Preliminaries

- Acknowledge my Bureau of Labor Statistics (BLS) colleagues
  - John Eltinge, Polly Phipps
  - Alex Measure, Stephen Pegula

- This work reflects the views of the speaker and does not reflect official policy, procedures, or applications of the U.S. Government.

- Objective – Provide an example of text analysis for data editing
Example – Accident Reports

- Downloaded accident reports from U.S. Department of Labor – Enforcement Data Website – OSHA:
- Records are Occupational Safety and Health Administration Fatality and Catastrophe Summaries
- Also known as Accident Investigation Summaries – OSHA 170 form
- Prepared once OSHA conducts an inspection after a catastrophe or fatality
Example – Accident Reports

Files have the following:

- Date of accident
- Event description – short phrase *
- Event keywords – Part, Source, etc *
- **Event type** – numeric label
- **Abstract** – unstructured text *
- More ...

* This is a text field.
Text Analysis for Editing

- Approach is to build a classifier
  - **Event Type** as the class label
  - **Abstract** as the predictor

- Use for data editing and verification tasks:
  - Examine misclassified records for editing
  - Explore consistency of labeling
  - Multiple and missing labels
Process the Data

- Extracted accident reports for November and December of 2011
- Preprocess text
  - Remove special characters
  - Remove stop words
- Yielded 358 accident reports
- Lexicon had 3841 words
The most common approach is the bag of words or term-document matrix.

The rows correspond to words.

The columns correspond to documents.

The \((i,j)\) -th entry in the matrix is the number of times the \(i\) -th word appears in the \(j\) -th document.
Preparing Data Set

- Event types were adjusted:
  - Combine two falling events – 04 -> 05
  - Combine two struck events – 06 -> 01
- Used codes 01, 02, 05, 14

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Struck by</td>
<td>08</td>
<td>Inhalation</td>
</tr>
<tr>
<td>02</td>
<td>Caught in or between</td>
<td>09</td>
<td>Ingestion</td>
</tr>
<tr>
<td>03</td>
<td>Bite/Sting/Scratch</td>
<td>10</td>
<td>Absorption</td>
</tr>
<tr>
<td>04</td>
<td>Fall (same level)</td>
<td>11</td>
<td>Rep. motion/pressure</td>
</tr>
<tr>
<td>05</td>
<td>Fall (from elevation)</td>
<td>12</td>
<td>Cardio/breathing failure</td>
</tr>
<tr>
<td>06</td>
<td>Struck against</td>
<td>13</td>
<td>Shock</td>
</tr>
<tr>
<td>07</td>
<td>Rubbed/abraded</td>
<td>14</td>
<td>Other</td>
</tr>
</tbody>
</table>
Data Set

- Data set included 332 records
- After processing – 3,665 words
- Classifier built – 01, 02, and 05

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>00&lt;sup&gt;a&lt;/sup&gt;</td>
<td>5</td>
</tr>
<tr>
<td>01</td>
<td>100</td>
</tr>
<tr>
<td>02</td>
<td>78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>05</td>
<td>106</td>
</tr>
<tr>
<td>14&lt;sup&gt;a&lt;/sup&gt;</td>
<td>27</td>
</tr>
</tbody>
</table>

Dupes<sup>a</sup> | 22 |

<sup>a</sup> Records in these categories need editing and verification
Reduce Dimensions

- Used Latent Semantic Analysis
- Based on singular value decomposition of TDM

\[ X = TSD^T \]

- Left singular vectors in \( T \) span the document space
- Right singular vectors in \( D \) span the word/term space
- Use matrix \( D \) to reduce dimensionality \( \sim \) PCA
Choosing the Number of Dimensions

- Use a scree plot
- Look for elbow in the curve
- Chose 4 dimensions
Build Classifier

- Many options:
  - Naïve Bayes
  - $K$ nearest neighbors
  - Classification Trees

- First removed:
  - Dupes
  - Codes 00, 14
Assess Results

Classification tree used for editing tasks:

- Determine event type labels for code 00 records
- Propose an event type for code 14 (Other)
- Suggest event type for duplicates
- Verify misclassified records

Code 00 records – two could be classified

<table>
<thead>
<tr>
<th>Suggested Code</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 – Struck by</td>
<td>Truck Tips over with Employee in It</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Employee Does Not Sustain Injuries in Fall from Ladder</td>
</tr>
</tbody>
</table>
Assess Results

- Predicted event type for records coded as 14 (*Other*)

<table>
<thead>
<tr>
<th>Suggested Code</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 – Struck</td>
<td>Employee Is Injured in Trash Truck Overturn</td>
</tr>
<tr>
<td>01 – Struck</td>
<td>Worker Lacerates Hand on Angle Grinder Used on Concrete</td>
</tr>
<tr>
<td>02 – Caught</td>
<td>Employee Gets Finger Laceration by Door Clamp</td>
</tr>
<tr>
<td>02 – Caught</td>
<td>Employee Gets Finger Amputations with a Table Saw</td>
</tr>
<tr>
<td>02 – Caught</td>
<td>Employee's Finger Is Amputated by Miter Saw</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Water Tank Crushes Employee</td>
</tr>
</tbody>
</table>
### Assess Results

**Predicted event type for duplicates**

<table>
<thead>
<tr>
<th>Suggested Code</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>05 – Fall</td>
<td>Elevated Scissor Lift Collapses and Injures Two Workers</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Two Employees Are Killed When Aerial Lift Collapses</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Two Employees Are Injured When Elevated Platform Collapses</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Worker Fractures Leg During Wall Board Movement</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Employee Sustains Head Injuries in Fall Off Roof</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>One Employee Is Killed and Four Injured When Floor Collapses</td>
</tr>
<tr>
<td>01 – Struck by</td>
<td>Two Employees Are Injured When Forklift Tip Over</td>
</tr>
<tr>
<td>01 – Struck by</td>
<td>Employee Is Killed When Crane Strikes Lift; Another Injured</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Employee Is Killed When Struck on Neck by Chain Saw</td>
</tr>
<tr>
<td>05 – Fall</td>
<td>Employee Dies from Crushing/drowning</td>
</tr>
<tr>
<td>01 – Struck by</td>
<td>Two Employees Killed in Fall from Basket Attached to Crane</td>
</tr>
<tr>
<td>01 – Struck by</td>
<td>Employee Is Injured When Hand Gets Caught in Equipment</td>
</tr>
</tbody>
</table>

*The predicted event types in the shaded rows seem plausible.*
Assess Results

- Look at 42 misclassified records – compare with short event description
  - Thirty (30) records seemed to be truly misclassified – based on event description
  - Six (6) records did not have correct label and our classifier was correct (my opinion).
    - “Roof framer falls from an unsecured beam and injures head”
      - Coded as 01 (Struck) and we coded as 05 (Fall)
  - Six (6) other records had event types where either seemed reasonable
Refining the Analysis

- Try stemming and term weights
- Add more data to create the classifier
- Build classifiers using other approaches
  - Random forests
  - Bagging and boosting
  - $K$ nearest neighbors
- Use alternative encoding that incorporates word order
Other Application

- Verify consistency of coding
- For example: Are investigators consistent when encoding accidents involving lacerations or amputations?
- Cluster abstracts – what are the event types?

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 – 52%</td>
<td>01 – 23%</td>
<td>01 – 41%</td>
</tr>
<tr>
<td>02 – 24%</td>
<td>02 – 9%</td>
<td>02 – 50%</td>
</tr>
<tr>
<td>05 – 24%</td>
<td>05 – 68%</td>
<td>05 – 9%</td>
</tr>
</tbody>
</table>
Other Applications

- Analyze OSHA abstracts each month to look for trends in accidents
- Use text to determine if OSHA coding depends on state
- Create a classifier that OSHA investigators can use to suggest event type
Useful Links

- LSA/LSI: [http://lsa.colorado.edu/papers/dp1.LSAintro.pdf](http://lsa.colorado.edu/papers/dp1.LSAintro.pdf)
- Text data mining: [http://projecteuclid.org/euclid.ssu/1216238228](http://projecteuclid.org/euclid.ssu/1216238228)
- Previous work in text Analysis: [http://www.fcsm.gov/events/prior.html](http://www.fcsm.gov/events/prior.html)
Contact Information

Wendy Martinez

Director, Mathematical Statistics Research Center
Office of Survey Methods Research
U.S. Bureau of Labor Statistics

www.bls.gov/osmr/home.htm
202-691-7400
martinez.wendy@bls.gov