shootings

CS-4984 Computational Linguistics

Arjun Chandrasekaran
Saurav Sharma
Peter Sulucz
Jonathan Tran

December 2014

Virginia Tech

Blacksburg, VA

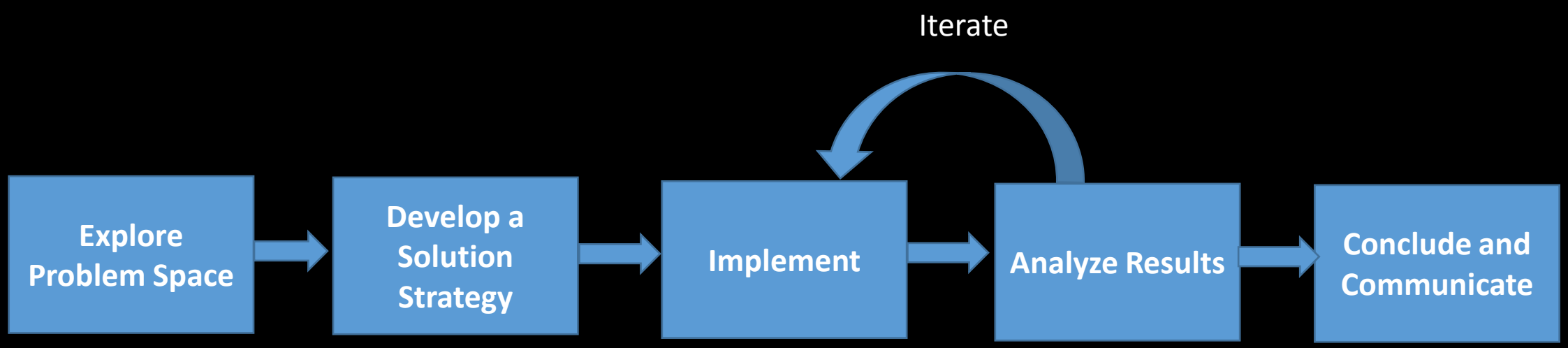
Outline

- Problem Statement
- Introduction
- Approach
  - Sanitizing Data
  - Filtering Data
  - Generating Paragraph Summaries
- Results
- Improvements
- Interesting Findings
- Lessons Learned
- Acknowledgements

Problem Statement

Generate an easily readable summary of an event given a large collection of webpages

Introduction



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

	Small Collection	**Large Collection**
Initial Size	4519	37829
Size after Sanitization	2626	5456
Percent reduction	41.9%	85.6%

Sanitizing Data

Classifying

- Small Event: 342 training files
  - 160 positive files
- Big Event: 77 training files
  - 20 positive files

- Training the Small Collection

Classifier	Accuracy	Time taken to train (s)
Naïve Bayes	0.9383	2.48
Decision Tree	0.9619	12.06
MaxEntropy	0.931	241.83
SVM	0.9938	4.99

- Training the Big Collection

Classifier	Accuracy	Time taken to train (s)
Naïve Bayes	0.8039	5.75
Decision Tree	0.9167	6.76
MaxEntropy	0.9566	597.41
SVM	1	6.02

Named Entity Recognition

Named Entities	Tags
lanza	Person
connecticut	Location
newtown	Location
sandy	Person
hook	Person

Process

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

Non-Terminals	Terminals
Summary	<intro> <weapon description> <killed> <wounded> <age range>
intro	On the <time> of <date>, <shooter> opened fire in <location>
weapon description	The <gunman plurality> fired <number> rounds out of his <weapon>.
killed	<number> (children | people) lost their lives.
wounded	<number> of the victims were hurt, and are being treated for their injuries.
age range	The victims were between the ages of <number> and <number>
weapon	<word> <weapon>
time	(morning | afternoon | night | evening)
date	<day of week>, <month> <number>
day of week	(monday | tuesday | wednesday | thursday | friday | saturday | sunday)
shooter	<word> <word>
location	<word>
number	[0-9]+
word	[a-z]+

Hadoop Job Run Timings

Small Collection

Job	Timing
Run 1	74 seconds
Run 2	70 seconds
Run 3	100 seconds
Average:	81.3 seconds

Big Collection

Job	Timing
Run 1	231 seconds
Run 2	213 seconds
Run 3	226 seconds
Average:	223.3 seconds

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