Applying Web Scraping to Kenyan Politics

A STUDY BY JOSH FROMM AND ASHLEY GUO
AN 127
29 MAY 2014
Outline

- Goals
- Background
- Method
- Issues
- Results
- Improvements
- Conclusion
Project Goals

- Estimate the corruption level of Kenyan Politicians
- Detect relationships between public figures and high profile individuals
- Identify politicians associated with Mega Scandals
- Collect data from multiple sources to avoid bias
Background

Previous literature:

- Clustering press releases based on content (categorizing documents)\(^1\)
- Developing automated text analysis & correcting for misleading errors\(^2\)
- Extracting social networks and reconstructing networks from a collection of documents\(^3\)
- Name disambiguation\(^4\)

---

1. Grimmer 2011
2. Grimmer 2013, Reinanda 2013
Background

Three major online newspapers:
Background

Cop tells court Lari MP punched him

THURSDAY, MAY 8, 2014 - 00:00 - BY ALPHONSE MUNG'AU

A police officer yesterday told a Nairobi court that Lari MP Joseph Mburu Kahangara attacked him while he was motorists at Globe Cinema roundabout.

Constable David Rutto, attached to traffic department told Acting Chief Magistrate Daniel Ochiria that he had
Background

Which politicians are we interested in?

Elected in 2013:
- Members of Parliament (MPs)
- Governors
- Senators
- Cabinet Ministers (CMs)

Elected in 2007: (precedes all online news articles)
- MPs only
Method and Implementation

SQL Database Schema

- SQL is most widely used database software
- Free and easy to set up / pass on
- Provides simple access and querying capabilities
- Can be used by many users at once
Method and Implementation

MP Table – used for names of interest and word tracking

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Unique identifier</td>
</tr>
<tr>
<td>Name</td>
<td>List of aliases for this politician</td>
</tr>
<tr>
<td>Corrupt</td>
<td>Count of references to corrupt</td>
</tr>
<tr>
<td>Graft</td>
<td>Count of references to graft</td>
</tr>
<tr>
<td>Indict</td>
<td>Count of references to indict</td>
</tr>
<tr>
<td>Scandal</td>
<td>Count of references to scandal</td>
</tr>
</tbody>
</table>
Method and Implementation

‘Good Guy’ Table – used to keep track of public figures who associate with a list of known non-corrupt individuals

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Unique identifier</td>
</tr>
<tr>
<td>Name</td>
<td>Name of public figure</td>
</tr>
<tr>
<td>Count</td>
<td>Number of times seen with group members</td>
</tr>
</tbody>
</table>
Method and Implementation

‘Bad Guy’ Table – used to keep track of public figures who associate with a list of known very corrupt individuals

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Unique identifier</td>
</tr>
<tr>
<td>Name</td>
<td>Name of public figure</td>
</tr>
<tr>
<td>Count</td>
<td>Number of times seen with group members</td>
</tr>
</tbody>
</table>
Method and Implementation

Scandals Table – used for names of interest and word tracking

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Unique identifier</td>
</tr>
<tr>
<td>Name</td>
<td>List of aliases for this politician</td>
</tr>
<tr>
<td>Scandal</td>
<td>Which scandal was detected</td>
</tr>
<tr>
<td>Count</td>
<td>Number of times this politician has been associated with this scandal</td>
</tr>
</tbody>
</table>
Method and Implementation

Web Table – Used to keep track of which articles have already been checked

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>URL of an already checked article</td>
</tr>
</tbody>
</table>
Software Implementation

Written in Perl

- Widely Distributed
- Easy to learn
- Excellent regular expression integration

Several Phases:

- Determine which URLs to check
- Pull the data from the web
- Partition data into relevant portions
- Check each chunk for desired patterns
- Update database with result
Software Implementation

Determine which URLs to check:

- Not as easy as you might expect
- Requires webpage to either have archive or predictable article labeling
- Standard names article using simple ID: [www.standardmedia.co.ke/article/20000122681](http://www.standardmedia.co.ke/article/20000122681)
- Star provides full archive of all articles, but does not have useful ID
- Nation does not provide either. Very difficult to work with
Software Implementation

Pull data from web and extract useful bits:

- Getting full HTML is easy
- Unfortunately, HTML contains tons of irrelevant information
- Look for special tags indicating the body, varies for each news source
  - ‘<field-name-body’ ARTICLE BODY ‘sharethis-button’
- break remaining body into paragraphs
Software Implementation

Pattern Parsing

- Check if paragraph has any words of interest: corrupt, graft, scandal, indict.
  - If so, check that paragraph for any politician name (or alias)

- Check if paragraph has any good guy or bad guy in it
  - If so, check for ALL other names in that paragraph

- Check if paragraph mentions any scandals
  - If so, find all politician names in that paragraph.
Software Implementation

Upload Information

- If any match is found, upload into the appropriate table
- Determine if this match is new or has been seen before
  - If new, create a new entry
  - If old, update a count
- Create and submit simple SQL query to the database
Issues

- Very difficult to differentiate direct and indirect association
  - ‘Ashley and Josh were seen together in French class’ vs ‘Ashley and Josh are both students’

- Also difficult to identify the polarity of an association
  - ‘Josh is my friend’ vs ‘Josh is my worst enemy’

- Virtually impossible to differentiate names from two adjacent capitalized words
  - ‘Ashley Guo’ vs ‘Transparency International’
Results

Word association (a selection):

<table>
<thead>
<tr>
<th>Politician name</th>
<th>Corrupt</th>
<th>Graft</th>
<th>Indict</th>
<th>Scandal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uhuru Kenyatta</td>
<td>38</td>
<td>5</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>Charity Ngilu</td>
<td>26</td>
<td>7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Amos Wako</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Nyaga Wambora</td>
<td>149</td>
<td>108</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Odinga Amolo</td>
<td>40</td>
<td>9</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Gideon Mbuvi (&quot;Sonko&quot;)</td>
<td>23</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
## Results

Scandal selection (a selection):

<table>
<thead>
<tr>
<th>Politician name</th>
<th>Scandal</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evans Kidero</td>
<td>Standard Gauge Railway</td>
<td>162</td>
</tr>
<tr>
<td>Mutula Kilonzo</td>
<td>Anglo-Leasing</td>
<td>20</td>
</tr>
<tr>
<td>Henry Rotich</td>
<td>Standard Gauge Railway</td>
<td>12</td>
</tr>
<tr>
<td>Alfred Keter</td>
<td>Standard Gauge Railway</td>
<td>13</td>
</tr>
<tr>
<td>Amos Wako</td>
<td>Goldenberg</td>
<td>5</td>
</tr>
<tr>
<td>Eugene Wamalwa</td>
<td>BVR (Biometric Voter Registration)</td>
<td>4</td>
</tr>
</tbody>
</table>
## Results

Good Guy association (a selection):

<table>
<thead>
<tr>
<th>Name recognized</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jimmy Wanjigi</td>
<td>2</td>
</tr>
<tr>
<td>Kibaki’s Personal Secretary Alfred Getonga</td>
<td>2</td>
</tr>
<tr>
<td>David Mwangi</td>
<td>2</td>
</tr>
<tr>
<td>Lucy Hannan</td>
<td>3</td>
</tr>
<tr>
<td>Gladwell Otieno</td>
<td>5</td>
</tr>
<tr>
<td>Zahid Rajan</td>
<td>4</td>
</tr>
<tr>
<td>Open Governance</td>
<td>3</td>
</tr>
<tr>
<td>Chapter Six</td>
<td>1</td>
</tr>
</tbody>
</table>
## Results

**Bad Guy association (a selection):**

<table>
<thead>
<tr>
<th>Name recognized</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Punjani</td>
<td>19</td>
</tr>
<tr>
<td>Martha Karua</td>
<td>17</td>
</tr>
<tr>
<td>Mike Sonko (Gideon Mbuvi)</td>
<td>29</td>
</tr>
<tr>
<td>William Kabogo</td>
<td>28</td>
</tr>
<tr>
<td>Rachel Shebesh</td>
<td>26</td>
</tr>
<tr>
<td>George Saitoti</td>
<td>16</td>
</tr>
<tr>
<td>Ferdinand Waititu</td>
<td>13</td>
</tr>
<tr>
<td>Times New Roman</td>
<td>1</td>
</tr>
</tbody>
</table>
Validity test

- How successful are our tests?
- Check articles ourselves to check validity of results

Word frequency matching:
  - Correct: 12
  - Incorrect: 3

Scandal association:
  - Correct: 7
  - Incorrect: 3

Good Group association:
  - Correct: 0
  - Incorrect: 3

Bad Group association:
  - Correct: 21
  - Incorrect: 2
Validity test

How successful are our tests?

Check articles ourselves to check validity of results

Word frequency matching:
- Correct: 12
- Incorrect: 3

Scandal association:
- Correct: 7
- Incorrect: 3

Good Group association:
- Correct: 0
- Incorrect: 3

Bad Group association:
- Correct: 21
- Incorrect: 2

Insufficient sample size, seems okay by close inspection

Tedious and time-consuming: Definitely can benefit from future work
Improvements

- Improve phrase detection
- Improve successful name detection – very difficult
- Expand past Kenya!
- Further verification of results
- Add more news sources
Conclusion

- Successfully developed software suite
- Can serve as a foundation for other web scraping applications
- Natural language processing is necessary but difficult
- Appears to have potential for determining whether individuals are corrupt
- Has useful applications for automatic social group extraction