Using Natural Language Processing to improve the Commodity Flow Survey

FedCASIC– April 2019

Christian Moscardi
Commodity Flow Survey

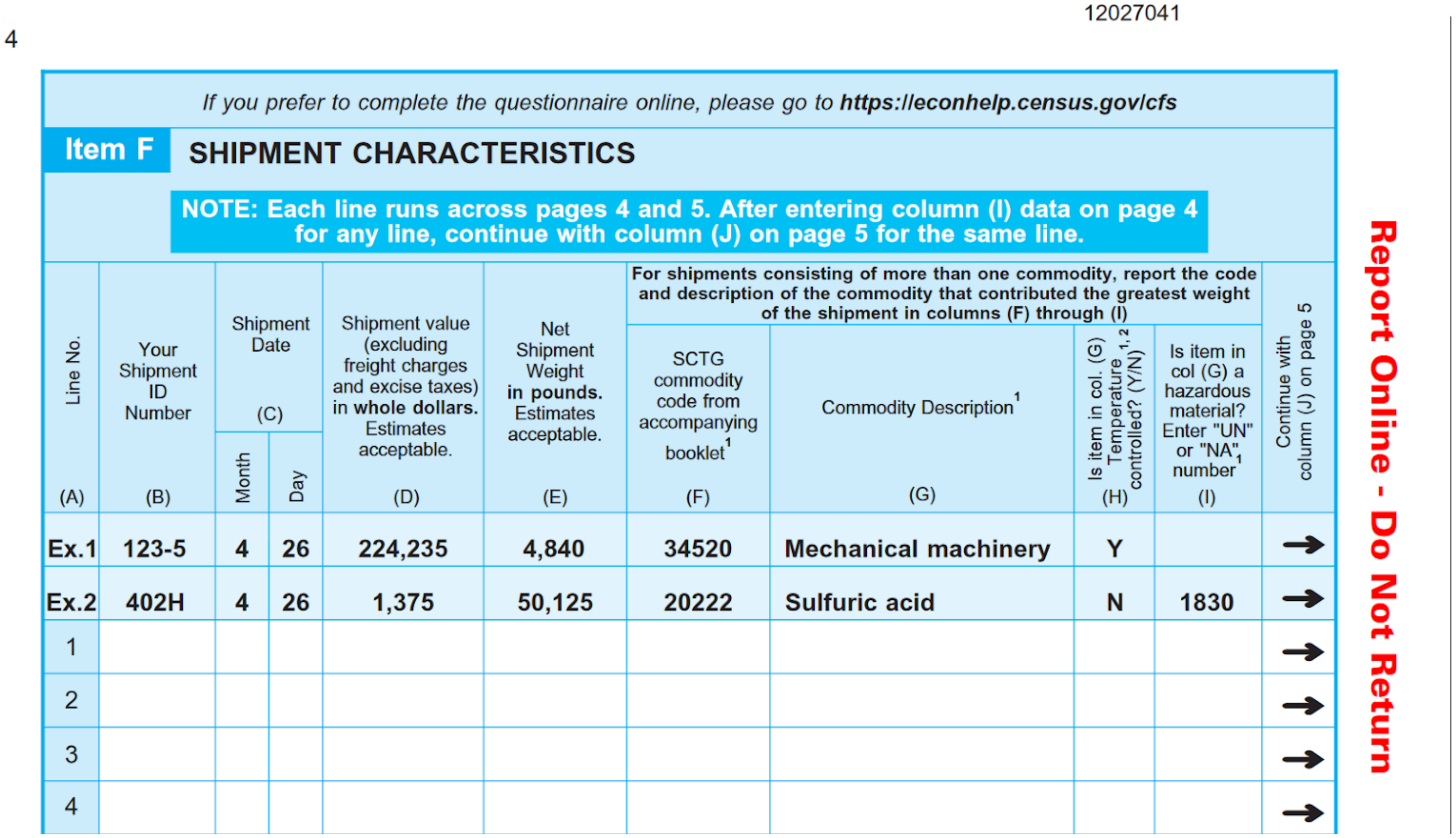
Mehdi Hashemipour, PhD
Bureau of Transportation Statistics

United States™ Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

United States Department of Transportation
Bureau of Transportation Statistics

Overview

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2017, 2022)
- Respondents provide sampling of shipments from each quarter



4

12027041

*If you prefer to complete the questionnaire online, please go to **https://econhelp.census.gov/cfs***

Item F **SHIPMENT CHARACTERISTICS**

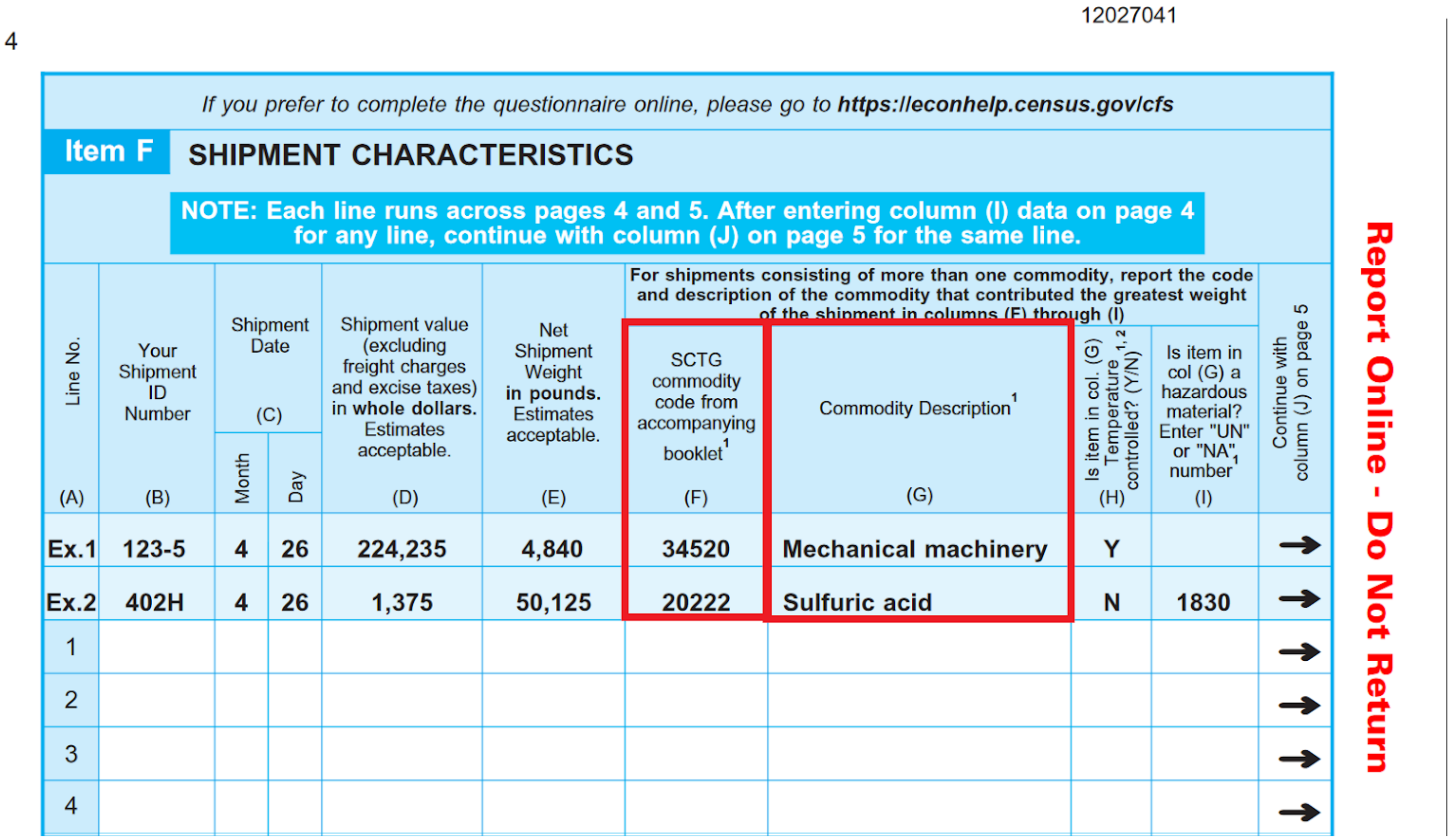
NOTE: Each line runs across pages 4 and 5. After entering column (I) data on page 4 for any line, continue with column (J) on page 5 for the same line.

Line No.	Your Shipment ID Number	Shipment Date (C)		Shipment value (excluding freight charges and excise taxes) in **whole dollars.** Estimates acceptable.	Net Shipment Weight **in pounds.** Estimates acceptable.	For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I)				Continue with column (J) on page 5
(A)	(B)	Month	Day	(D)	(E)	SCTG commodity code from accompanying booklet[1] (F)	Commodity Description[1] (G)	Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H)	Is item in col (G) a hazardous material? Enter "UN" or "NA" number[1] (I)	
Ex.1	**123-5**	**4**	**26**	**224,235**	**4,840**	**34520**	**Mechanical machinery**	**Y**		→
Ex.2	**402H**	**4**	**26**	**1,375**	**50,125**	**20222**	**Sulfuric acid**	**N**	**1830**	→
1										→
2										→
3										→
4										→

Report Online - Do Not Return

Overview

- Commodity Flow Survey
- Commissioned by BTS
- Conducted every 5 years (2017, 2022)
- Respondents provide sampling of shipments from each quarter



4

12027041

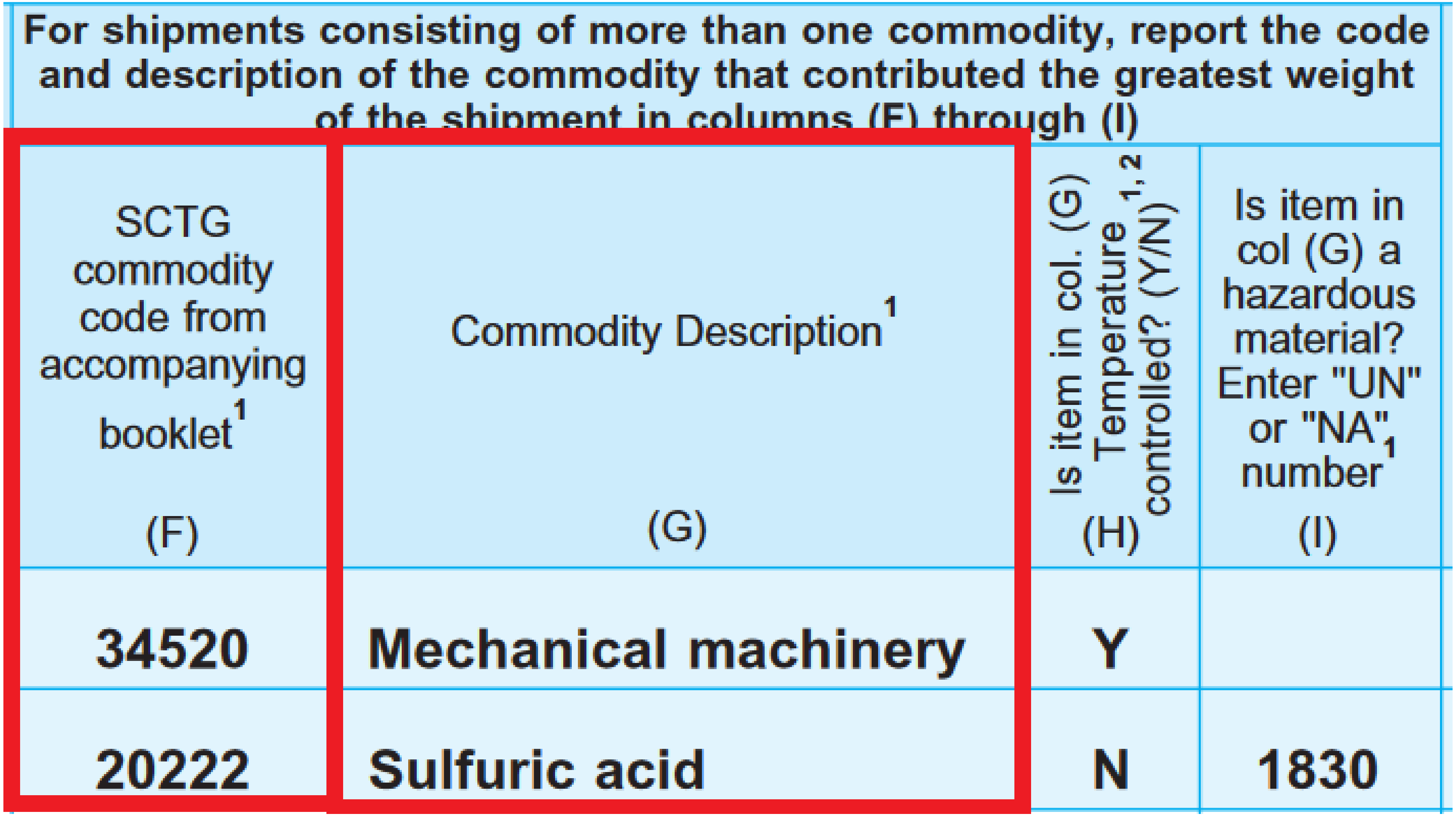
*If you prefer to complete the questionnaire online, please go to **https://econhelp.census.gov/cfs***

Item F **SHIPMENT CHARACTERISTICS**

NOTE: Each line runs across pages 4 and 5. After entering column (I) data on page 4 for any line, continue with column (J) on page 5 for the same line.

Line No. (A)	Your Shipment ID Number (B)	Shipment Date (C)		Shipment value (excluding freight charges and excise taxes) in **whole dollars.** Estimates acceptable. (D)	Net Shipment Weight **in pounds.** Estimates acceptable. (E)	**For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I)**				Continue with column (J) on page 5
		Month	Day			SCTG commodity code from accompanying booklet[1] (F)	Commodity Description[1] (G)	Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H)	Is item in col (G) a hazardous material? Enter "UN" or "NA"[1] number (I)	
Ex.1	**123-5**	**4**	**26**	**224,235**	**4,840**	**34520**	**Mechanical machinery**	**Y**		→
Ex.2	**402H**	**4**	**26**	**1,375**	**50,125**	**20222**	**Sulfuric acid**	**N**	**1830**	→
1										→
2										→
3										→
4										→

Report Online - Do Not Return



For shipments consisting of more than one commodity, report the code and description of the commodity that contributed the greatest weight of the shipment in columns (F) through (I)

SCTG commodity code from accompanying booklet[1] (F)	Commodity Description[1] (G)	Is item in col. (G) Temperature controlled? (Y/N)[1,2] (H)	Is item in col (G) a hazardous material? Enter "UN" or "NA", number[1] (I)
34520	**Mechanical machinery**	**Y**	
20222	**Sulfuric acid**	**N**	**1830**

Overview

- ITEM G - Other Clarifying Information

"Pulling this information was a huge spend of time and resources.“

"Just glad this is over!!"

Overview

Using Machine Learning, can we automate the assignment of SCTG codes to shipment records?

(Yes.)

Initial Model

- Preprocessing
    1. “Throw out” SCTG 40999, 43999
        - These are miscellaneous SCTG codes
    2. Spell-check, stem, de-duplicate
    3. Left with ~400,000 unique training records
- Feature engineering
    1. “Bag-of-words” + TF-IDF scores
- Modelling
    - Logistic Regression, “elastic net” regularization
    - Cross-validate, hold out test set, etc.

28 STEEEL BEAM,S

28 STEEEL BEAM S

STEEEL BEAMS

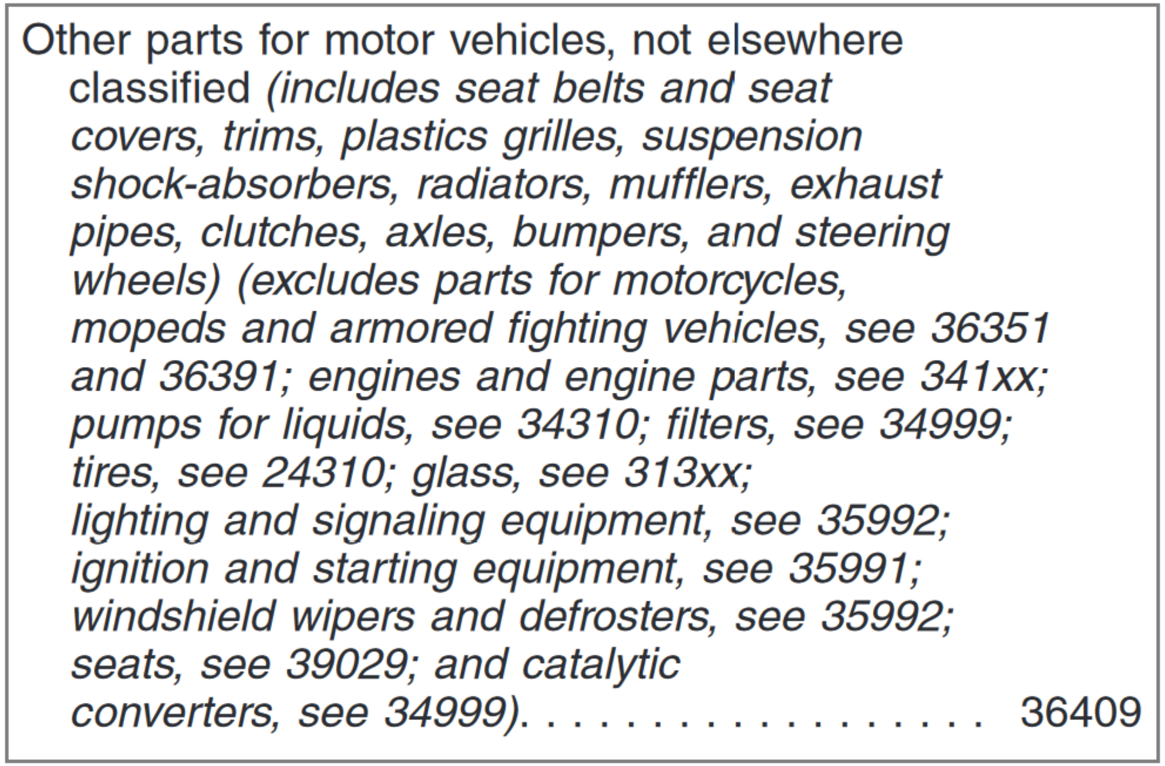
steeel beams

steel beams

steel beam

Further Investigation

- Initial results: ~50% “accuracy”
  - What does that mean?
- Should we use a more complex pipeline?
- Aside from 40999, 43999, ~80 more “other” codes
  - Remove these codes, recovery jumps to 64%



Other parts for motor vehicles, not elsewhere classified *(includes seat belts and seat covers, trims, plastics grilles, suspension shock-absorbers, radiators, mufflers, exhaust pipes, clutches, axles, bumpers, and steering wheels) (excludes parts for motorcycles, mopeds and armored fighting vehicles, see 36351 and 36391; engines and engine parts, see 341xx; pumps for liquids, see 34310; filters, see 34999; tires, see 24310; glass, see 313xx; lighting and signaling equipment, see 35992; ignition and starting equipment, see 35991; windshield wipers and defrosters, see 35992; seats, see 39029; and catalytic converters, see 34999).* 36409

Further Investigation

- E.g. **40994**
  - Sewing and knitting needles (includes for machines) crochet hooks, hook and eye **fasteners**, safety pins, straight pins, buttons, buckles and clasps, tubular and bifurcated rivets, **snap-fasteners**, zippers, and similar notions.

Image courtesy Wikimedia commons

Further Investigation

- **Model's prediction**
- **33310**
  - Nails, screws, bolts, nuts, washers, staples except in strips, and similar **fastening** articles
- What was the NAICS Code?

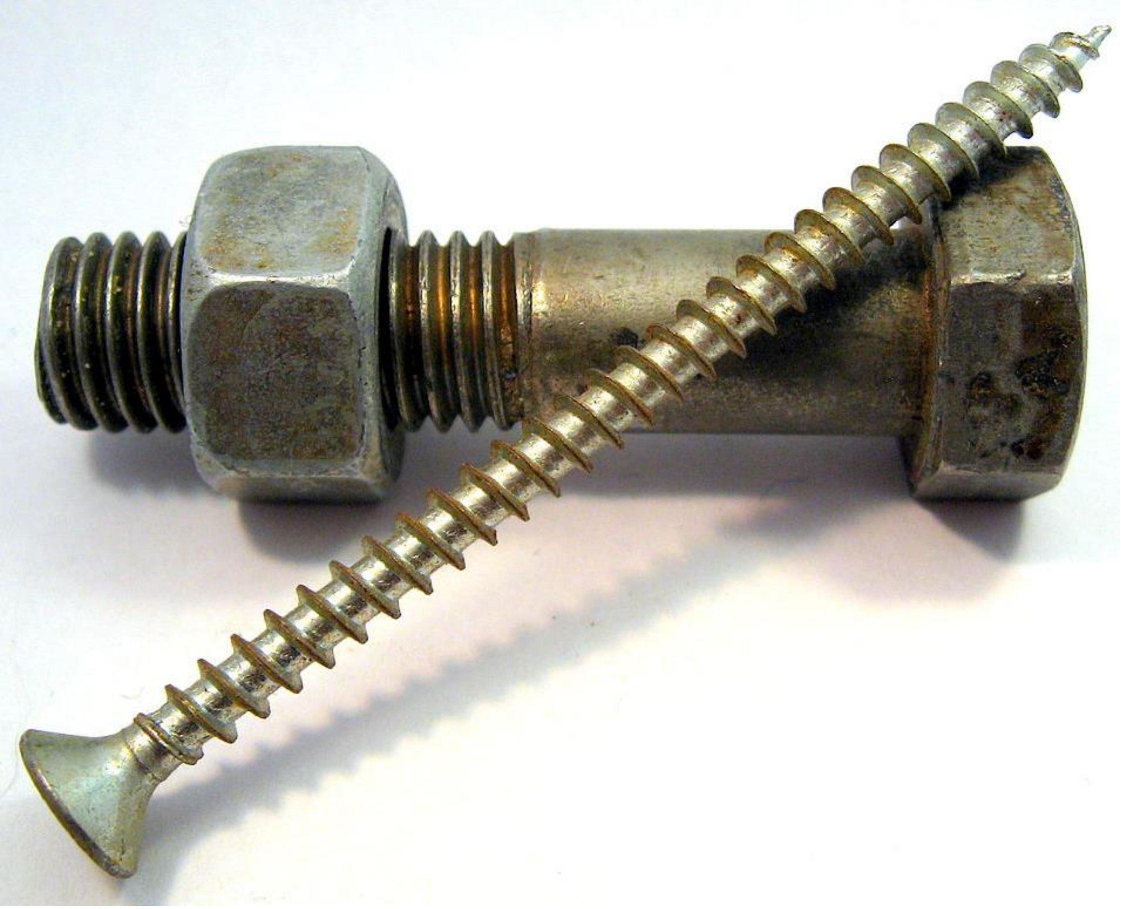
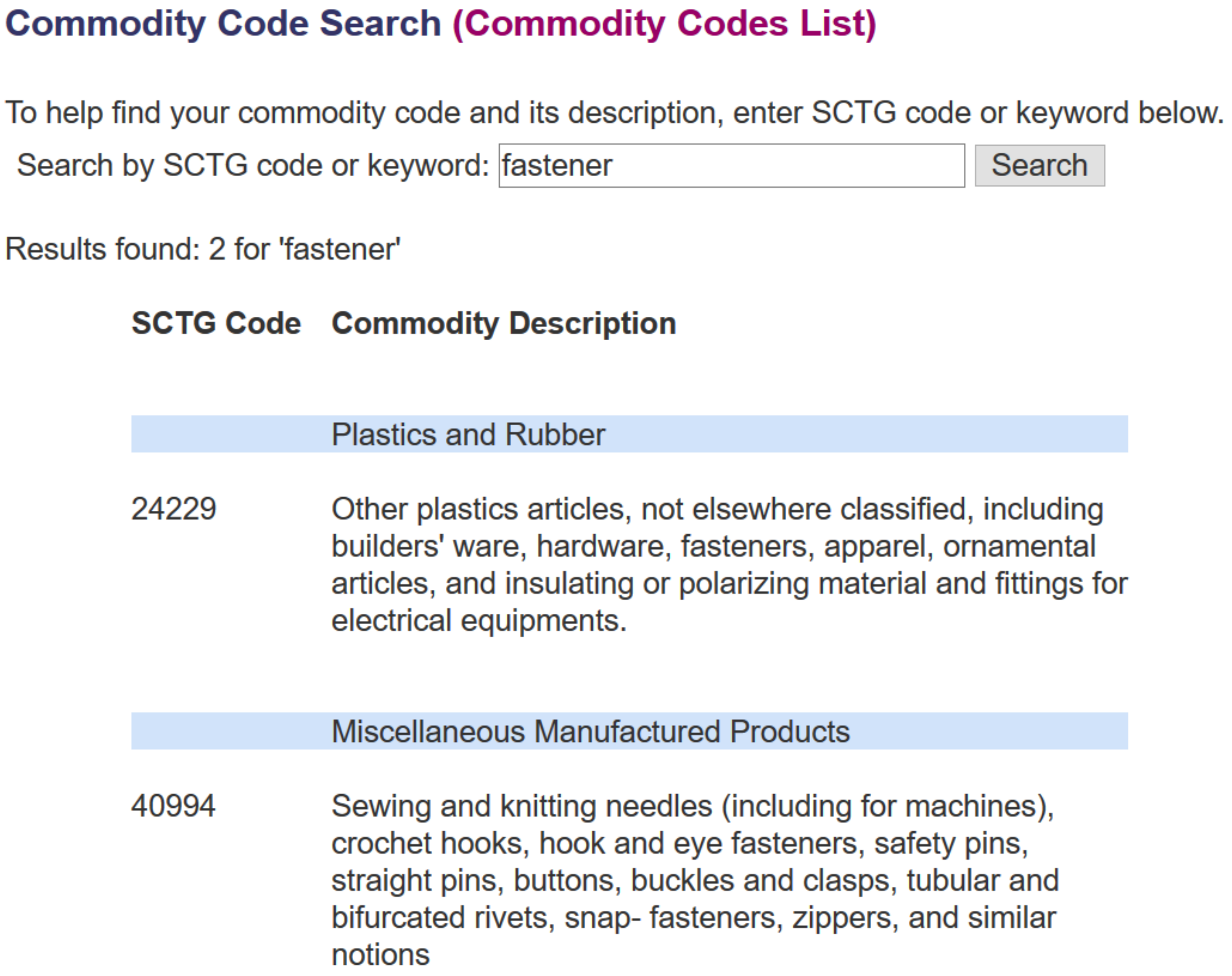


Image courtesy Wikimedia commons

Further Investigation

- Manually validating, about 50% of items labelled 40994 by respondents were miscoded.
- However, the model was getting it right!
- **We can see** the workflow which led to these miscodings



Commodity Code Search (Commodity Codes List)

To help find your commodity code and its description, enter SCTG code or keyword below.

Search by SCTG code or keyword: fastener

Results found: 2 for 'fastener'

SCTG Code	Commodity Description
	Plastics and Rubber
24229	Other plastics articles, not elsewhere classified, including builders' ware, hardware, fasteners, apparel, ornamental articles, and insulating or polarizing material and fittings for electrical equipments.
	Miscellaneous Manufactured Products
40994	Sewing and knitting needles (including for machines), crochet hooks, hook and eye fasteners, safety pins, straight pins, buttons, buckles and clasps, tubular and bifurcated rivets, snap- fasteners, zippers, and similar notions

Let’s Experiment

- Proof-of-concept: ran model on 170,000 unlabeled/invalid records
- 70,000 with probability score above predefined threshold [.5 – 1)
  - Determined by coarse inspection
- CFS Analysts validate a sample of 350 unique records

- Also wanted to determine accuracy in the [0 - .5) threshold
- Took sampling of the other 100,000 unlabeled / invalid records.
  - Model probability ranges [0 - .5)
  - 60 from each range

Results

- Validation: **89% accurate in [.5 -1); 80% in [.4 - .5)**
  - "Accuracy" definition
- Batch-edits have saved **~1000 hours** of manual editing time
- Deploying model as-is would save >$2M of respondent time

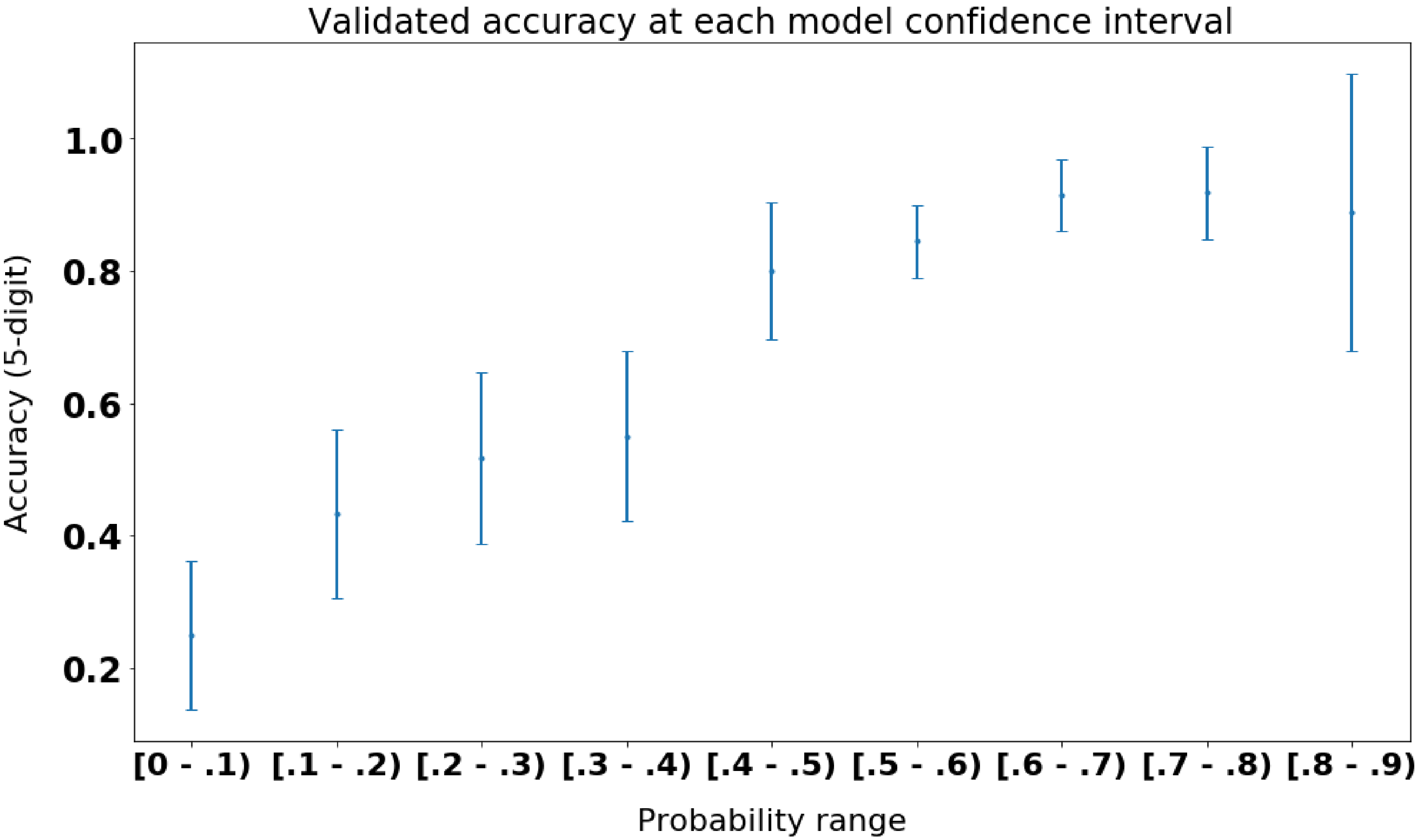
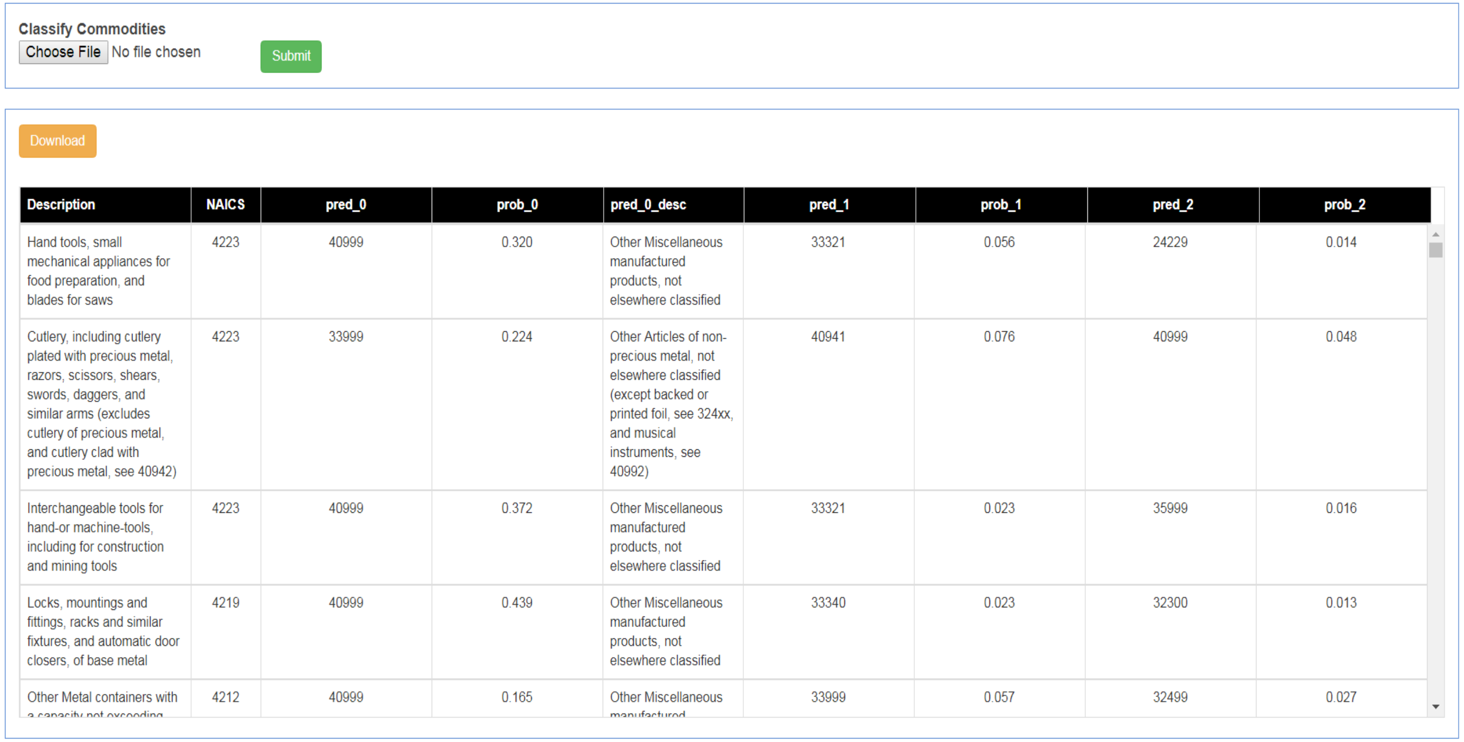


Figure: validation accuracy for each model probability / confidence range. Bars are 95% Bernoulli CI

Batch Cleaner and Batch Classifier Application

Batch Classifier App:

- Get a dataset that contains any number of shipment records
- Run the Machine Learning Model
- Provide An output file contain:
  - ❑ top 3 suggested SCTG5,
  - ❑ and their model confidence probabilities.
  - ❑ standard descriptions for each predicted SCTG,

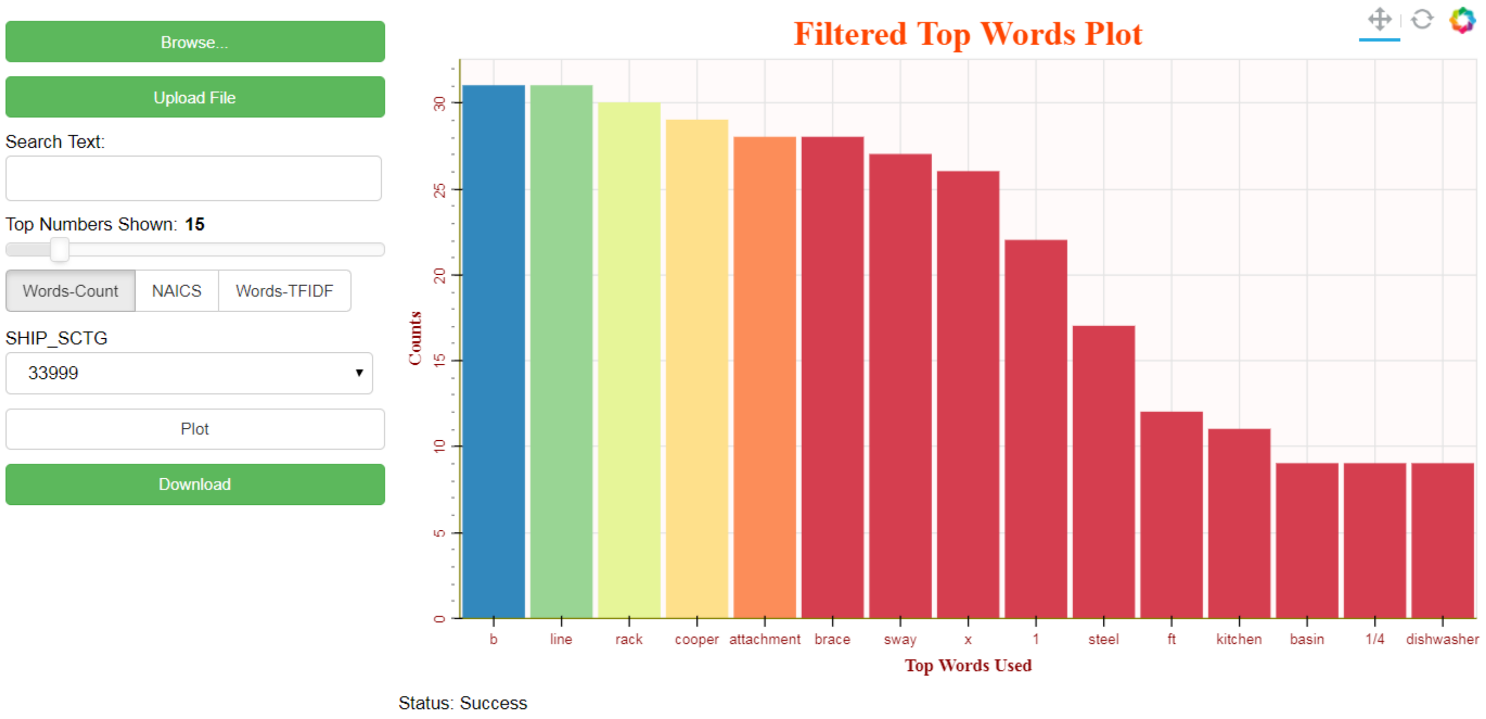


Description	NAICS	pred_0	prob_0	pred_0_desc	pred_1	prob_1	pred_2	prob_2
Hand tools, small mechanical appliances for food preparation, and blades for saws	4223	40999	0.320	Other Miscellaneous manufactured products, not elsewhere classified	33321	0.056	24229	0.014
Cutlery, including cutlery plated with precious metal, razors, scissors, shears, swords, daggers, and similar arms (excludes cutlery of precious metal, and cutlery clad with precious metal, see 40942)	4223	33999	0.224	Other Articles of non-precious metal, not elsewhere classified (except backed or printed foil, see 324xx, and musical instruments, see 40992)	40941	0.076	40999	0.048
Interchangeable tools for hand-or machine-tools, including for construction and mining tools	4223	40999	0.372	Other Miscellaneous manufactured products, not elsewhere classified	33321	0.023	35999	0.016
Locks, mountings and fittings, racks and similar fixtures, and automatic door closers, of base metal	4219	40999	0.439	Other Miscellaneous manufactured products, not elsewhere classified	33340	0.023	32300	0.013
Other Metal containers with a capacity not exceeding	4212	40999	0.165	Other Miscellaneous manufactured	33999	0.057	32499	0.027

SCTG Analysis App

SCTG Analysis Application: Using this tool help users to input any excel or CSV dataset contains Product Descriptions and Label Codes into the app and perform a visual analysis.

For the selected SCTG5, the app interface illustrates plots of Top Words, top NAICS codes and top TF-IDF words (that reflect how important a word is to a document in a collection or corpus).



Search Text:

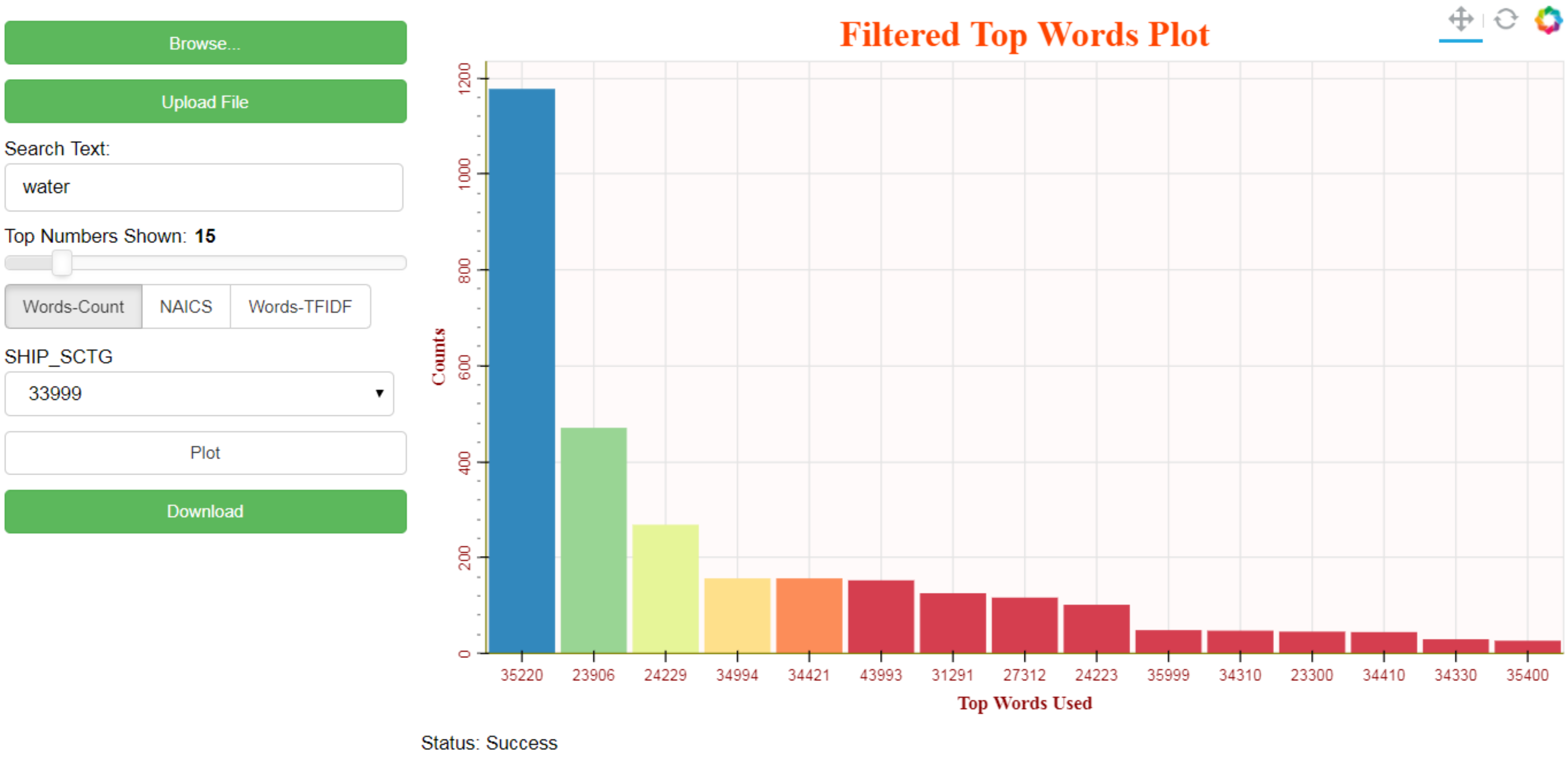
Top Numbers Shown: 15

SHIP_SCTG

33999

Status: Success

SCTG Analysis App cont.



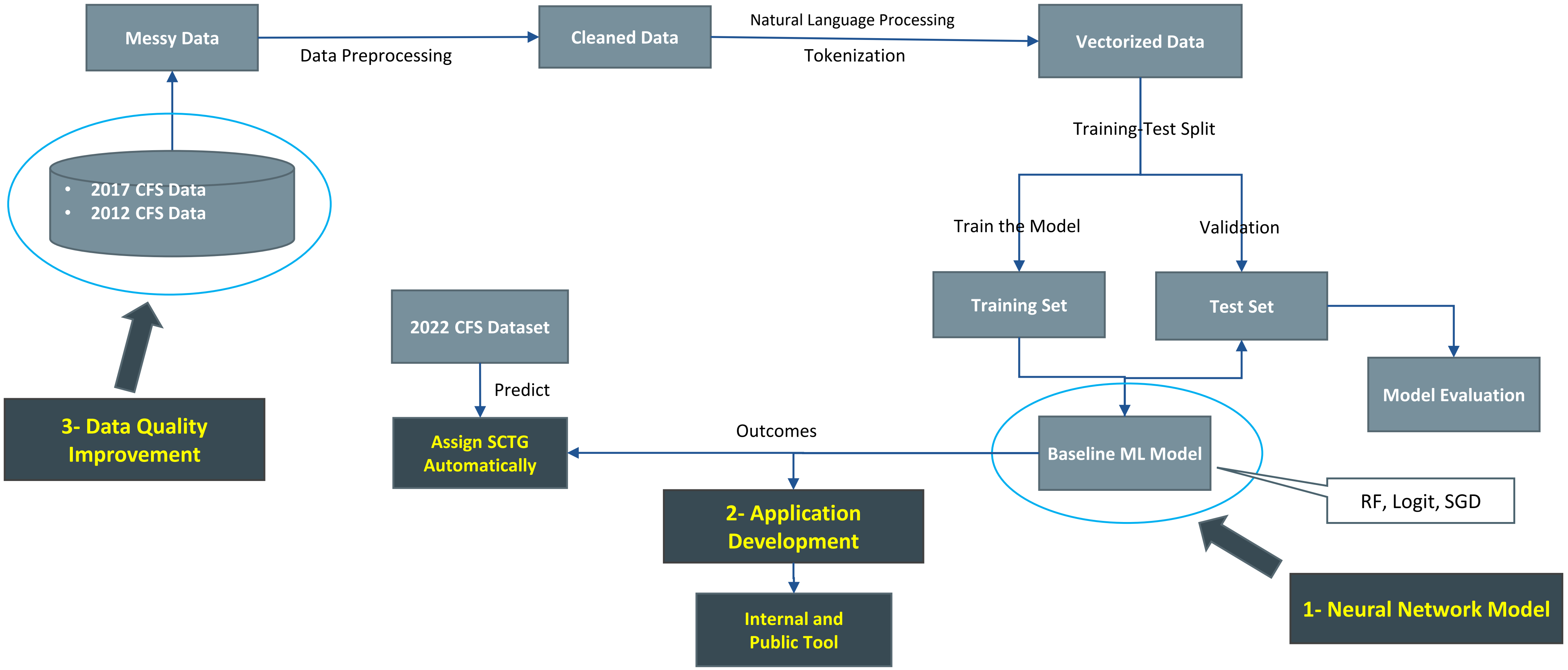
Search Text: water

Top Numbers Shown: **15**

SHIP_SCTG: 33999

Status: Success

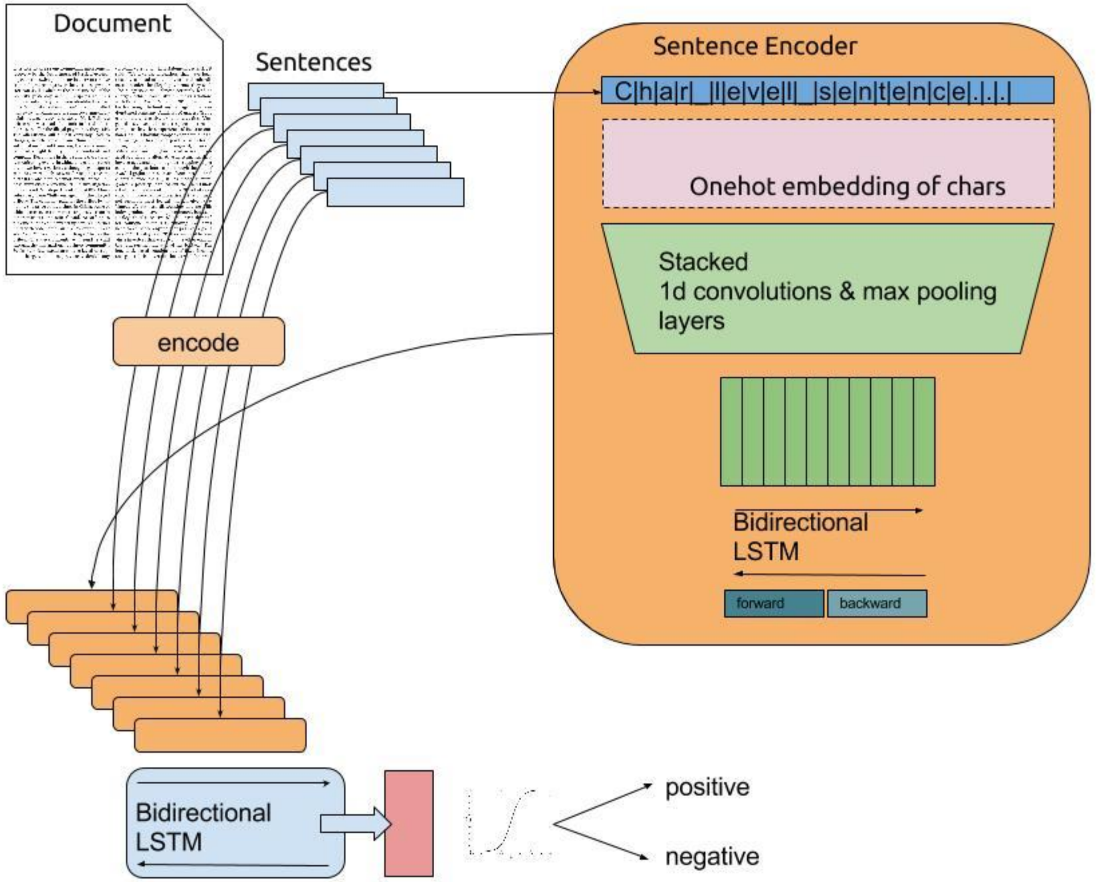
CFS – SCTG Machine Learning Pipeline & Future Works



Neural Network Model - Character Level LSTM

This model is reading characters one by one, to create an embedding of the of a given text/product description. As such our neural network will try to learn that specific sequences of letters form words separated by spaces or other punctuation points.

Our goal is to encode text from character level, so we'll begin by splitting the text into words. Then encode each word to characters. We use a bi-directional LSTM to read word by word and create a complete document encoding.



Data Quality Improvement

Better Training Set == Better Model == Better classification & More Savings

Motivation: Using cleaner data and collecting more product description improve the ML model performance

Objective: Mapping more product descriptions into SCTG codes

- Web Scraping
- Amazon's Mechanical Turk

Web Scraping for CFS SCTG Project

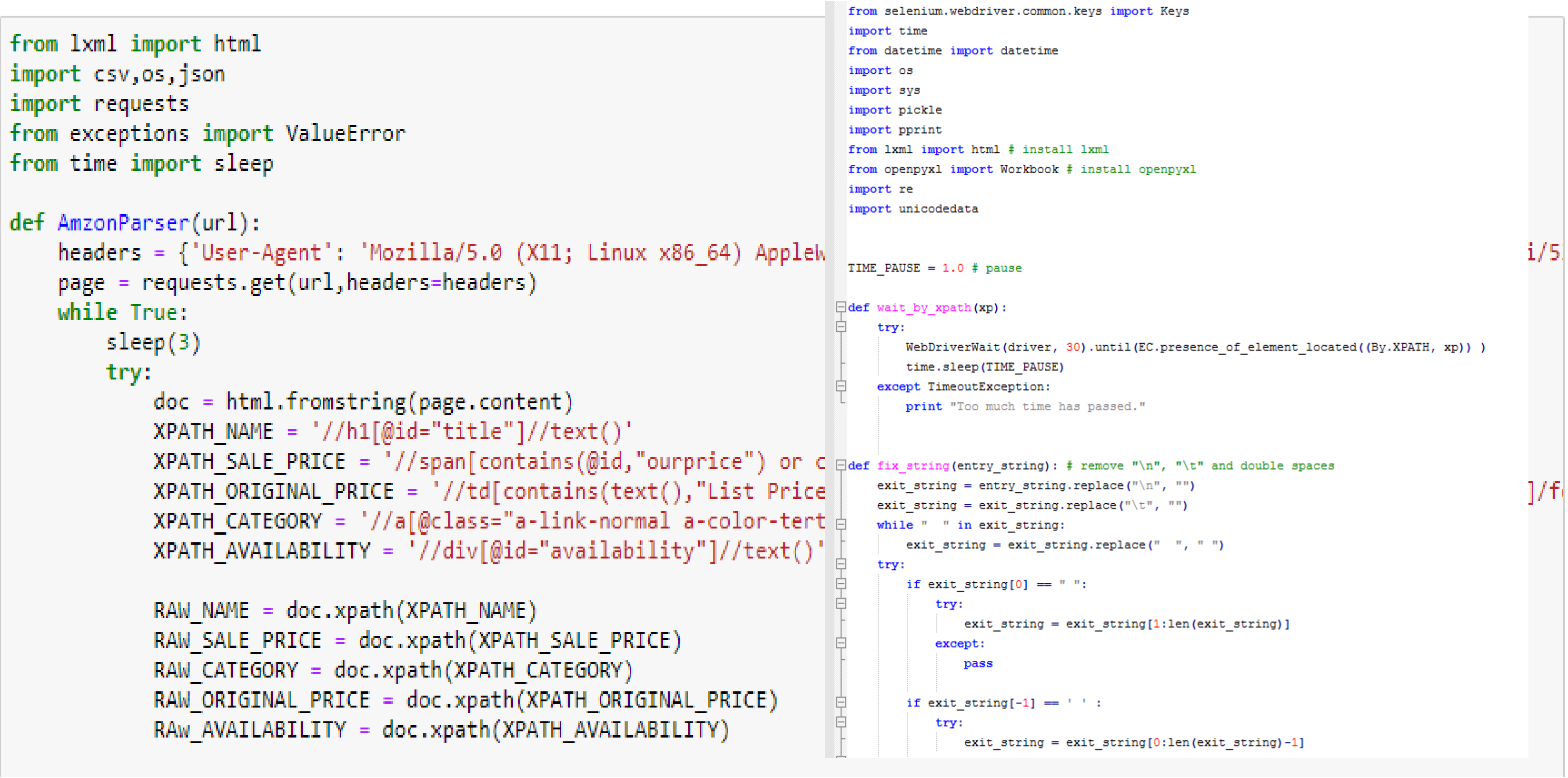
- **Development:**
  - Selected SCTG classes that are not performing well
  - Targeted e-commerce and manufacturers websites to extract the desired product descriptions

    exp: DEWALT, Ferguson, Home Depot, IKEA, Alibaba, etc.
  - Developed Web Crawler Scripts for each targeted ecommerce

Web Scraping for CFS SCTG Project cont.

So far:

Collected ≈ 250,000 product's description for the low performing class of SCTGs in the ML model

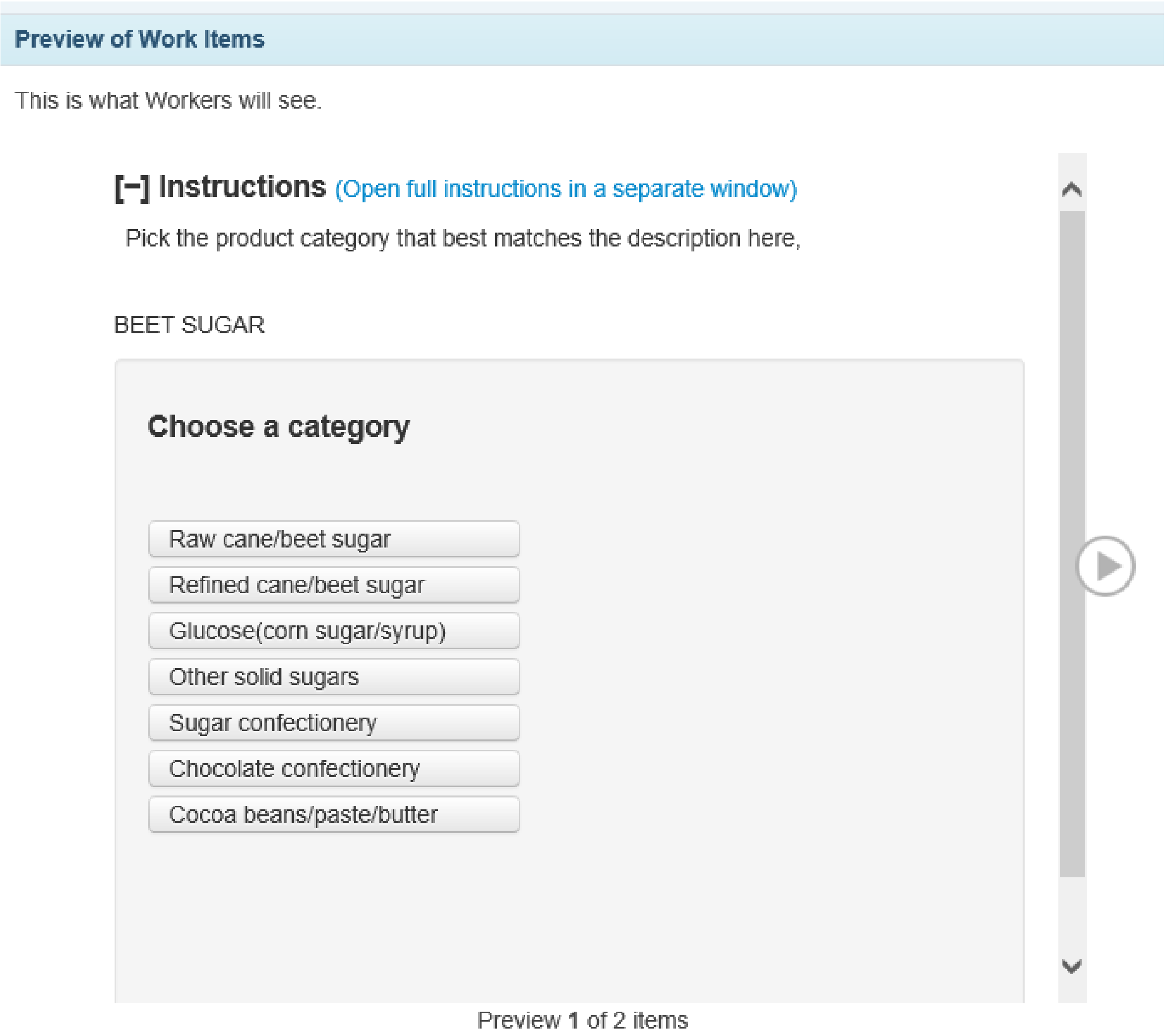


```
from lxml import html
import csv,os,json
import requests
from exceptions import ValueError
from time import sleep

def AmzonParser(url):
    headers = {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleW
    page = requests.get(url,headers=headers)
    while True:
        sleep(3)
        try:
            doc = html.fromstring(page.content)
            XPATH_NAME = '//h1[@id="title"]//text()'
            XPATH_SALE_PRICE = '//span[contains(@id,"ourprice") or c
            XPATH_ORIGINAL_PRICE = '//td[contains(text(),"List Price
            XPATH_CATEGORY = '//a[@class="a-link-normal a-color-tert
            XPATH_AVAILABILITY = '//div[@id="availability"]//text()'

            RAW_NAME = doc.xpath(XPATH_NAME)
            RAW_SALE_PRICE = doc.xpath(XPATH_SALE_PRICE)
            RAW_CATEGORY = doc.xpath(XPATH_CATEGORY)
            RAW_ORIGINAL_PRICE = doc.xpath(XPATH_ORIGINAL_PRICE)
            RAw_AVAILABILITY = doc.xpath(XPATH_AVAILABILITY)
```

```
from selenium.webdriver.common.keys import Keys
import time
from datetime import datetime
import os
import sys
import pickle
import pprint
from lxml import html # install lxml
from openpyxl import Workbook # install openpyxl
import re
import unicodedata


TIME_PAUSE = 1.0 # pause

def wait_by_xpath(xp):
    try:
        WebDriverWait(driver, 30).until(EC.presence_of_element_located((By.XPATH, xp)) )
        time.sleep(TIME_PAUSE)
    except TimeoutException:
        print "Too much time has passed."


def fix_string(entry_string): # remove "\n", "\t" and double spaces
    exit_string = entry_string.replace("\n", "")
    exit_string = exit_string.replace("\t", "")
    while "  " in exit_string:
        exit_string = exit_string.replace("  ", " ")
    try:
        if exit_string[0] == " ":
            try:
                exit_string = exit_string[1:len(exit_string)]
            except:
                pass

        if exit_string[-1] == ' ' :
            try:
                exit_string = exit_string[0:len(exit_string)-1]
```

MTurk - Crowdsourcing

- Labelling NAICS code index file. ~10k records.
- Turkers choose among top 7-10 predictions from model .



Preview of Work Items

This is what Workers will see.

[–] Instructions (Open full instructions in a separate window)

Pick the product category that best matches the description here,

BEET SUGAR

Choose a category

Preview **1** of 2 items

MTurk - Implementation

- **How do we ensure quality?**
- Gateway task
  - Label 50 “gold standard” records
  - Must be at least 60% accurate on min. 5 records
- “Quadruple-key entry”
  - 4 workers label each record
  - Take a vote
  - Total disagreement? This record needs manual investigation.
- Continuous Validation
  - Inter-rater agreement
  - Include more gold standard during actual task

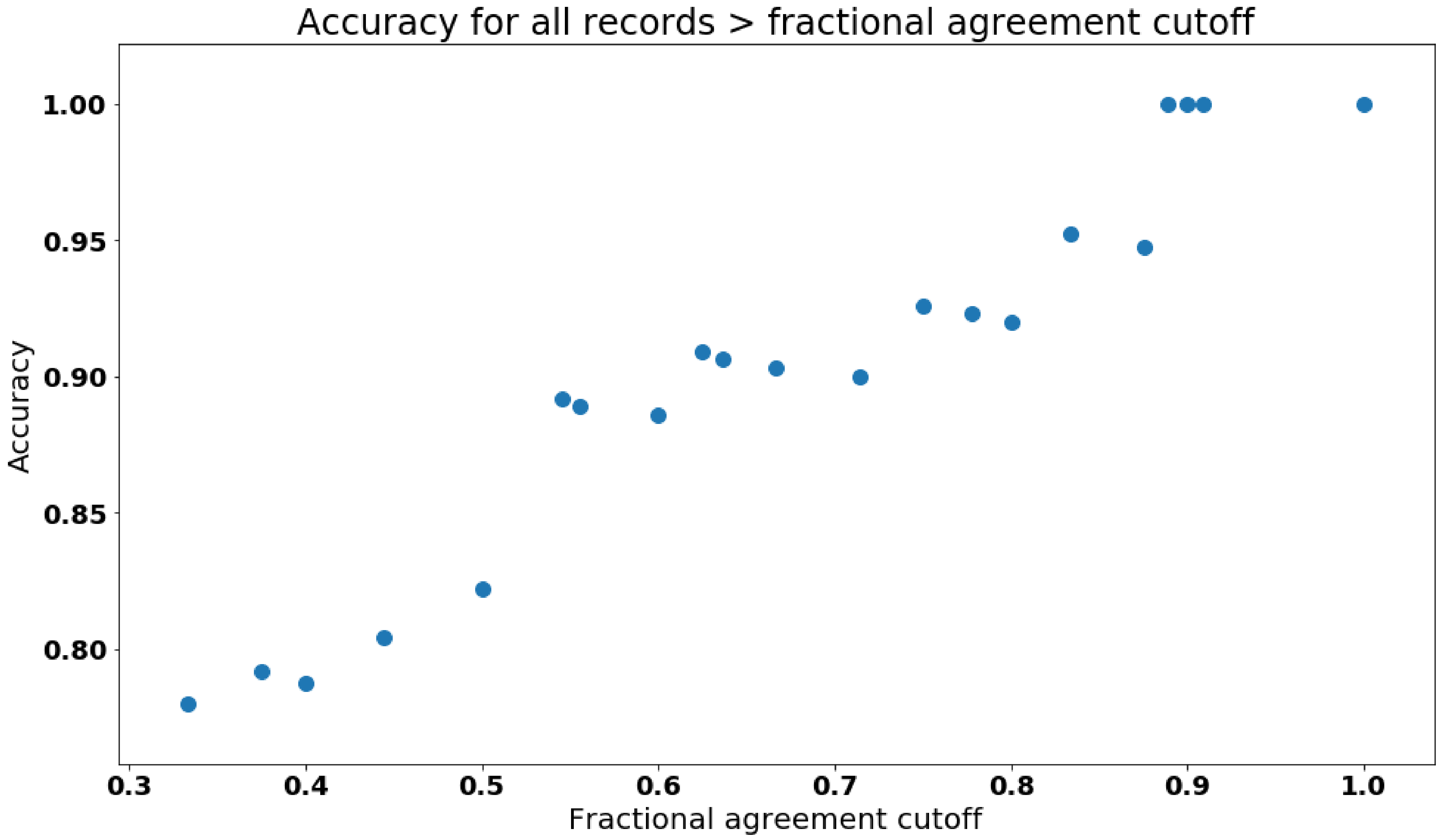


Figure caption in slide notes.

Thank you!

- **Christian**: Christian.L.Moscardi@census.gov
- **Mehdi:** m.hashemipour.ctr@dot.gov