Knowledge Graph For Researchers
Kyle Vincent and Emma Meno
CS 4624: Multimedia, Hypertext, and Information Access
Dr. Edward Fox
Virginia Tech, Blacksburg, VA 24061
April 28, 2020
OUTLINE

- Project Deliverables [3]
- Accomplishments [5]
- Testing and Evaluation [20]
- Lessons Learned [22]
- Future Work [26]
- Acknowledgements [28]
- References [29]
Deliverables
Project Team → Project Reports → Dataset - Algo - Dataset Tuples

Knowledge Graph → And Make Up (Input Data Set, Output Data Set) → Workflow Diagram

Subject Matter Experts → Query Service → Database

Reads → To Generate → Stored In

To Return
Project Deliverables

- Build ontology for Twitter-based project info an SME might want to query, answer or calculate
- Build knowledge graph based on created ontology
- Build set of APIs to trigger set of network algorithms based on info queried to graph
Accomplishments
Literature Review: Twitter Data Source

- 2016
- 2017

CS5604: Information Retrieval

BROWSE BY
By Issue Date  Authors  Titles  Subjects

Search within this collection:

This collection contains the final projects of the students in various offerings of the course Computer Science 5604: Information Retrieval. This course is taught by Professor Ed Fox. Analyzing, indexing, representing, storing, searching, retrieving, processing and presenting information and documents using fully automatic systems. The information may be in the form of text, hypertext, multimedia, or hypermedia. The systems are based on various models, e.g., Boolean logic, fuzzy logic, probability theory, etc., and they are implemented using inverted files, relational thesauri, special hardware, and other approaches. Evaluation of the systems' efficiency and effectiveness.
# Built Ontology

## 2016 Reports (CS 5604: Information Retrieval)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task ID</td>
<td>Task Name</td>
<td>Input File(s)</td>
<td>Output File(s)</td>
<td>Dependent Task ID</td>
<td>Functions / Libraries</td>
<td>Manual (Y/N)</td>
<td>Data Collection</td>
<td>Command Line Parameters</td>
<td>Project</td>
<td>Report Names</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Remove stop words and lemmatize</td>
<td>Cleanedtext from HBase (RDD of tweets)</td>
<td>Cleaned Lemmatized Tweets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>table name data type for each row key desired columns</td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Manually Classify Tweets</td>
<td>Cleaned Lemmatized Tweets AND Set of Events Classes</td>
<td>Tweet Training Data</td>
<td>1</td>
<td>N/A</td>
<td>Y</td>
<td></td>
<td>collection ID batch size</td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Feature Selection</td>
<td>Tweet Training Data</td>
<td>Word Feature Representation</td>
<td>1</td>
<td>Word2Vec Model</td>
<td>N</td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Feature Selection</td>
<td>Tweet Training Data</td>
<td>Word Feature Representation</td>
<td>1</td>
<td>Association Rules Based Classifier (provided by Dr. Pereira)</td>
<td>N</td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Tweet Classification</td>
<td>Word Feature Representation</td>
<td>Predicted Class of the Given Tweet // Associated Probabilities of Each Class</td>
<td>2.3</td>
<td>Multi-class logistic regression classifier</td>
<td>N</td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Convert to CSV</td>
<td>Cleaned Tweets in HBase</td>
<td>CSV Format Tweets</td>
<td>37</td>
<td>Pig Script</td>
<td>N</td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604: Information and Storage Retrieval Fall 2016 - CMT (Collection Management Tweets)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Record User, URL, and Tweet Relationships</td>
<td>CSV Format Tweets</td>
<td>Social Network Matrix</td>
<td>5</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604: Information and Storage Retrieval Fall 2016 - CMT (Collection Management Tweets)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Calculate Importance Factors for Nodes</td>
<td>Social Network Matrix</td>
<td>Weighted Social Network Matrix</td>
<td>6</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td><a href="https://vtechworks.lib.vt.edu/.handle/10919/37373">https://vtechworks.lib.vt.edu/.handle/10919/37373</a></td>
<td>CS5604: Information and Storage Retrieval Fall 2016 - CMT (Collection Management Tweets)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Nodes.csv

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>field</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>Cleantext from HBase (RDD of tweets)</td>
</tr>
<tr>
<td>2</td>
<td>Cleaned Lemmatized Tweets</td>
</tr>
<tr>
<td>3</td>
<td>Set of Events Classes</td>
</tr>
<tr>
<td>4</td>
<td>Tweet Training Data</td>
</tr>
<tr>
<td>5</td>
<td>Word Feature Representation</td>
</tr>
<tr>
<td>6</td>
<td>Predicted Class of the Given Tweet</td>
</tr>
<tr>
<td>7</td>
<td>Associated Probabilities of Each Class</td>
</tr>
<tr>
<td>8</td>
<td>Cleaned Tweets in HBase</td>
</tr>
</tbody>
</table>

**Diagram:**

![Diagram showing the relationship between Dataset, Algo, and Dataset](image-url)
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>taskid</td>
<td>name</td>
<td>inputid</td>
<td>outputid</td>
<td>functionsAndLibraries</td>
<td>manual</td>
<td>commandLineParameters</td>
<td>reportURL</td>
<td>reportName</td>
<td>domainCollection</td>
</tr>
<tr>
<td>2</td>
<td>Remove stop words and lemmatize</td>
<td>1</td>
<td>2</td>
<td>StanfordNLP</td>
<td>Y</td>
<td></td>
<td>table name; data type for each row-key; desired columns</td>
<td><a href="https://vtechworks.lib.vt.edu/handle/10919/73713">https://vtechworks.lib.vt.edu/handle/10919/73713</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td>Kentucky Shooting, Newton Shooting, Firefighter Shooting, Hurricane Arthur, Hurricane Sandy, Hurricane Isaac, China Factory Explosion, Texas Plant Explosion, Manhattan Explosion</td>
</tr>
<tr>
<td>3</td>
<td>2 Manually Classify Tweets</td>
<td>2</td>
<td>4</td>
<td>N/A</td>
<td>Y</td>
<td></td>
<td>collection ID; batch size</td>
<td><a href="https://vtechworks.lib.vt.edu/handle/10919/73713">https://vtechworks.lib.vt.edu/handle/10919/73713</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td>Kentucky Shooting, Newton Shooting, Firefighter Shooting, Hurricane Arthur, Hurricane Sandy, Hurricane Isaac, China Factory Explosion, Texas Plant Explosion, Manhattan Explosion</td>
</tr>
<tr>
<td>4</td>
<td>3 Manually Classify Tweets</td>
<td>3</td>
<td>4</td>
<td>N/A</td>
<td>Y</td>
<td></td>
<td>collection ID; batch size</td>
<td><a href="https://vtechworks.lib.vt.edu/handle/10919/73713">https://vtechworks.lib.vt.edu/handle/10919/73713</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td>Kentucky Shooting, Newton Shooting, Firefighter Shooting, Hurricane Arthur, Hurricane Sandy, Hurricane Isaac, China Factory Explosion, Texas Plant Explosion, Manhattan Explosion</td>
</tr>
<tr>
<td>5</td>
<td>4 Feature Selection</td>
<td>4</td>
<td>6</td>
<td>Word2Vec Model</td>
<td>N</td>
<td></td>
<td>path_to_jar_file; support threshold; block size; cluster node; base_dir on HDFS; path_to_training_file; path_to_test_file; stopwords_file; output_directory</td>
<td><a href="https://vtechworks.lib.vt.edu/handle/10919/73713">https://vtechworks.lib.vt.edu/handle/10919/73713</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td>Kentucky Shooting, Newton Shooting, Firefighter Shooting, Hurricane Arthur, Hurricane Sandy, Hurricane Isaac, China Factory Explosion, Texas Plant Explosion, Manhattan Explosion</td>
</tr>
<tr>
<td>6</td>
<td>6 Feature Selection</td>
<td>4</td>
<td>6</td>
<td>Pereira</td>
<td>N</td>
<td></td>
<td>path_to_jar_file; support threshold; block size; cluster node; base_dir on HDFS; path_to_training_file; path_to_test_file; stopwords_file; output_directory</td>
<td><a href="https://vtechworks.lib.vt.edu/handle/10919/73713">https://vtechworks.lib.vt.edu/handle/10919/73713</a></td>
<td>CS5604 Fall 2016 Classification Team Final Report</td>
<td>Kentucky Shooting, Newton Shooting, Firefighter Shooting, Hurricane Arthur, Hurricane Sandy, Hurricane Isaac, China Factory Explosion, Texas Plant Explosion, Manhattan Explosion</td>
</tr>
</tbody>
</table>

**Diagram:**

```
Dataset --> Algo --> Dataset
```
## Path Metrics

<table>
<thead>
<tr>
<th>Path Length</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
</tr>
</tbody>
</table>

![Histogram of Path Lengths](chart.png)
Interim Project Toolkit - The GRANDstack

GraphQL  React  Apollo  Neo4j Database
Implemented Project Toolkit - Grakn

- Database storage/interface
- Graql queries for reads, writes, & analysis
- Three Client APIs:
  - Node.js
  - Python
  - Java
The And/Or Problem

- With traditional node-edge-node relationships, we weren’t sure how to differentiate between two scenarios present in our data:
  - A certain output requires Input 1 AND Input 2 to be produced
  - You can use either Input 1 OR Input
- This means the difference between returning 2 paths and 1 path
Solution: The Hypergraph

- Recommended by our Client
- A hyperedge can connect any number of entities
- Now we can distinguish between hyperedges and normal edges
Grakan Database Schema

- Our client recommended a tool called Grakan.
  - Designed for Knowledge Graphs specifically
  - Open Source
  - Allows Hyper-Relationships
- Sample of our Schema:
Migration Scripts

- With schema established, starting migrating the data
- Grakn has three clients:
  - Python
  - Java
  - Node.js
- Wrote data migration scripts that could use these clients and migrate from CSV → Grakn DB
- This proved way more challenging than it had been with Neo4j
Possible User Interface: Console

- One possible way to query the data is the Grakn Console
- It is included with the download of the software
- Uses the Graql language.
Testing and Evaluation
Testing and Evaluation

- First test to be conducted by client shortly after project submission
- Next test will be to have a subject matter expert use both the console and the Workbase UI
Lessons Learned
Challenges Faced

- Waiting on network science data
- Gathering sufficient material from project reports
- Selecting an appropriate project toolkit
- Transition to virtual learning
Solutions

- Focus on different data → Twitter
- Including reports from multiple years → 2016 & 2017
- Shift to Grakn
- Use of Zoom and other online platforms
Overall Lessons

● Asking more questions earlier in the design process
● Communicating with the professor, client, and team
● Documentation and project tools
Future Work
Future Work

- Upload the rest of the data to the graph
- Build upon schema to aid better path queries
- Use Node.js Cient to establish a website as front end
- Automate process of parsing through reports
Acknowledgements

**Client:** Prashant Chandrasekar
- Fifth-year PhD student
- Works with Dr. Fox
- *Our Work:* Proof of Concept for PhD

**Professor:** Dr. Edward Fox
- Director of Digital Library Research Laboratory
References

”Grakn.AI – Text Mined Knowledge Graphs”,
https://www.stockwerk.co.at/event/grakn-ai-textmined-knowledge-graphs/

”Grakn Schema Overview”, https://dev.grakn.ai/docs/schema/overview

Directed Hypergraph and Applications:

Grakn Documentation: https://grakn.ai/
Questions?