Building Intelligence analysis systems

Hands-on with nutch, solr, lucene, maltego/netglub, and more

Agenda

- INTRO:Intelligence analysis systems
- ARCHITECTURE:components
- HANDS-ON:what's on your VM
- MOD01:data scapping
- MOD02:data storage
- MOD03:data viz (Maltego/Netglub: building transforms)
- FINI:other stuff and Q/A

INTRO:intelligence analysis

- IA is a way of reducing ambiguity in highly ambiguous situations (wiki)

INTO:Intelligence analysis

- Or rather - making large volumes of information available at your fingertips for your personal joy of owning personal custom google ;-)

हैकर

Scrappers

- HTTP: web crawlers, RSS feed parsers, forum crawlers, social media etc
- IRC bots
- ... Yer own ..

HANDS ON: on prepared VM
You'll find some samples, which we are
Going to play with. Roll your sleeves :-)

Data storage

- Small amounts of data: files (local, HDFS)
- SQL databases: works but scale poorly
- Non-sql key value storage works too and scales well

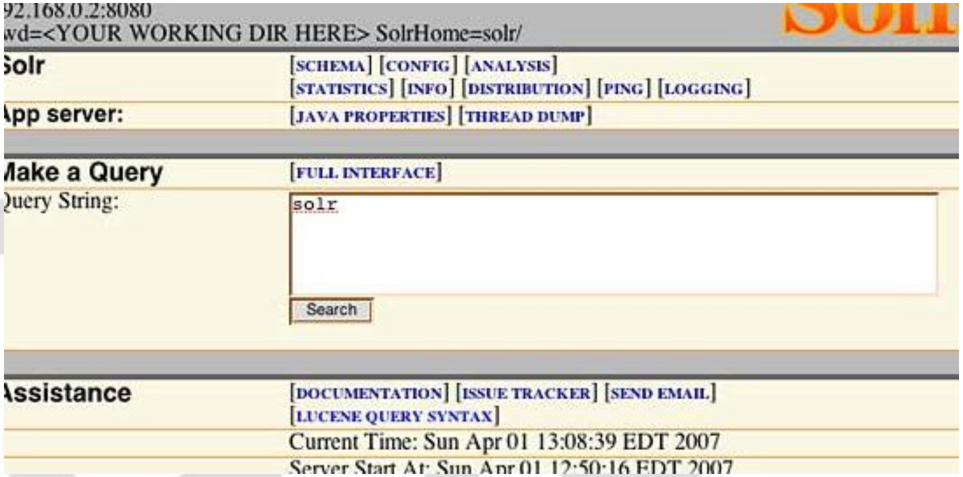
HANDS-ON: we’ve got a bit of both
On VM

Post-analysis

- Language correction (slang, misspellings etc)
- Language translation (taking chinese/russian/.. Feeds)
- Custom “synonymous” word matching
- Similarity hashing functions and more

Post analysis tools

- Hands on:
  - SOLR
  - RIAK Search

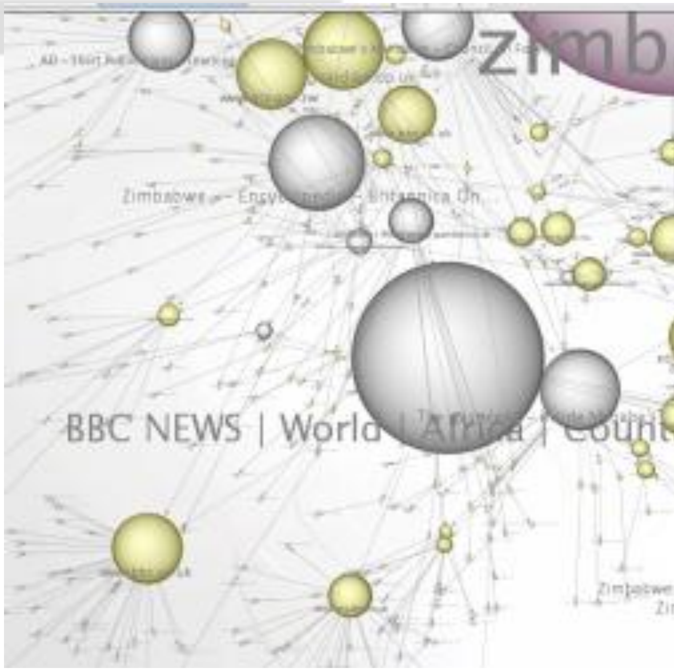


```
92.168.0.2:8080
wd=<YOUR WORKING DIR HERE> SolrHome=solr/
```

Solr	[SCHEMA] [CONFIG] [ANALYSIS] [STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING]
App server:	[JAVA PROPERTIES] [THREAD DUMP]
Make a Query	[FULL INTERFACE]
Query String:	solr Search
Assistance	[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL] [LUCENE QUERY SYNTAX]
	Current Time: Sun Apr 01 13:08:39 EDT 2007
	Server Start At: Sun Apr 01 12:50:16 EDT 2007

UI and viz

- A few tools that we are going to play with:
  - Custom web UI
  - Maltego (including building custom transforms)
  - Netglub (if have time)


Zimb
Zimbabwe – Encyclopedia Britannica On
BBC NEWS | World | Africa | Count

HANDS-ON

- So boot VM and lets get started :-)

Some codez!

yeh!!!!!!

HANDS-ON

- On your VM:
  - – Instructions in docs folder
  - – MOD01 MOD02 and MOD03 are different
  - – Sections that we are going to play with
  - – You will need internet connection and some URLz to play with. You'll get idea

Objectives

- To get the sh* working ;)
- To write some code (maybe)
- To exchange ideas
- Did I say beer? ;)

MOD01

- Doing scrappers:
  - – Nutch
    - Customization and custom plugins
    - Custom scrapping and indexing
    - Data into solr
  - – Ebot
    - Data into RIAK storage
  - – (if we have time, we’ll look into more)

MOD02

- Storage and processing:
  - SOLR (details in doc)

MOD02.2

- Key-value -> RIAK and ERL
  - – Details in doc

MOD03

- Extracting and making use of data
  - Custom UI in 3 minutes (doc #1)
  - Using maltego client to eat your data
  - Transforms - custom builds and tweaking

MOD03.2

- Netglub - opensource maltego on drugs

Other topics of interest

- NLP
- Nilsimsa hashing and applications
- Language correction algorithms

Questions?

Fygrave@o0o.nu

http://www.o0o.nu