# A Combined Natural Language Annotation and Visualization Tool for the Exploratory Analysis of Federal Survey Response and Note Documents

Haley Hunter-Zinck
Center for Optimization and Data Science (CODS)
U.S. Census Bureau

FCSM 2024 Research and Policy Conference
October 22, 2024

United States® Census Bureau

# Disclaimer

- Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau.
- This presentation does not contain sensitive information including Title 13, Title 26, Title 5, other controlled unclassified information, or administratively restricted information.

# Exploratory analysis of survey documents requires human-computer interaction to derive insights

Motivation: facilitate exploratory analysis of survey documents

- Automate standardized annotations
- Make annotations accessible

Example use cases

- Extract information from survey notes for downstream processes
- Summarize content from an open response survey question

# SNARE: Survey Note And Response Explorer

## An interactive annotation and visualization tool

- Objectives
  - Automated routine annotations
  - Visualizations and interactivity
- Two parts
  - 1. Annotation pipeline
  - 2. Dashboard
- Technical requirements
  - Approved
  - Open source
  - Free
  - Minimal computing resources
- Out of scope
  - Applied Programming Interface (API) based tools
  - Generative Artificial Intelligence (AI) models

United States® Census Bureau

# We use a publicly accessible corpus of comments to a notice soliciting input for 2030 Census preliminary research

**DEPARTMENT OF COMMERCE**

**Census Bureau**

**[Docket Number 220526–0123]**

**Soliciting Input or Suggestions on 2030 Census Preliminary Research**

**AGENCY:** Census Bureau, Department of Commerce.

**ACTION:** Notice and request for comment.

**SUMMARY:** Early planning for the 2030 Census program began in Fiscal Year 2019 with building the program foundation and preparing for the official program kick-off and start of the Design Selection Phase in October 2021. The primary goal of the Design Selection Phase is to conduct the research, testing, and operational planning and design work to inform the selection of the 2030 Census operational design. We are issuing this notice to engage with our stakeholders on the development and implementation strategies that improve the way people participate in the 2030 Census. This notice also includes specific topics of interest to help guide input from stakeholders and other members of the public.

**DATES:** Comments on this notice must be received by November 15, 2022.

https://www.regulations.gov/document/USBC-2022-0004-0001

| Attribute | Counts |
|---|---|
| Documents | 1,445 |
| Unique documents | 1,295 |
| Median words | 101 |
| Median characters | 531 |

Note: attachments are excluded

# We focus on annotations with greatest relevance to survey use cases

1. Language identification
2. Named entity recognition (NER)
3. Sentiment analysis
4. Topic modeling

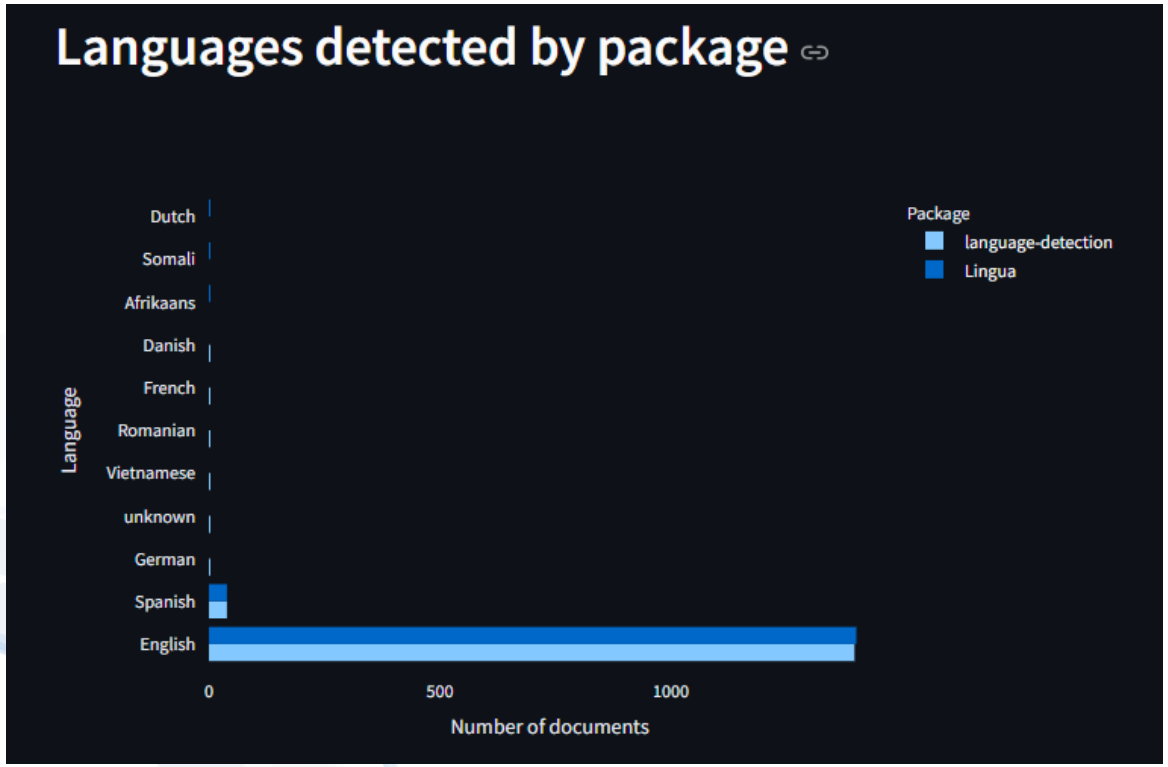
# Language identification

- Identify the language in which the text is written
- Important for quality checking and routing for downstream processing

- Python packages
  - lingua
  - language-detection

# Language identification

Annotations indicate responses are mostly English and Spanish.




Languages annotations other than English and Spanish are errors.

Language detection package

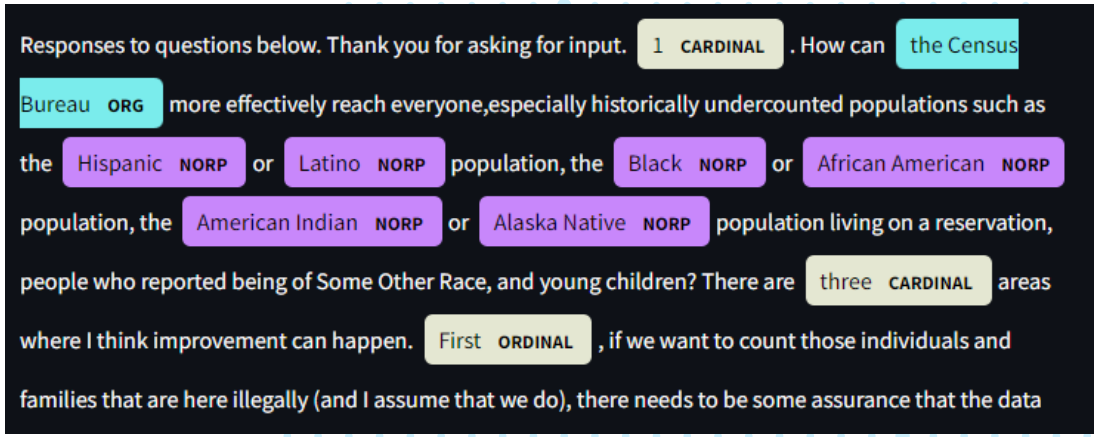
lingua

| Language | Probability | Example |
|---|---|---|
| Afrikaans | 0.0 | 41:16-41:20. |
| English | 1.0 | Ideas on: 1-) NEW DATA SOURCES The Family -Prevailing structures and trends for each classification. i-Intact (mother and father). ii-Single parent by sex (male or female). iii-Non-traditional. iv-Other. Child rearing practices. -Prevailing trends. i-Strict. ii-Permissive. iii-Erratic. iv-Science based. For each combination of categories above, determine resulting: "Personality Type" and "Socio-Economic Strata" Results will be of significant interest to "Hispanic or Latino population" and "Academic Researchers" on national socio-economic and political advancements. |
| Spanish | 1.0 | En el censo las personas esconden informaciones. Pero nosotros con nuestra acienci y arte de convencimientos logramos al maximo los posibles datos que se piden en el censo, pedimos que reguntas como usted est legal en este pas, es usted ciudadano , cules son sus actividades fuera del trabajo, tiene usted familiares ilegales en este pas etc nuestro trabajo que hacemos en la calle es muy bello , pues encontramos persona bien educada y humanas, nos hacen entender que realizamos una gran labor para el bien de la nacin. |
| Dutch | 0.06 | #NAME? |
| Somali | 0.07 | Ask me if Im gay! |

# Named entity recognition (NER)

- Label spans of texts corresponding to predefined entities
- Automated extraction of names, organizations, etc.

- Python package and models
  - spaCy small English language model (en_core_web_sm)
  - spaCy transformer English language model (en_core_web_trf)




Highlighted entities in document ID USBC-2022-0004-0153 using the spaCy transformer English language model

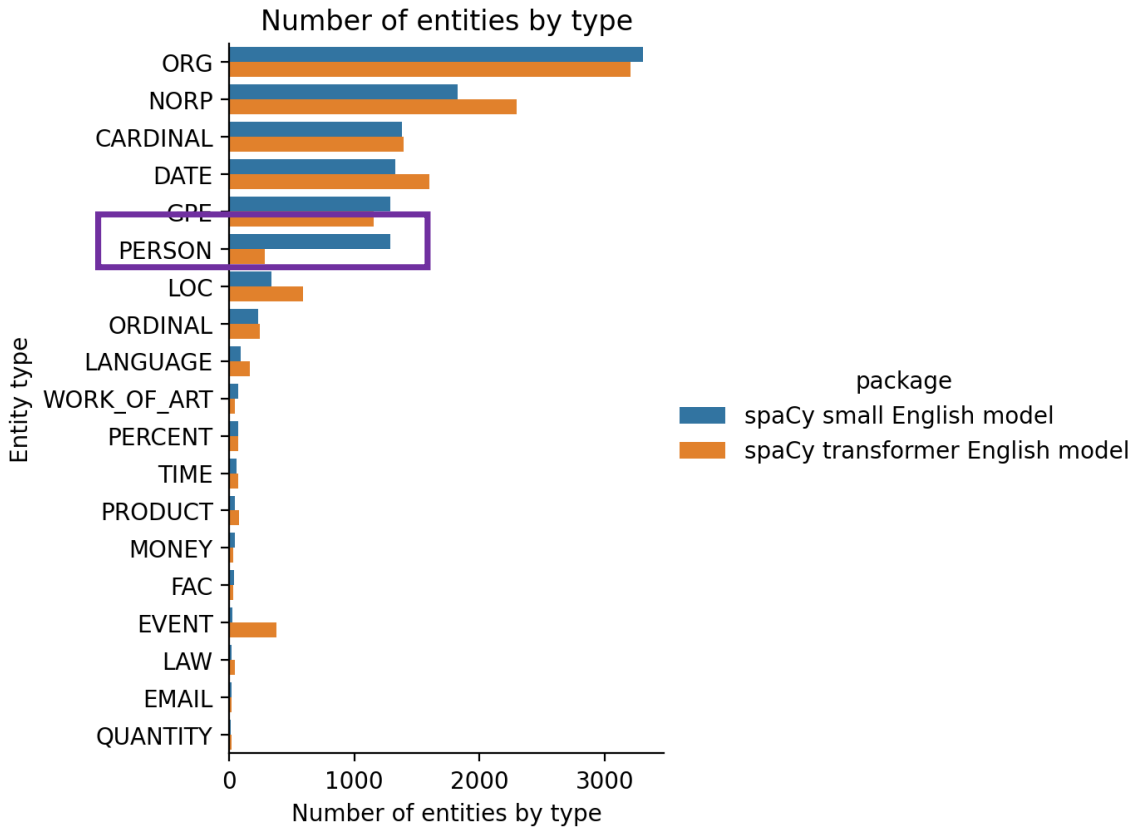
| Tag | Explanation |
|---|---|
| CARDINAL | Numerals that do not fall under another type |
| ORG | Companies, agencies, institutions, etc. |
| NORP | Nationalities or religious or political groups |
| ORDINAL | "first", "second", etc. |
| LANGUAGE | Any named language |
| DATE | Absolute or relative dates or periods |

spaCy NER tag explanations

United States® Census Bureau

# Named entity recognition (NER)

General ranking of entity types is similar between models




Some models perform better for some entity types than others (e.g. PERSON).

NER model: en_core_web_sm

NER tag: PERSON

| PERSON: People, including fictional | count |
|---|---|
| Census | 596 |
| Latino | 23 |
| Covid | 15 |
| Alaska Native | 12 |
| Garifuna | 10 |
| Que | 7 |
| Census 2030 | 5 |
| Santos | 5 |
| Recruiter | 5 |
| MQA | 4 |

NER model: en_core_web_trf

NER tag: PERSON

| PERSON: People, including fictional | count |
|---|---|
| Trump | 9 |
| Santos | 5 |
| I | 4 |
| Dan Bouk | 3 |
| danah boyd | 3 |
| Trumps | 3 |
| Donald Trump | 2 |
| Gareth Casady | 2 |
| Nancy Wolanski | 2 |
| Covid | 2 |

United States® Census Bureau

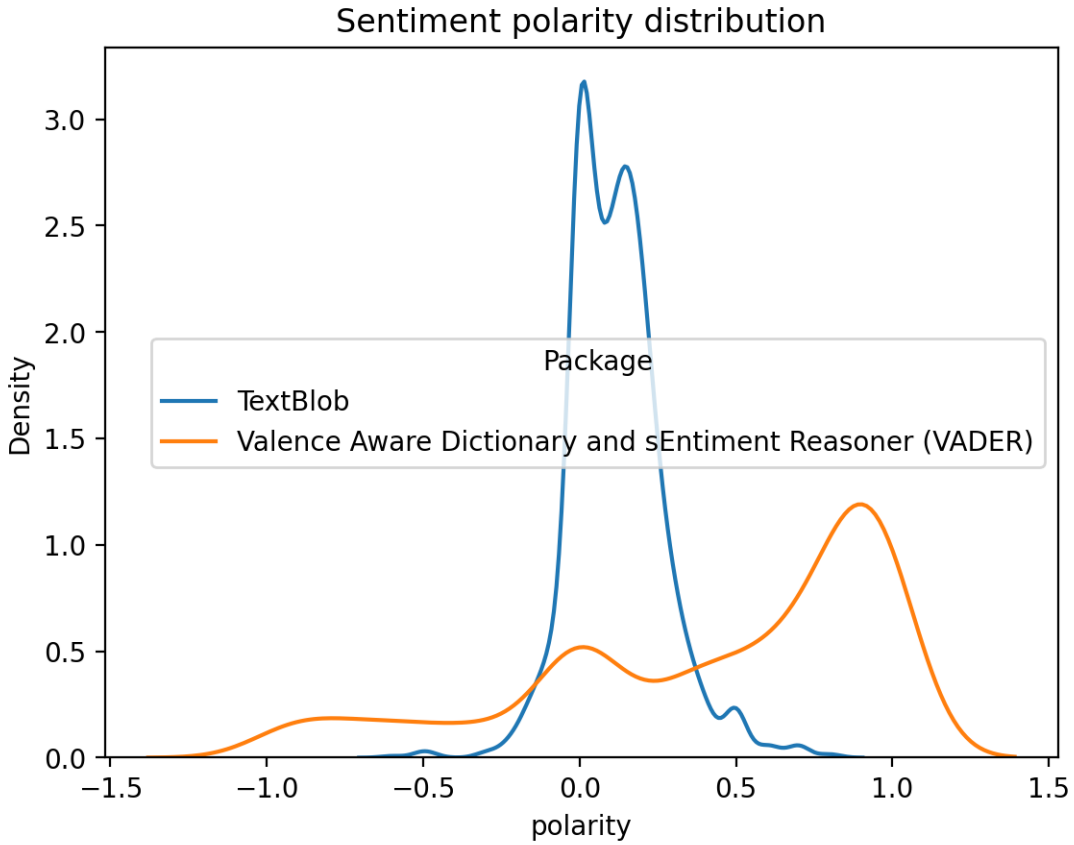
# Sentiment analysis

- Classify a document's content as positive, neutral, or negative
- Flag notes with extreme polarity
- Get sentiment distribution in a topic

- Python packages
  - textblob
  - VADER

# Sentiment analysis

Distributions of sentiment polarity for each package are different




Neither package may generalize well to our corpus

| Polarity | TextBlob | Valence Aware Dictionary and sEntiment Reasoner (VADER) |
|---|---|---|
| Highest | (USBC-2022-0004-0713) It would be great to add Neurodiversity to the census, because this group of people is growing. | (USBC-2022-0004-1399) I gather one of the considerations for the 2030 census is how to increase participation of underrepresented populations. I would like to share my story in the hopes that it will be of assistance - among the thousands of comments you probably are receiving-. I am a Hispanic woman who was very concerned about the safety of undocumented people. Provided the White House in 2018 and 2019 was attempting clear scare tactics against undocumented people, with the idea of asking people about their citizenship. I reached out to the Census bureau to ask about the safety of the personal information of participants. I recall getting a rather vague response that did not in unequivocal ways assure me that if I encouraged undocumented people to participate that their information would not be used against them. The general situation of the effort to 1) interfere in the kind of questions to be posed by the Census and 2) use the census as a means of intimidation and political gaming, leads me to believe that it is very important for the Census to be treated as an autonomous institution that is sustained by the federal government, but not dictated by it. Perhaps it is, but that did not come across well in communications from the media. This leads me to believe it is not as autonomous as it should - like the concept of the autonomous university in Mexico. There are other such autonomous entities in the federal government, therefore, it would not be unprecedented to create a clear legal framework for such stronger autonomy from the Census. My other observation is that by not building close community networks the Census is unable to gain the trust of the general population and hence to increase participation. In my city, Census Bureau people went to the farmers market. They were very kind and informative. However, they were at a farmers market where mostly middle and upper class people go (things are rather expensive there). So, they were not going to reach the underrepresented populations most effectively there. It seems to |

# Topic modeling

- Identify themes discussed in documents
- Label documents with theme

- Python packages
  - BERTopic

# Topic modeling

Automatically derived summaries provide useful information for selecting hyperparameters.

| Topic model | Number of topics | Number of outlier documents | Coherence |
|---|---|---|---|
| BERTopic (min topic size 30) | 5 | 0 | 0.499532556927852 |
| BERTopic (min topic size 10) | 10 | 7 | 0.478497041941295 |
| BERTopic (min topic size 100) | 2 | 0 | 0.335175912471618 |

# Topic modeling

Selecting hyperparameters in an automated fashion means we could miss important insights.

Highest coherence model

Top 10 topics for 'BERTopic (min topic size 30)'

| label | n |
|---|---|
| 0_the_to_and_of | 1,180 |
| 1_arab_the_and_that | 107 |
| 2_and_the_census_bureau | 60 |
| 3_attached_file_see_41 | 58 |
| 4_que_de_la_para | 40 |

Second highest coherence model

Top 10 topics for 'BERTopic (min topic size 10)'

| label | n |
|---|---|
| 0_the_to_and_in | 1,154 |
| 1_bureau_and_census_the | 59 |
| 2_arab_americans_that_and | 55 |
| 3_file_attached_see_files | 46 |
| 4_que_de_la_para | 40 |
| 5_arab_americans_that_the | 28 |
| 6_prison_prisoners_incarcerated_of | 26 |
| 7_mena_white_category_not | 18 |
| 8_attached_see_file_41 | 12 |

United States® Census Bureau

# Topic modeling

Topic identifies respondents voicing concern about where incarcerated residents are counted in the decennial census

| | |
|---|---|
| USBC-2022-0004-0250 | Hello! I am a resident of Minnesota and wanted to call attention to the practice of counting incarcerated people in the places where they are incarcerated rather than where they usually live. This causes issues with funding and resourcing and has many downstream negative implications. I request that the Census Bureau shift to a more equitable way of counting incarcerated individuals. There has been a huge increase in the number of incarcerated individuals which has caused many issues. Thank you for your consideration |
| USBC-2022-0004-0224 | Hello, Please count incarcerated people at their home, not their facility location, and end the practice of "prison gerrymandering". See https://www.prisonersofthecensus.org/ Thank you! Emily |
| USBC-2022-0004-0601 | More than 1.4 million people are incarcerated in state and federal prisons and nearly 33% of them are Black Americans. Prison gerrymandering dilutes the representational clout of large numbers of Black Americans. I sincerely hope the Census will end this unfair counting system. |

# Conclusions

- We developed an annotation pipeline and dashboard for survey document exploration.
- Interactive summarization and visualization aids in deriving insights.

- Next steps
  - Add additional annotations and tools
  - Enhance visualizations
  - Interactive labeling of topics

# Acknowledgements


- **Arezou Koohi**, CODS, U.S. Census Bureau (now at FDIC)




External

# QUESTIONS?



haley.s.hunter-zinck@census.gov