Outline

- Problem Statement
- Introduction
- Approach
  - Sanitizing Data
  - Filtering Data
  - Generating Paragraph Summaries
- Results
- Improvements
- Interesting Findings
- Lessons Learned
- Acknowledgements
Generate an easily readable summary of an event given a large collection of webpages
Introduction

- Explore Problem Space
- Develop a Solution Strategy
- Implement
- Analyze Results
- Conclude and Communicate

Iterate
Data Set

Team J – Shootings

Small Collection
Adam Lanza fatally shoots 20 children and 6 adults at Sandy Hook Elementary School in Newtown Connecticut

Big Collection
Jared Lee Loughner shoots Congresswoman Gabriel Giffords in the head, in Tucson Arizona
Sanitizing Data
Stop-words List

• Stop-words
  • Using Default NLTK stop-words
• Website specific words added to a comprehensive stop-words list
  • News
  • Points
  • Comment
  • Password
  • Login
Sanitizing Data
Removing Files

• Lower case - consistency
• Positively classified documents – Provided strong training set
• Removed suspicious sentences – Sentences with words over 16 characters.
• Removed tiny files – primarily files with hyperlinks

<table>
<thead>
<tr>
<th></th>
<th>Small Collection</th>
<th>Large Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Size</td>
<td>4519</td>
<td>37829</td>
</tr>
<tr>
<td>Size after Sanitization</td>
<td>2626</td>
<td>5456</td>
</tr>
<tr>
<td>Percent reduction</td>
<td>41.9%</td>
<td>85.6%</td>
</tr>
</tbody>
</table>
Sanitizing Data

Classifying

• Small Event: 342 training files  
  • 160 positive files

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Time taken to train (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.9383</td>
<td>2.48</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.9619</td>
<td>12.06</td>
</tr>
<tr>
<td>MaxEntropy</td>
<td>0.931</td>
<td>241.83</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9938</td>
<td>4.99</td>
</tr>
</tbody>
</table>

• Big Event: 77 training files  
  • 20 positive files

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Time taken to train (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.8039</td>
<td>5.75</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.9167</td>
<td>6.76</td>
</tr>
<tr>
<td>MaxEntropy</td>
<td>0.9566</td>
<td>597.41</td>
</tr>
<tr>
<td>SVM</td>
<td>1</td>
<td>6.02</td>
</tr>
</tbody>
</table>

Training the Small Collection

Training the Big Collection
## Named Entity Recognition

<table>
<thead>
<tr>
<th>Named Entities</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>lanza</td>
<td>Person</td>
</tr>
<tr>
<td>connecticut</td>
<td>Location</td>
</tr>
<tr>
<td>newtown</td>
<td>Location</td>
</tr>
<tr>
<td>sandy</td>
<td>Person</td>
</tr>
<tr>
<td>hook</td>
<td>Person</td>
</tr>
</tbody>
</table>
Process

Extract Data → Filter Data → Apply Grammar
Generating Paragraph Summaries
Regular Expressions

• Regular expressions to match our regular grammar
  Examples:
  ( shooter | murderer | killer | gunman )\[s,]+([a-z]+\s[a-z]+)
  ( ([0-9]+\s(injured | wounded | hurt | damaged | lived)) )

• Date
  • (Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday)

• Time of day
  • (Morning | Afternoon | Evening | Night)
Generating Paragraph Summaries

Filtering

• Names
  • POS tagging (backoff tagger)
  • Stop-words

• Numbers
  • Extract and Verify value
    • Verify if 1 <= date <= 31
  • Stop-words
## Generating Paragraph Summaries

### Regular Grammar

<table>
<thead>
<tr>
<th>Non-Terminals</th>
<th>Terminals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>&lt;intro&gt; &lt;weapon description&gt; &lt;killed&gt; &lt;wounded&gt; &lt;age range&gt;</td>
</tr>
<tr>
<td>intro</td>
<td>On the &lt;time&gt; of &lt;date&gt;, &lt;shooter&gt; opened fire in &lt;location&gt;</td>
</tr>
<tr>
<td>weapon description</td>
<td>The &lt;gunman plurality&gt; fired &lt;number&gt; rounds out of his &lt;weapon&gt;.</td>
</tr>
<tr>
<td>killed</td>
<td>&lt;number&gt; (children</td>
</tr>
<tr>
<td>wounded</td>
<td>&lt;number&gt; of the victims were hurt, and are being treated for their injuries.</td>
</tr>
<tr>
<td>age range</td>
<td>The victims were between the ages of &lt;number&gt; and &lt;number&gt;</td>
</tr>
<tr>
<td>weapon</td>
<td>&lt;word&gt; &lt;weapon&gt;</td>
</tr>
<tr>
<td>time</td>
<td>(morning</td>
</tr>
<tr>
<td>date</td>
<td>&lt;day of week&gt;, &lt;month&gt; &lt;number&gt;</td>
</tr>
<tr>
<td>day of week</td>
<td>(monday</td>
</tr>
<tr>
<td>shooter</td>
<td>&lt;word&gt; &lt;word&gt;</td>
</tr>
<tr>
<td>location</td>
<td>&lt;word&gt;</td>
</tr>
<tr>
<td>number</td>
<td>[0-9]+</td>
</tr>
<tr>
<td>word</td>
<td>[a-z]+</td>
</tr>
</tbody>
</table>
## Hadoop Job Run Timings

### Small Collection

<table>
<thead>
<tr>
<th>Job</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 1</td>
<td>74 seconds</td>
</tr>
<tr>
<td>Run 2</td>
<td>70 seconds</td>
</tr>
<tr>
<td>Run 3</td>
<td>100 seconds</td>
</tr>
<tr>
<td><strong>Average:</strong></td>
<td><strong>81.3 seconds</strong></td>
</tr>
</tbody>
</table>

### Big Collection

<table>
<thead>
<tr>
<th>Job</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 1</td>
<td>231 seconds</td>
</tr>
<tr>
<td>Run 2</td>
<td>213 seconds</td>
</tr>
<tr>
<td>Run 3</td>
<td>226 seconds</td>
</tr>
<tr>
<td><strong>Average:</strong></td>
<td><strong>223.3 seconds</strong></td>
</tr>
</tbody>
</table>
Results

- Newton School Shooting:
  On the morning of Saturday, December 15, Adam Lanza opened fire in Connecticut. The gunman fired 100 rounds out of his rifle. 27 children lost their lives. 2 of the children were hurt, and are being treated for their injuries. The victims were between the ages of 6 and 7.

- Tucson Gabrielle Giffords Shooting:
  On the night of Sunday, January 9, Jared Lee opened fire in Tucson. The suspect fired 5 rounds out of his rifle. 6 people lost their lives. 32 of the people were hurt, and are being treated for their injuries. The victims were between the ages of 40 and 50.
Improvements

• Common Sense Verification
• Utilizing Dependencies and Context
• Better Candidacy Selection
• Iterative Slot Filling
• Better Cleaning up Big Collection
• Capitalized NEs
• More informative summary
Interesting Findings

- Recall vs. Precision
- Selecting the best candidate
- Killer's name vs victims' names
- 100 vs. Hundreds
- POS vs. NER
Lessons Learned

• Allocate time for iterations
• Cleaning a collection is vital to getting good results
• Small files can be considered noise
• Simplicity is key
• Reuse Code
• File Structure for Organization
Acknowledgements

Dr. Edward A. Fox
fox@vt.edu

Xuan Zhang
xuancs@vt.edu

Tarek Kanan
tarekk@vt.edu

Mohamed Magdy Gharib Farag
mmagdy@vt.edu

With support from NSF DUE-1141209 and IIS-1319578
Questions?

Arjun Chandrasekaran
carjun@vt.edu

Saurav Sharma
svsharma@vt.edu

Peter Sulucz
peters1@vt.edu

Jonathan Tran
jtran372@vt.edu