

Top N: A Tool to Simplify Coding of Open-Ended Text Survey Responses

Curtiss Chapman

U.S. Census Bureau

FedCASIC 2025 - Virtual

April 22, 2025



This presentation is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not those of the U.S. Census Bureau. The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. CBDRB-FY25-CBSM004-005.

Problem

- Analyzing Open-ended Survey Questions
 - Time Consuming
 - Technical knowledge (NLP, programming)

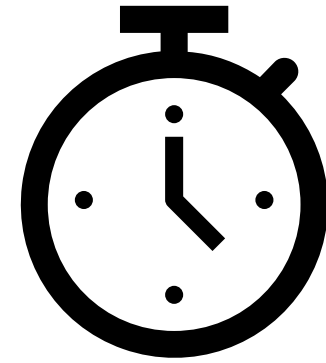
Q: What is another important reason to fill out the census?

Type response here

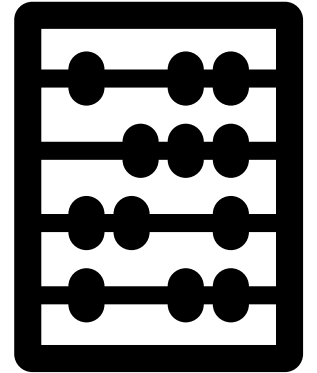
- Options:
 - Include, deal with drawbacks
 - Include, but never analyze
 - Exclude

Solution!

- Top Words Labeling Assistant
(aka Top N tool)
 - Decreases time required for labeling
 - Utilizes NLP tools
 - User only needs Excel knowledge



NLP



The Labeling Process

Q: What is another important reason to fill out the census?

	Response	Label
1	can be useful for scientific research	
2	i was told i had to do it	
3	tbh i dont know much about this	
4	if i dont, you send people to bother me	
5	genealogy purposes	
6	i don't know i just fill it out	
7	genealogy research	
8	historical records for future generations	
9	genealogy research for future generations	

The Labeling Process

Q: What is another important reason to fill out the census?

	Response	Label
1	can be useful for scientific research	Research
2	i was told i had to do it	Required
3	tbh i dont know much about this	I don't know
4	if i dont, you send people to bother me	Required
5	genealogy purposes	Research
6	i don't know i just fill it out	I don't know
7	genealogy research	Research
8	historical records for future generations	Records
9	genealogy research for future generations	Research

The Labeling Process

Q: What is another important reason to fill out the census?

	Response	Label
1	can be useful for scientific research	Research
5	genealogy purposes	Research
7	genealogy research	Research
9	genealogy research for future generations	Research
2	i was told i had to do it	Required
4	if i dont, you send people to bother me	Required
3	tbh i dont know much about this	I don't know
6	i don't know i just fill it out	I don't know
8	historical records for future generations	Records

Confirm
VS.
Decide

Top Words Labeling Assistant

	Response	Label
1	can be useful for scientific research	
2	i was told i had to do it	
3	tbh i dont know much about this	
4	if i dont , you send people to bother me	
5	genealogy purposes	
6	i dont know i just fill it out	
7	genealogy research	
8	historical records for future generations	
9	genealogy research for future generations	

Top Words Labeling Assistant



	Response	Label	wd_genealogy	wd_research	wd_dont
5	genealogy research for future generations		1	1	0
7	genealogy research		1	1	0
9	genealogy purposes		1	0	0
1	can be useful for scientific research		0	1	0
4	if i dont, you send people to bother me		0	0	1
3	tbh i dont know much about this		0	0	1
6	i dont know i just fill it out		0	0	1
8	historical records for future generations		0	0	0
2	i was told i had to do it		0	0	0

Top Words Labeling Assistant



	Response	Label	wd_genealogy	wd_research	wd_dont
5	genealogy research for future generations	Research	1	1	0
7	genealogy research	Research	1	1	0
9	genealogy purposes	Research	1	0	0
1	can be useful for scientific research	Research	0	1	0
4	if i dont, you send people to bother me	Required	0	0	1
3	tbh i dont know much about this	I don't know	0	0	1
6	i don't know i just fill it out	I don't know	0	0	1
8	historical records for future generations	Records	0	0	0
2	i was told i had to do it	Required	0	0	0

DEMO

Thank you!

Feel free to contact me at:

curtiss.a.chapman@census.gov

Streamlit

- Creating web apps with Python! No need to be a full stack developer!
- Used in data science groups to create web applications to support Data Science / Analytics.

Compatible with

BASICALLY EVERYTHING!

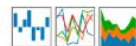
bokeh

Altair

PyTorch

OpenCV

DECK.GL

pandas 

Vega-Lite

matplotlib

NumPy

learn

TensorFlow

plotly

Keras