Implementing Interactive Classification Tools in the 2022 Economic Census

Emily Wiley
Daniel Whitehead
U.S. Census Bureau
October 27, 2022

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System (DMS) number: P-7504847, subproject P-7514952; Disclosure Review Board (DRB) approval number: CBDRB-FY22-ESMD009-007).
Outline

• Background and motivation
• BEACON overview
• SINCT overview
• Field Test results
• Lessons learned and conclusions
North American Industry Classification System (NAICS)

- Business establishments are classified by NAICS code based on primary business activity.
- There are prelisted descriptions, but the respondent also has the option of writing in a business description.

Source: 2017 Economic Census
North American Product Classification System (NAPCS)

- Businesses classify their revenue into NAPCS codes based on specific products and services.
- Similar to NAICS, respondents choose from a prelist or provide a write-in.

Source: 2017 Economic Census
• Goal: Develop ML applications and integrate them into the electronic reporting instrument for the 2022 Economic Census
What is BEACON?

• **Business Establishment Automated Classification of NAICS**

• A machine learning tool developed by the Economic Statistical Methods Division (U.S. Census Bureau) to classify NAICS for establishments based on a write-in business description

1. Respondent provides write-in description
2. Text is outputted to BEACON API
3. API returns most relevant NAICS codes to respondent
BEACON: Goals

• Assist respondents in self-designating their NAICS codes

• Improve accuracy of self-designated NAICS codes

• Reduce manual coding of write-ins
BEACON: Training Data

- Historic write-in responses to the Economic Census (EC)
- Frequent write-in text that was autocoded during 2017 EC
- Business descriptions from IRS SS-4 forms
- Classification Analytical Processing System (CAPS) items
- Harmonized System commodity description

<table>
<thead>
<tr>
<th>Business Description Text</th>
<th>NAICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a car dealership.</td>
<td>441110</td>
</tr>
<tr>
<td>R&amp;D lab – medical/health</td>
<td>541715</td>
</tr>
<tr>
<td>we mainly repair furniture, some sales</td>
<td>811420</td>
</tr>
</tbody>
</table>
BEACON: Methodology

• Text cleaning
  • Remove common words and phrases (e.g., “the”, “has”, “for instance”)
  • Correct common misspellings

• Dictionary
  • Words and word combinations that BEACON recognizes
    • Words are cleaned, stemmed, and meet minimum frequency requirements
  • Associations between words and NAICS codes in the training data
    • “tutor” is highly associated with NAICS 611691 – Exam Preparation and Tutoring
    • “retail” occurs in many NAICS codes and is therefore less predictive
BEACON: Methodology

• Three separate information retrieval models

• Models are applied hierarchically
  • First: 2-digit (sector level)
  • Then: 6-digit (industry level)
BEACON: Methodology

• Purity Weights
  • The NAICS distributions of the various words/combs are averaged using “purity weights” that give more weight to the NAICS distributions of words/combs that are more pure/predictive
  • The purity weight is a function of the maximum proportion

• Relevance scores
  • Range in value between 0 and 100
  • Reflect how confident BEACON is that the NAICS code is correct
BEACON: Results

- From these scores, BEACON returns a ranked list of NAICS codes at the 6-digit level.
What is SINCT?

• **Smart Instrument NAPCS Classification Tool**

• Developed by Economy-Wide Statistics Division

• Two distinct versions
  • SINCT 1.0: TF-IDF
  • SINCT 2.0: Doc2Vec

Source: istockphoto.com
SINCT: Training Data

- Reclassified write-ins from the 2012 and 2017 Economic Census
- Classification Analytical Processing System (CAPS) items
- Subject matter expert examples
- NAPCS title file

<table>
<thead>
<tr>
<th>NAICS</th>
<th>Search Term</th>
<th>NAPCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>423860</td>
<td>Airplane</td>
<td>4002000003</td>
</tr>
<tr>
<td>481111</td>
<td>Airplane</td>
<td>7003075000</td>
</tr>
</tbody>
</table>
SINCT 1.0: Methodology

• Term frequency-inverse document frequency (TF-IDF)
  • Compute the frequency of a word or phrase within a document, and then adjust that calculation to account for common words

1. Compute TF-IDF scores for every word in training data
2. Compare the search term with every matching word in the training data
3. Add the TF-IDF scores for each word, and return the ten highest-scored NAPCS codes
SINCT 2.0: Methodology

• Doc2Vec machine learning model
  • Neural network model
  • Represent words, phrases, sentences, or paragraphs as a vector

Text cleaning and spell check

Convert the search string into a numerical representation

Use cosine similarity scores to identify the ten closest training examples, and return those NAPCS codes
SINCT: Results

- SINCT returns the ten highest-scored NAPCS codes to respondents
- Respondents can select from those results, perform a new search, or leave their search term as a write-in
2021 Industry Classification Report Field Test

• The 2021 Economic Census Industry Classification Report (Refile) was repurposed as a field test for BEACON and SINCT

• Approximately 37,000 establishments
  • 12,000 truth deck
  • 25,000 non-truth deck
Field Test: Results

• BEACON
  • Returned correct NAICS code 90% of the time
  • Respondents selected it 83% of the time

• SINCT
  • Returned correct NAPCS code 74% of the time
  • Respondents only selected it 50% of the time
Lessons Learned

• BEACON performed as expected
  • Speed and concurrent request requirements were met
  • 90% accuracy rate

• SINCT performance was lower than expected
  • 6% timeout rate
  • 74% accuracy rate
  • Solution: overhaul the model, improving speed and accuracy
Lessons Learned

• Respondents often select “none of these”, even if the correct result is presented
  • Solution: update wording and instructions for main mailing

• Additional training data
  • Added 7,000 examples each to BEACON and SINCT’s training data
Conclusions

BEACON and SINCT should significantly reduce the number of unclassified write-ins in the 2022 Economic Census

Analysts will spend less time and will code a higher percentage of write-ins

Data quality will be improved
Thank You!

Emily Wiley
Emily.L.Wiley@census.gov

Daniel Whitehead
Daniel.Whitehead@census.gov