CES DIARY AUTOCODER

Division of Consumer Expenditure Surveys

Janel Brattland — Chief, Branch of Production and Control

CIVIC DIGITAL FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

MICHELL LI
Carnegie Mellon University
Management Information Systems

MOTIVATION

The Bureau of Labor Statistics wants to automatically assign item codes to the Diary portion of the Consumer Expenditure Survey.

The process is currently labor intensive and expensive.

The existing autocoder is a rule based system.

Creating a new autocoder using machine learning can reduce costs and improve accuracy.

CONTRIBUTIONS

Models

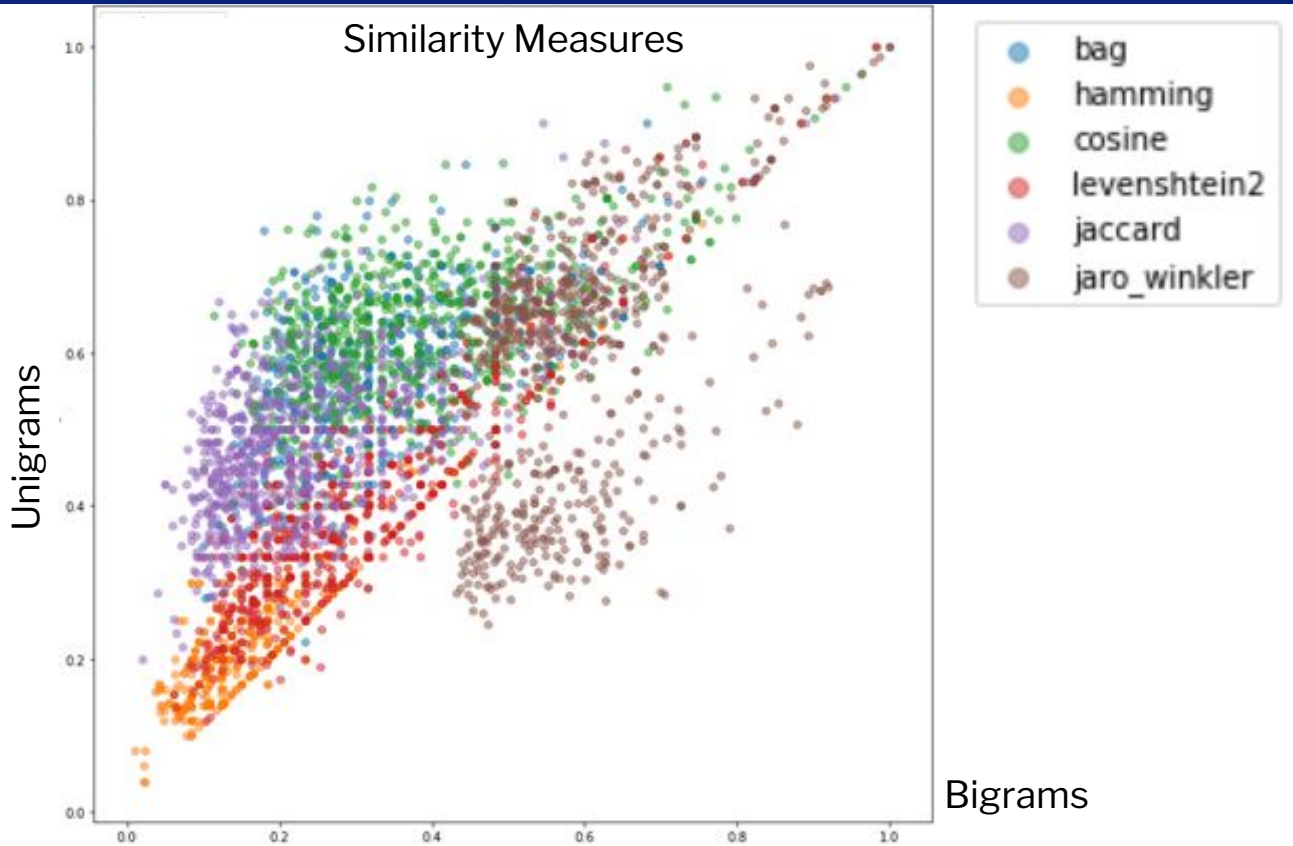
- Four Models
  - ECLO, EFDB, EOTH - Random Forest, L1 Regularization
  - EMLS - Decision Tree, Random Forest

Data and Analysis

- Spell Checker created using Levenshtein, Jaro Winkler and QWERTY distance
- Module for imputing Unclassifiable diary items
- Analytics Dashboard UI using Dash

CIVIC DIGITAL FELLOWSHIP

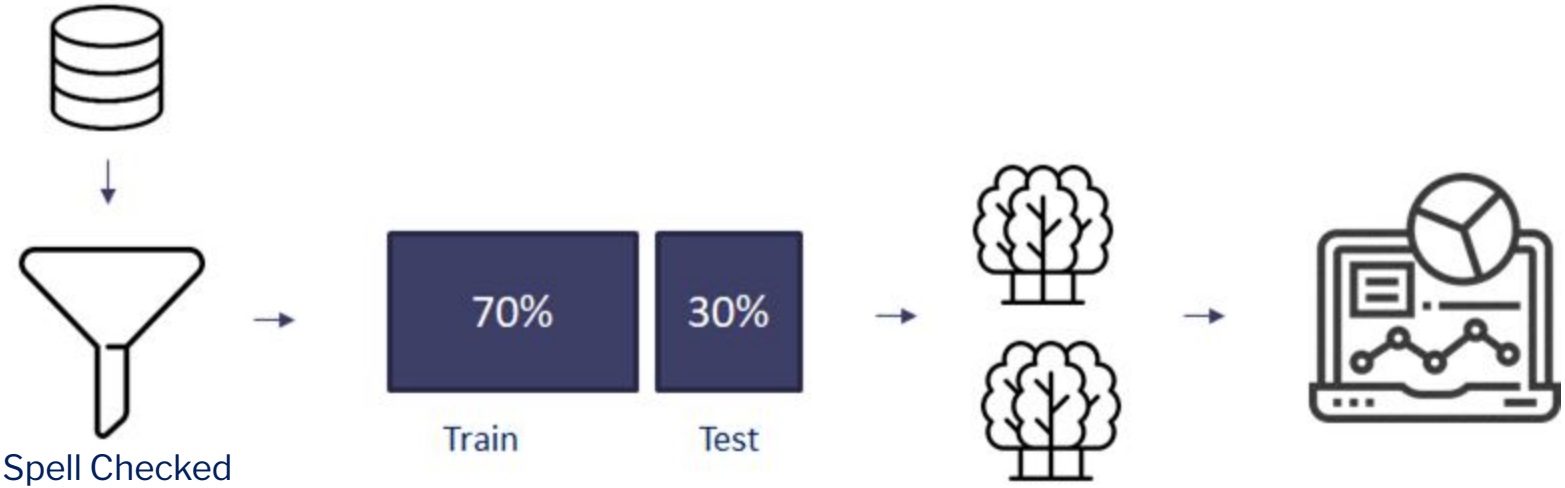
SPELL CHECKER



CIVIC DIGITAL FELLOWSHIP

MODEL ARCHITECTURE - ECLO

3 years SAS data



Spell Checked
TfidfVectorizer ngrams = (2,3)

n_estimators = 400
max_depth = 100
criterion = entropy

CIVIC DIGITAL FELLOWSHIP

RESULTS

Model	Precision	Recall	F1 Score	Accuracy
Clothing	0.91	0.9	0.9	0.9
Food and Beverages	0.95	0.94	0.94	0.94
Meals	0.98	0.98	0.98	0.98
Other	0.85	0.83	0.83	0.83

Calculated using weighted macro averages

CIVIC DIGITAL FELLOWSHIP