Using Open-Ended Text to Evaluate “Other” Disposition Codes

Victoria R. Narine

Josh Langeland

Office of Survey Methods Research

Bureau of Labor Statistics

FCSM Research & Policy Conference 2024

Wednesday, October 23, 2024

Occupational Employment and Wage Statistics (OEWS)

Produces wage and employment estimates for 800+ occupations

Cooperative effort between the BLS and State Workforce Agencies (SWA)

2 semiannual panels in May and November

Establishments in panel contacted via mail, email, or telephone; responses collected via mail, online, email, or telephone

Project Motivation

- Conducted an experiment on contact strategies in Massachusetts using OEWS collections
- Obtained contact log with disposition codes
  - **Disposition codes:** labels attributed to type of contact made with establishment

1. Collected data
2. Contact refinement
3. Data clarification
4. Left message
5. Promised data
6. Refusal
7. Submitted online contact form
8. Other

Project Motivation

- Conducted an experiment on contact strategies in Massachusetts using OEWS collections
- Obtained contact log with disposition codes
  - **Disposition codes:** labels attributed to type of contact made with establishment

1. Collected data
2. Contact refinement
3. Data clarification
4. Left message
5. Promised data
6. Refusal
7. Submitted online contact form
8. Other

Motivation

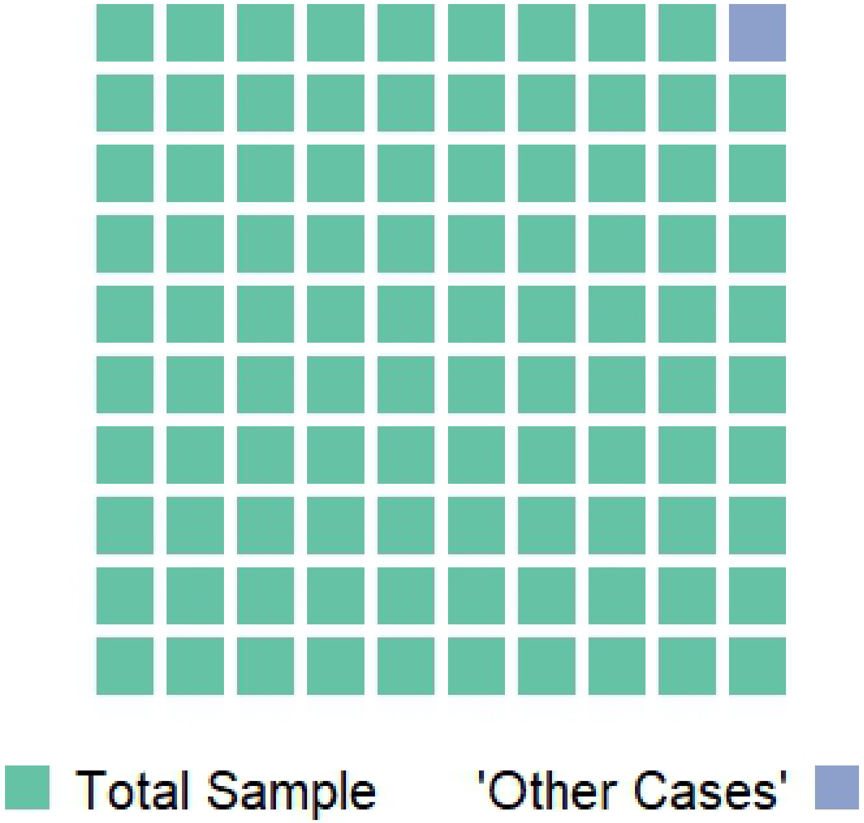
- If disposition code is “Other,” data collectors asked to write short, explanatory notes in contact log
  - Purpose of notes in contact log is to provide additional context to data collection
- Notes are unstructured, individualized open-text
  - ~1 or 2 sentences

Current Research

- **Research Question:** Can contact log notes for “Other” disposition codes be leveraged to determine whether the standard disposition codes are sufficient or in need of improvement?
    1. Can these “Other” cases be coded into the standard disposition codes?
    2. Does the content of the “Other” notes indicate codes need to be revised?

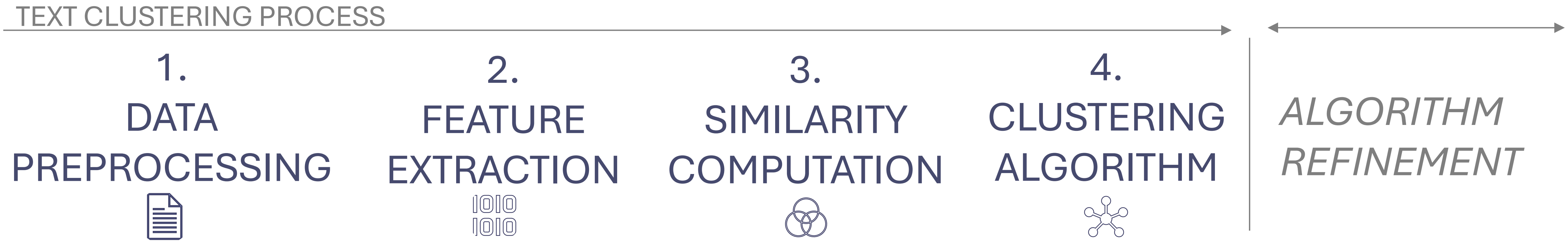
Research Strategy

- Total sample included **30,294** cases from MA
- **222** cases remained after filtering for “Other” disposition codes
  - ~.73% of entire sample



Research Strategy

- Analyzed content of open-ended text using **unsupervised text clustering** to automatically detect patterns in text



TEXT CLUSTERING PROCESS

1. DATA PREPROCESSING

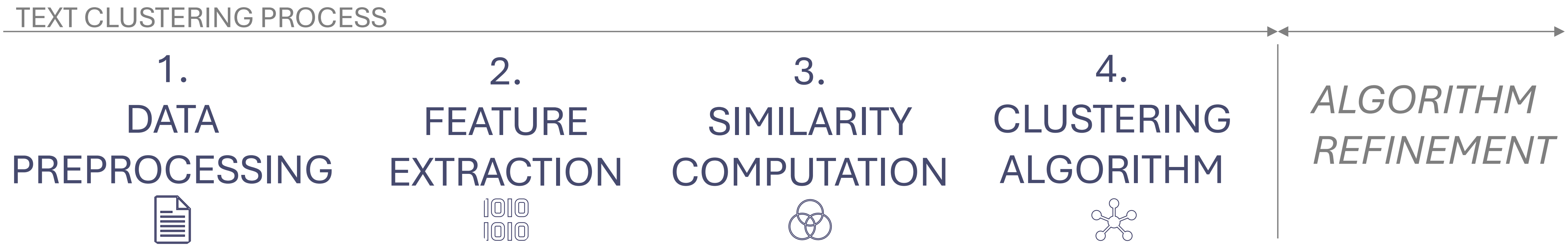
2. FEATURE EXTRACTION

3. SIMILARITY COMPUTATION

4. CLUSTERING ALGORITHM

ALGORITHM REFINEMENT

Research Strategy



TEXT CLUSTERING PROCESS

1. DATA PREPROCESSING

2. FEATURE EXTRACTION

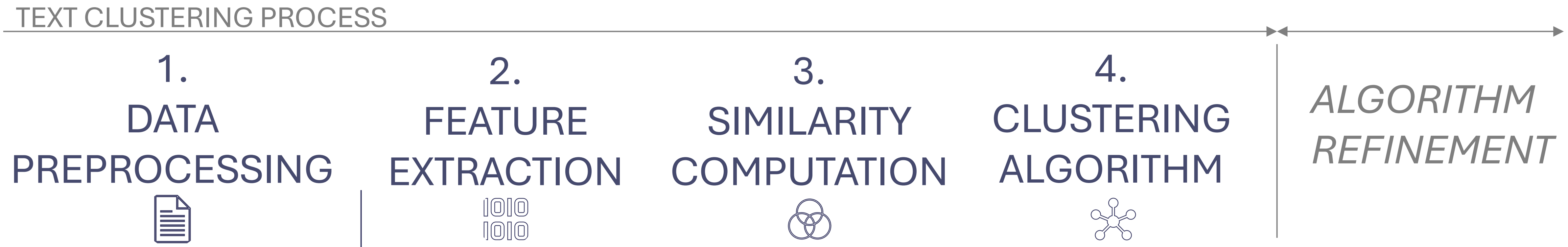
3. SIMILARITY COMPUTATION

4. CLUSTERING ALGORITHM

ALGORITHM REFINEMENT

- Removed stop words, punctuation, and conducted stemming
- Manual preprocessing included removing dates from comments

Research Strategy



TEXT CLUSTERING PROCESS

1. DATA PREPROCESSING

2. FEATURE EXTRACTION

3. SIMILARITY COMPUTATION

4. CLUSTERING ALGORITHM

ALGORITHM REFINEMENT

- Removed stop words, punctuation, and conducted stemming
- Manual preprocessing included removing dates from comments

- Constructed a TF-IDF (Term Frequency – Inverse Document Frequency) matrix

Research Strategy

TEXT CLUSTERING PROCESS

1. DATA PREPROCESSING

- Removed stop words, punctuation, and conducted stemming
- Manual preprocessing included removing dates from comments

2. FEATURE EXTRACTION

- Constructed a TF-IDF (Term Frequency – Inverse Document Frequency) matrix

3. SIMILARITY COMPUTATION

- Applied **cosine similarity** to compare extracted features of log comments and to quantify similarity

4. CLUSTERING ALGORITHM

ALGORITHM REFINEMENT

Research Strategy

TEXT CLUSTERING PROCESS

1. DATA PREPROCESSING

- Removed stop words, punctuation, and conducted stemming
- Manual preprocessing included removing dates from comments

2. FEATURE EXTRACTION

- Constructed a TF-IDF (Term Frequency – Inverse Document Frequency) matrix

3. SIMILARITY COMPUTATION

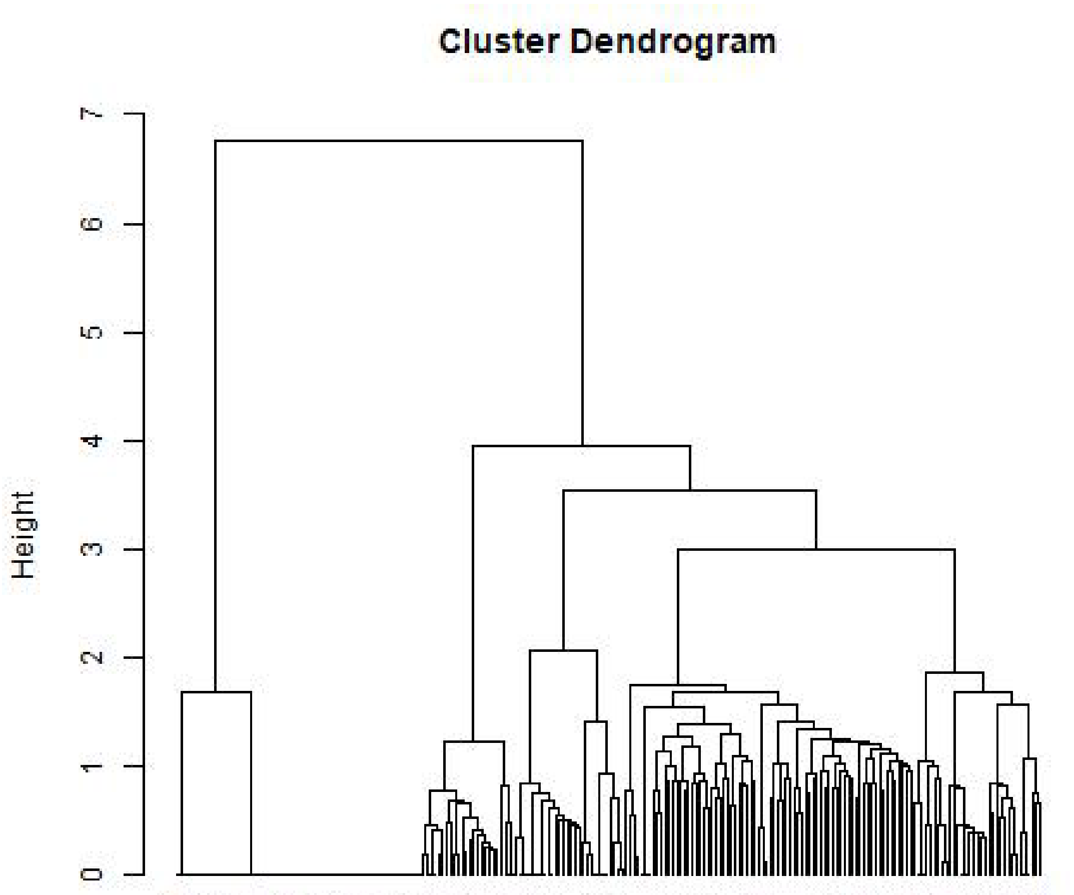
- Applied **cosine similarity** to compare extracted features of log comments and to quantify similarity

4. CLUSTERING ALGORITHM

- Used **hierarchical clustering algorithm** to group comments based on similarity

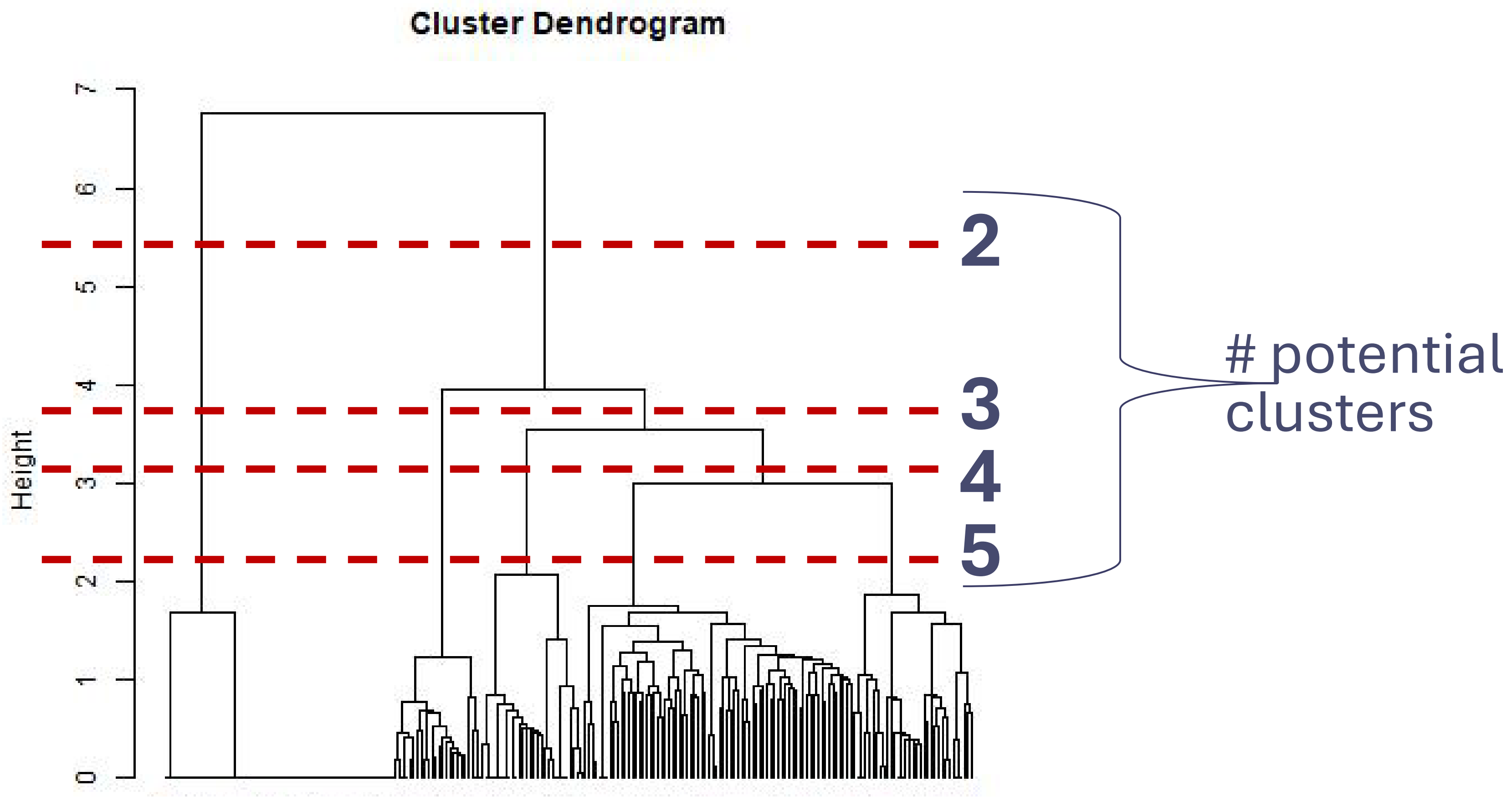
ALGORITHM REFINEMENT

Project Outcomes



- Hierarchical clustering visualized as a dendrogram, which shows the hierarchical relationships between the comments

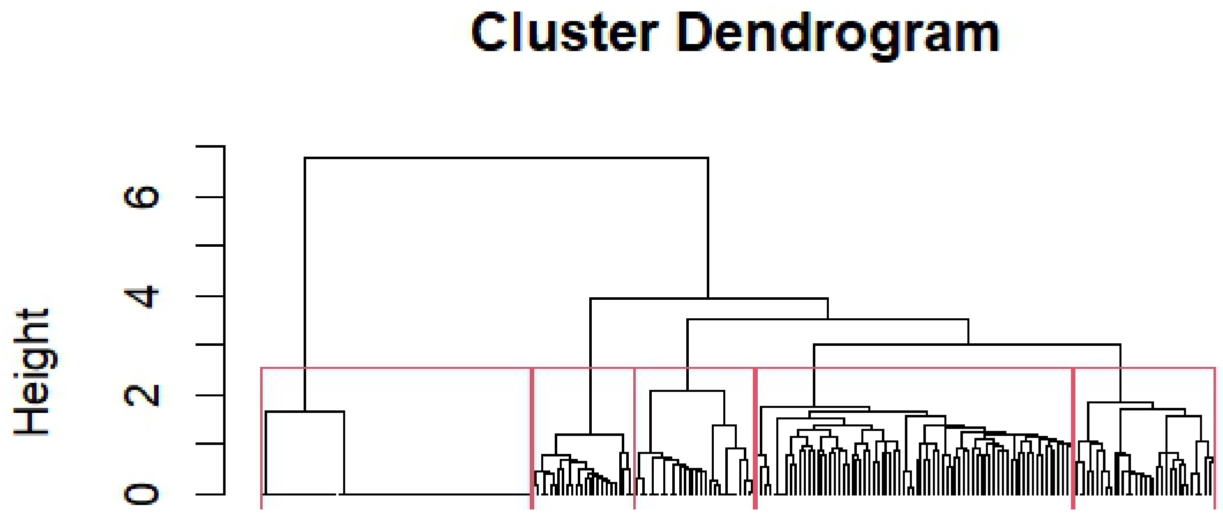
Project Outcomes



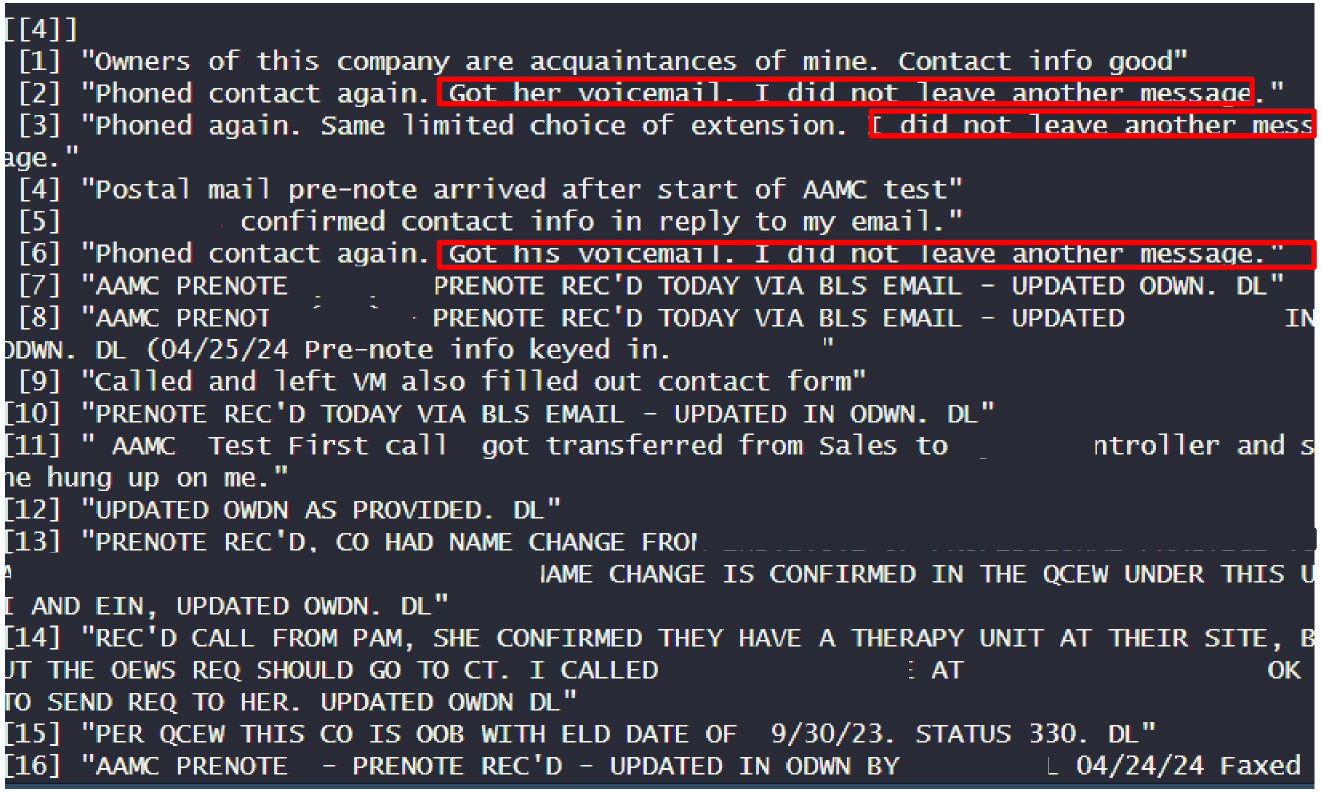
- Draw horizontal line and count number of times it intersects with vertical line
- # intersections = number of clusters

Project Outcomes

- Chose 5 clusters as a starting point
- Manually inspected comments in clusters



Project Outcomes



```
[[4]]
 [1] "Owners of this company are acquaintances of mine. Contact info good"
 [2] "Phoned contact again. Got her voicemail. I did not leave another message."
 [3] "Phoned again. Same limited choice of extension. I did not leave another mess
age."
 [4] "Postal mail pre-note arrived after start of AAMC test"
 [5] [redacted] confirmed contact info in reply to my email."
 [6] "Phoned contact again. Got his voicemail. I did not leave another message."
 [7] "AAMC PRENOTE [redacted] PRENOTE REC'D TODAY VIA BLS EMAIL - UPDATED ODWN. DL"
 [8] "AAMC PRENOT [redacted] PRENOTE REC'D TODAY VIA BLS EMAIL - UPDATED [redacted] IN
ODWN. DL (04/25/24 Pre-note info keyed in. [redacted] "
 [9] "Called and left VM also filled out contact form"
[10] "PRENOTE REC'D TODAY VIA BLS EMAIL - UPDATED IN ODWN. DL"
[11] " AAMC  Test First call  got transferred from Sales to [redacted] ntroller and s
he hung up on me."
[12] "UPDATED OWDN AS PROVIDED. DL"
[13] "PRENOTE REC'D, CO HAD NAME CHANGE FRO [redacted]
[redacted] IAME CHANGE IS CONFIRMED IN THE QCEW UNDER THIS U
I AND EIN, UPDATED OWDN. DL"
[14] "REC'D CALL FROM PAM, SHE CONFIRMED THEY HAVE A THERAPY UNIT AT THEIR SITE, B
UT THE OEWS REQ SHOULD GO TO CT. I CALLED [redacted] AT [redacted] OK
TO SEND REQ TO HER. UPDATED OWDN DL"
[15] "PER QCEW THIS CO IS OOB WITH ELD DATE OF  9/30/23. STATUS 330. DL"
[16] "AAMC PRENOTE  - PRENOTE REC'D - UPDATED IN ODWN BY [redacted] 04/24/24 Faxed
```

*PII redacted

- As a first step in exploratory analysis, clustering algorithm separated out meaningful comments
  - Preliminary evidence that disposition codes could be modified

Discussion

- Algorithm effectively clustered contact log comments based on similarity
  - Because few state analysts in MA, little within cluster variation for some clusters
- Most comments specific to experiment, limiting the amount of meaningful information that could be gleaned
- Depending on program office needs, some evidence that there may be an opportunity to refine disposition codes

Implications + Future Research

- Text clustering could be a viable method of analyzing unstructured survey data
  - Clustering algorithm effectively clustered based on comment similarities
- Method may produce more informative results with “better” data
- Finetuning algorithm
  - Requested additional data from program office to rerun algorithm
  - With additional data, make conclusions about revising disposition codes

Contact Information

Victoria R. Narine

Research Statistician

Behavioral Science Research Center

Office of Survey Methods Research

Narine.Victoria@bls.gov

Josh Langeland

Research Statistician

Behavioral Science Research Center

Office of Survey Methods Research

Langeland.Joshua@bls.gov