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
3. Pouliquen 2008, Baker 1993, Pujol 2002, Tang 2007
4. Reinanda 2013, Tang 2007

Background

Three major online newspapers:

Background

theSTAR

HOME | NATIONAL | LOCAL | BUSINESS | OPINIONS | SPORTS | LIFESTYLE | SOCIETY

STAR LIVE | SKYCAM | BRAZIL 2014

YOU ARE HERE: HOME » COP TELLS COURT LARI MP PUNCHED HIM

Cop tells court Lari MP punched him

THURSDAY, MAY 8, 2014 - 00:00 -- BY ALPHONCE MUNG'AHU

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

Constable David Ruto, attached to traffic department, told acting chief magistrate Daniel Ochenja that he had

Background

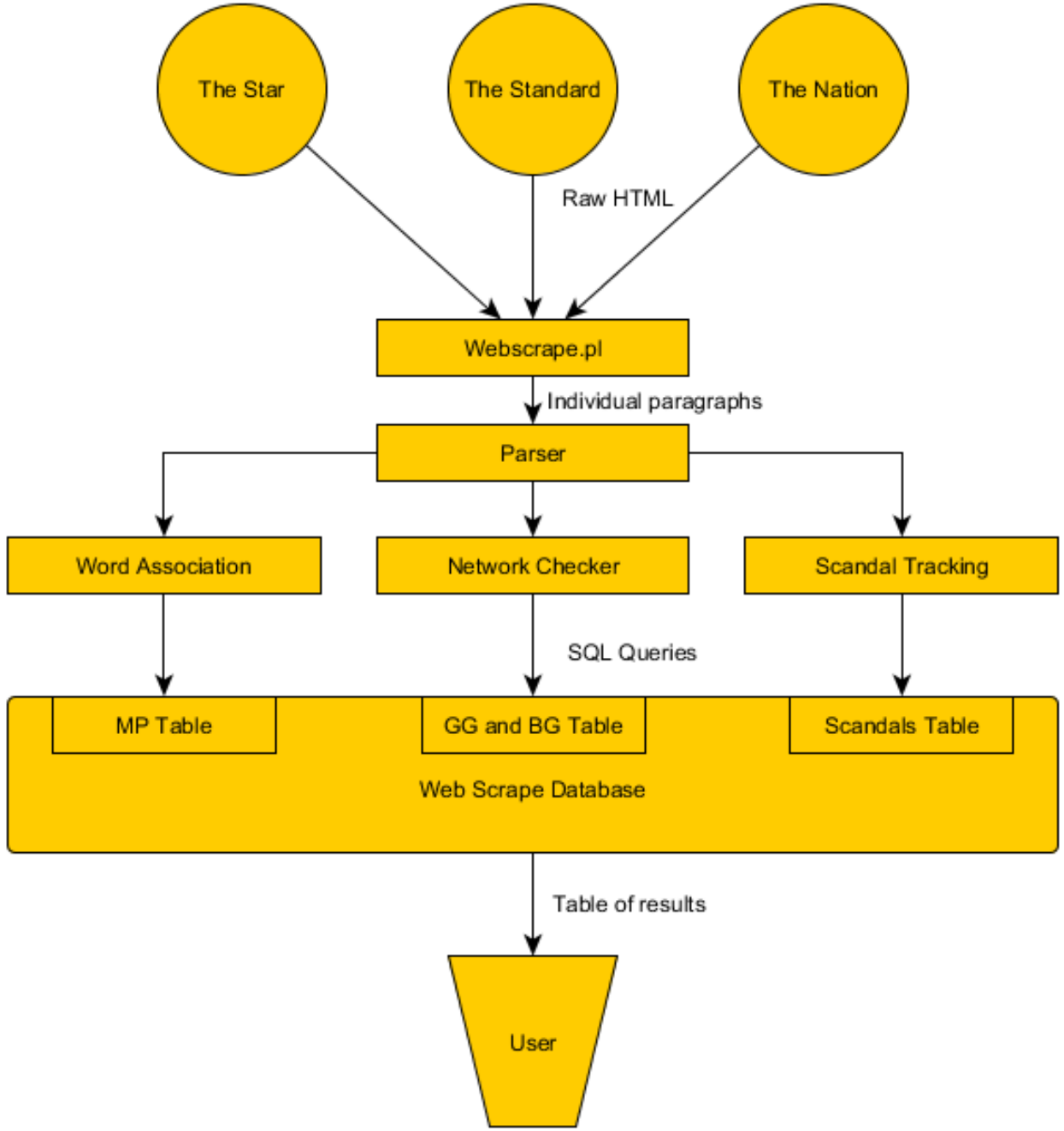
Which politicians are we interested in?

Elected in 2013:

- Members of Parliament (MPs)
- Governors
- Senators
- Cabinet Ministers (CMs)

Elected in 2007: (precedes all online news articles)

- MPs only


The Star
The Standard
The Nation
Raw HTML
Webscrape.pl
Individual paragraphs
Parser
Word Association
Network Checker
Scandal Tracking
SQL Queries
MP Table
GG and BG Table
Scandals Table
Web Scrape Database
Table of results
User

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

Field	Description
ID	Unique identifier
Name	List of aliases for this politician
Corrupt	Count of references to corrupt
Graft	Count of references to graft
Indict	Count of references to indict
Scandal	Count of references to scandal

Method and Implementation

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

Field	Description
ID	Unique identifier
Name	Name of public figure
Count	Number of times seen with group members

Method and Implementation

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

Field	Description
ID	Unique identifier
Name	Name of public figure
Count	Number of times seen with group members

Method and Implementation

Scandals Table – used for names of interest and word tracking

Field	Description
ID	Unique identifier
Name	List of aliases for this politician
Scandal	Which scandal was detected
Count	Number of times this politician has been associated with this scandal

Method and Implementation

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

Field	Description
URL	URL of an already checked article

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
- 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 if 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):

Politician name	Corrupt	Graft	Indict	Scandal
Uhuru Kenyatta	38	5	35	16
Charity Ngilu	26	7	1	3
Amos Wako	13	0	0	4
Nyaga Wambora	149	108	0	5
Odinga Amolo	40	9	11	12
Gideon Mbuvi (“Sonko”)	23	3	1	0

Results

Scandal selection (a selection):

Politician name	Scandal	Count
Evans Kidero	Standard Gauge Railway	162
Mutula Kilonzo	Anglo-Leasing	20
Henry Rotich	Standard Gauge Railway	12
Alfred Keter	Standard Gauge Railway	13
Amos Wako	Goldenberg	5
Eugene Wamalwa	BVR (Biometric Voter Registration)	4

Results

Good Guy association (a selection):

Name recognized	Count
Jimmy Wanjigi	2
Kibaki's Personal Secretary Alfred Getonga	2
David Mwangi	2
Lucy Hannan	3
Gladwell Otieno	5
Zahid Rajan	4
Open Governance	3
Chapter Six	1

Results

Bad Guy association (a selection):

Name recognized	Count
Ali Punjani	19
Martha Karua	17
Mike Sonko (Gideon Mbuvi)	29
William Kabogo	28
Rachel Shebesh	26
George Saitoti	16
Ferdinand Waititu	13
Times New Roman	1

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

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

- 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

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