS

Write-in Responses

- Washington Gas
- Washington Gas Light
- Washington Gas Light Company
- WGL

- Washington
- Wash Gas
- DC Gas
- ...

Natural Language Processing to the Rescue!

- > Search for variations on supplier names
- > Python script
- Compare HH-provided supplier name against reference list
- > Calculate <u>Levenshtein distance</u> between input text and reference list candidates
 - Value between 0 and 1
 - 0 = identical

the number of single-character edits – including insertions, deletions, and substitutions – to transform the input by the respondent into a given candidate on the reference list

Natural Language Processing to the Rescue!

- > Set threshold for *likely* matches
 - Score between 0.0 and 0.2: likely match
 - Score between 0.2 and 1.0: no likely match

- Set output guidelines
 - If there's a likely match: output 1 (best) candidate
 - If there's no likely match: output 10 candidates with lowest distance score

Reviewing the Output

MATCH

NO MATCH

ESSID	Supplier	Distance	Expanded Lookup Supplier	Lookup State	Lookup Supplier	Matched State	Matched	Project Supplier ID	State
10000000000	HIS, INDIVID		NOS. ENERGY	100	NOS, ENERGY			10.08004	
100000000000	NOS, ENERGY	1.275757575757	WISL ENGINEE	100	WILL ENGINEE			10.00	
1000000000000	NOS, ENERGY	1.270/12/12/12/13	OPE BRETON	791	OPE ENERGY			10.0000	
100000000000000000000000000000000000000	NOS, ENERGY	4.2797272727274	STE ENGINEE	-	DUE EMBROOM			10.0000	
100000000000000000000000000000000000000	NIS, INDEX		SCHOOL ENGINEE	- 04	SCHOOL SHERVY			15 46500	
100000000000000000000000000000000000000	KISS, ENGINEE		WTHIL ENGINEE	-	WTHIS DIRECTO			10 8000	
10000000000000	NURS EMERGY	120140014	VECTOR ENGINEE		NUMBER OF THE REST			10.000	
100000000000000000000000000000000000000	NISS, ENGINEE	0.0000000000000000000000000000000000000	THE EMPTOR	-	OLD DISERSON			19.07500	
100000000000000000000000000000000000000	NOS, ENERGY	1.0000000000	No disease	-	No departure			10.01108	
100000000000000000000000000000000000000	NUS. DIRECT	E-2000/00/2008	PECO BRETON	The State of the S	PECO ENERGY			19.61100	

Improving the Odds, Reducing False Positives

> Expand the reference list

- Manually: add known aliases
- Programmatically:
 - Expand common abbreviations (e.g., "CO" to "COMPANY")
 - Create acronyms or other shortened names (e.g., "Washington Gas Light" to "WGL")





> Add additional rules

Use other data elements

 (e.g., check HH state against reference list state)

Assessing the Results

> Even with 100% review of output, still much faster than matching manually



> Category flag results

Category flag	# Cases	% Cases (all)	% Cases (1-3)	
1. Likely match – confirmed	10,213	34%	42%	
2. Possible match – confirmed	4,901	16%	20%	
3. No confirmed match	9,456	32%	38%	
4. No supplier name given	5,419	18%		

Further Implications and Applications

- **>** How could we improve on our results?
 - Expand the reference list
 - Refine the rules

Category flag	# Cases
1. Likely match – confirmed	10,213
2. Possible match – confirmed	4,901
3. No confirmed match	9,456
4. No supplier name given	5,419

Further Implications and Applications

- > Looking beyond the initial data
 - Do the results from this early stage in the study correlate to any results from the later stages? YES!

Category flag	% Disavowed
1. Likely match – confirmed	4%
2. Possible match – confirmed	5%
3. No confirmed match	8%
4. No supplier name given	22%

Category flag	% Completed
1. Likely match – confirmed	97%
2. Possible match – confirmed	95%
3. No confirmed match	92%
4. No supplier name given	84%



Thank You!

MeghanMartin@westat.com

CindyGood@westat.com

MichelleAmsbary@westat.com

Francisco.Cifuentes@eia.gov







