

# Katrina Database

## SearchKat

University: Virginia Tech

Course: CS 4624 (Multimedia, Hypertext, and Information Access)

Professor: Edward A. Fox

Date: April 30, 2015

Group: Matthew Chittum, Kyle He, Gary Li, Tanvir Rahman

Location: Blacksburg, VA 24061, USA

# Introduction

- A cross-disciplinary project combining linguistics with Computer Science.
- Create a searchable database of interviews of Hurricane Katrina victims.
- Supports Thematic Searching
  - Search by word association
    - synonyms, antonyms, etc. - query expansion
    - client specific word association

# Client

- Dr. Katie Carmichael
  - Assistant Professor in the English Department
    - College of Liberal Arts and Human Sciences at VT
  - Ph.D from Ohio State University, master's and bachelor's degree from Tulane University
  - Works in 407 Shanks Hall
  - 181 Turner St NW, Blacksburg, VA 24061, United States
  - Phone: 540-231-7712
  - email: [katcarm@vt.edu](mailto:katcarm@vt.edu)
- Search 'Katie Carmichael' at <http://www.vt.edu/>

# Project Phases

- Phase 1: Removing Markings
- Phase 2: Word Count
- Phase 3: Search Implementation
  - Basic Searching
  - Thematic Searching/ Query Expansion

# Phase 1: Remove Markings

- Remove unnecessary markings from documents.

Before

```
1 {SL}
2 {BR}
3 So went to Lafayette
4 sp
5 very good friends
6 sp
7 we had
8 sp
9 maybe five households of people
10 sp
11 friends and
12 sp
13 and their relatives that
```

After

```
So went to Lafayette
very good friends
we had
maybe five households of people
friends and
and their relatives that
```

# Phase 2: Word Count

- Find word frequency across all documents.
  - Using the results, Dr. Carmichael picked certain words to be used in query expansion.
  - ex: 'people' appeared 31 times in document X

	A	B	C	D	E	F	G
1	Narrator	word #					
2		people 31					
3		because 26					
4		didn't 16					
5		that's 13					
6		things 12					
7		pictures 10					
8		little 9					
9		couldn't 9					
10	Acme	through 9					

# Word Groupings

- Social class
  - (Social class, academic, college, contest, income, palmisano, tulane, tuition).
- Victims/refugees
  - (Damage, damaged, damages, denied, destroyed, devastating, refugee, refugees).

# Phase 3: Search Implementation

- First run for searching
  - Customize desired displayed category for with search results.
  - Customize desired display categories for search results.
  - Display the line of text that contains the search term
  - Many results are too short, reflect very little contextual information



# Phase 3: Search Implementation

**file:** Bella-NT.txt

**line:** 416

**line-content:** Katrina

**file:** Luke-NT.txt

**line:** 425

**line-content:** Katrina

**file:** Super-NT.txt

**line:** 29

**line-content:** Katrina

**file:** Greg-NT.txt

**line:** 252

**line-content:** before Katrina and then a couple times after Katrina we would see each other

# Thematic Searching

- Able to search by word groupings.
- Group religion contains: pastor, prayed, prayer, prayers, etc..

29 results found in 2ms Page 1 of 3

**File-Name:** Maria-NT.txt

**Line-Number:** 345

**Content-Snippet:** Three prayers that cure the arts of the dying person Just everything you know we just

**File-Name:** Maria-NT.txt

**Line-Number:** 346

**Content-Snippet:** Just everything you know we just Prayed and prayed and prayed And gave him I gave h

**File-Name:** Maria-NT.txt

**Line-Number:** 342

**Content-Snippet:** And so we started praying on you know Prayers and we start praying every prayer that

**File-Name:** Maria-NT.txt

**Line-Number:** 347

**Content-Snippet:** Prayed and prayed and prayed And gave him I gave him a spiritual communion Be caus

# Interesting Results

- Some initial query results were only one word.
- Solr needs fields for its search results. If not specified, it will try to retrieve them from the source files.
- Converting original documents to CSV files with specified fields (filename, line content, etc.) helps Solr to return desired search results.

# Lessons Learned

- Learned to utilize Solr
- Learned about major concepts of search engine: indexing, result filtering, running query, query expansion, etc.
- Project reports suck

# Acknowledgements

- Katie Carmichael
- Edward A. Fox
- Mohamed Magdy Gharib Farag