Webis Student Presentations WS2014/15

- Argumentation Analysis in Newspaper Articles
- Morning Morality
- The Super-document
- Netspeak Query Log Analysis
- Informative Linguistic Knowledge Extraction from Wikipedia
- Elastic Search and the Clueweb
- Passphone Protocol Analysis with Avispa
- Beta Web
- SimHash as a Service: Scaling Near-Duplicate Detection
- One Class Classification of Vandalism in the Wikipedia

Modeling Information Extraction Problems using Argumentation Theory

Speakers:

Philip Drewes
Jonas Köhler

Motivation:

- Opinion mining
- Summaries of large texts
- Rating the validity of arguments in texts
- Search for arguments for a given hypothesis

⇒ We want to have a computable model of argumentation for human language.

A computational model of argumentation

nodes: **argumentative units**
(Claims, Premises)

arcs: **relations between arguments**
(Attacks, Supports, ...)

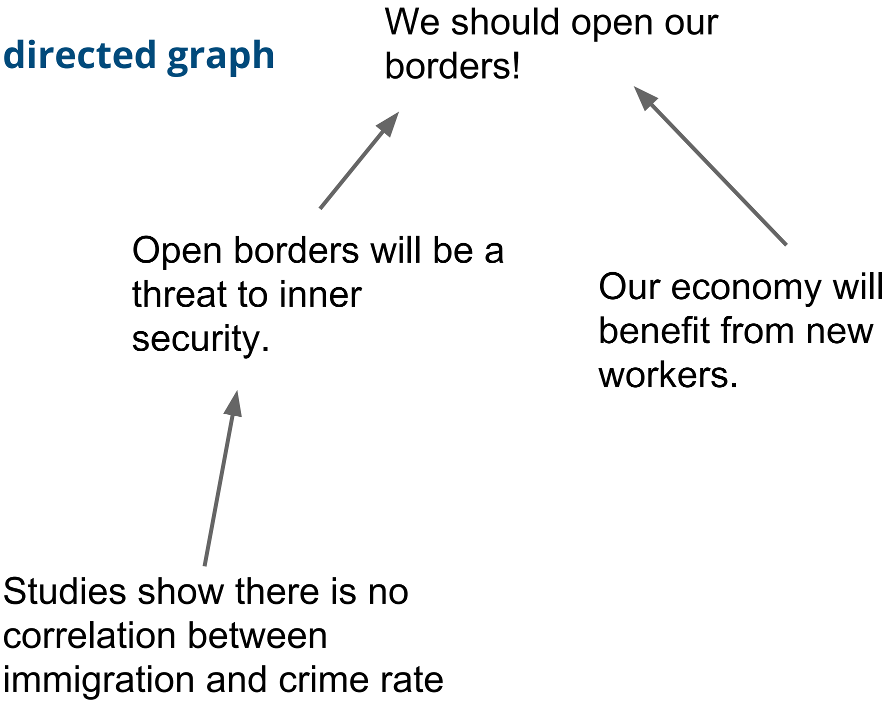
Questions:

When do arguments contradict?

How are arguments related?

What are important arguments?

directed graph



We should open our borders!

Open borders will be a threat to inner security.

Our economy will benefit from new workers.

Studies show there is no correlation between immigration and crime rate

A computational model of argumentation

Searching for arguments involves the task of detecting them

Classification:

Is a part of a text an argumentative unit?	⇒	**binary**	**{** yes, no **}**
What type of argumentative unit?	⇒	**nominal**	**{** claim, premise, … **}**
Are two argumentative units related?	⇒	**binary**	**{** yes, no **}**
What type of relation is it?	⇒	**nominal**	**{** attack, support, … **}**

…

⇒ Supervised learning problem

⇒ which features?

A computational model of argumentation

Features (mostly NLP based):

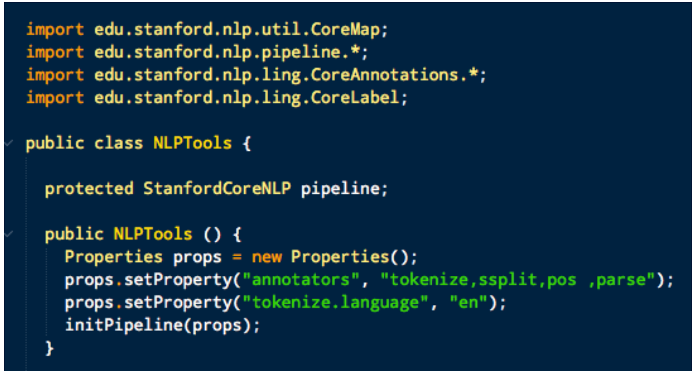
Lexical: number of punctuation marks in a part of text

Syntactic: depth of the parse tree (linguistics)

Indicators: are discourse marker present?

Contextual: number of sub clauses in the sentences around the part of interest

Heavy use of the **Stanford NLP** Java library:



```
import edu.stanford.nlp.util.CoreMap;
import edu.stanford.nlp.pipeline.*;
import edu.stanford.nlp.ling.CoreAnnotations.*;
import edu.stanford.nlp.ling.CoreLabel;

public class NLPTools {

  protected StanfordCoreNLP pipeline;

  public NLPTools () {
    Properties props = new Properties();
    props.setProperty("annotators", "tokenize,ssplit,pos ,parse");
    props.setProperty("tokenize.language", "en");
    initPipeline(props);
  }
```

⇒ training data?

⇒ human annotation!

Creating a corpus for an argument classifier

Annotation:

Humans will annotate argumentative texts by hand.

The texts are taken from online newspapers (opinion section).

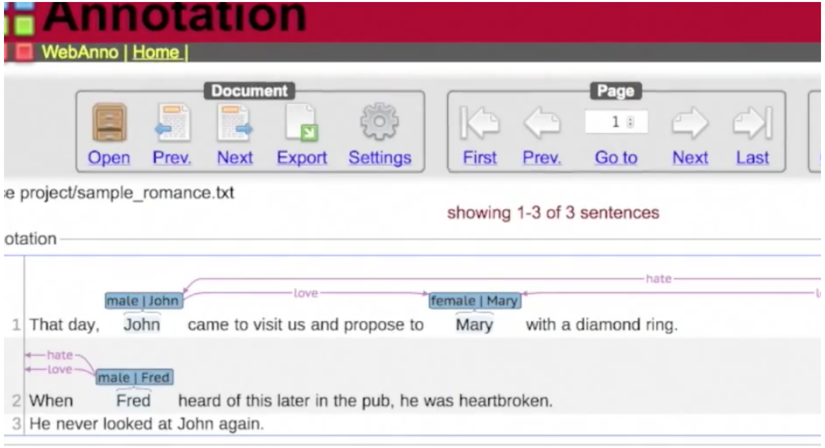
The tool for annotation is web-based. The annotations are saved to XML files.

Question:

Don’t we need 1000s of annotations?

Who will do all this work?

⇒ **Crowdsourcing!**



1 That day, John came to visit us and propose to Mary with a diamond ring.

2 When Fred heard of this later in the pub, he was heartbroken.

3 He never looked at John again.

Outlook:

What we have done so far:

- Implementing a classification framework, which is
  - Calculating the feature vectors
  - Reproducing the state of the art in classification
    - Stab et al.[1] achieve ~72% precision on an essay corpus
    - We are able to achieve ~68%
- Gathering the text data (automated web scraping)
- Designing the annotation job for the digital crowd.

[1] Stab C., Gurevych, I., Identifying Argumentative Discourse Structures in Persuasive Essays *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, p. 46-56, Association for Computational Linguistics, October 2014.

Outlook:

What we will do until February 2015 / What may come in the long term

- Let the crowd annotate our texts and build the training corpus

- Add additional features and improve the classification
  - Extend the model? Refine the classification?

- Analyze the data
  - Which questions may arise?

- Search for argument components
  - Only possible if there is a good model + classification

Thank you for your listening!

Questions?

Morning Morality on the Web

Webis presentation

2014-12-18

Morning Morality on the Web

-

foundation

Project foundation and discussion starter:

- Kouchaki, Maryam, and Isaac H. Smith. "The Morning Morality Effect The Influence of Time of Day on Unethical Behavior." Psychological science 25 (2013):95-102

Content:

- People's ethical behaviour is changing throughout the day.
- There is a „self-regulatory“ resource, which depletes the longer someone is behaving good.
- Therefore, a person is more likely to cheat and lie in the afternoon or evening than in the morning.

Morning Morality on the Web

-

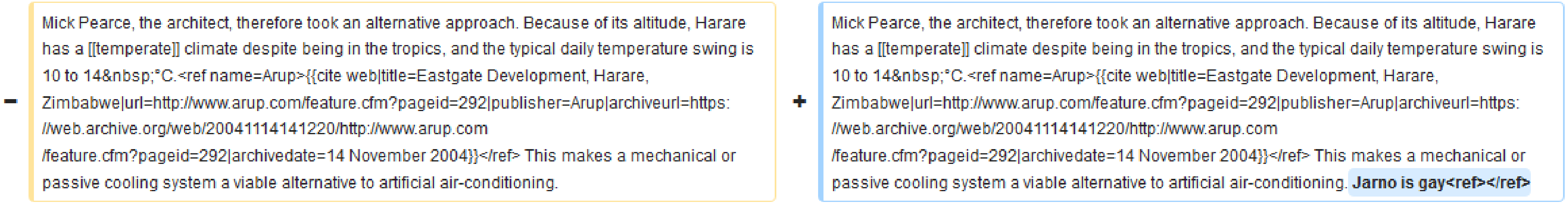
previous work

Is such a phenomenon measurable on the Web?

- In an effort to show such an effect, Wikipedia-Vandalism cases where analyzed.

What is Wikipedia-Vandalism?

Inappropriate change, addition or removal of Wikipedia content, like adding irrelevant, abusive words, deleting pages or purposely adding false information.



−

Mick Pearce, the architect, therefore took an alternative approach. Because of its altitude, Harare has a [[temperate]] climate despite being in the tropics, and the typical daily temperature swing is 10 to 14 °C.<ref name=Arup>{{cite web|title=Eastgate Development, Harare, Zimbabwe|url=http://www.arup.com/feature.cfm?pageid=292|publisher=Arup|archiveurl=https://web.archive.org/web/20041114141220/http://www.arup.com/feature.cfm?pageid=292|archivedate=14 November 2004}}</ref> This makes a mechanical or passive cooling system a viable alternative to artificial air-conditioning.

+

Mick Pearce, the architect, therefore took an alternative approach. Because of its altitude, Harare has a [[temperate]] climate despite being in the tropics, and the typical daily temperature swing is 10 to 14 °C.<ref name=Arup>{{cite web|title=Eastgate Development, Harare, Zimbabwe|url=http://www.arup.com/feature.cfm?pageid=292|publisher=Arup|archiveurl=https://web.archive.org/web/20041114141220/http://www.arup.com/feature.cfm?pageid=292|archivedate=14 November 2004}}</ref> This makes a mechanical or passive cooling system a viable alternative to artificial air-conditioning. **Jarno is gay<ref></ref>**

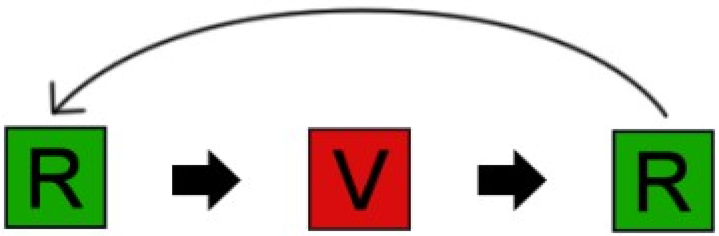
Morning Morality on the Web

-

Wikipedia Vandalism

How to get Wikipedia-Vandalism data?

- Scan through the history of edits for a Simple Vandalism Pattern.
- A revert back to a revision befor an edit(V) is most often a case of vandalism.
- Detection is done by users and bots.



Morning Morality on the Web

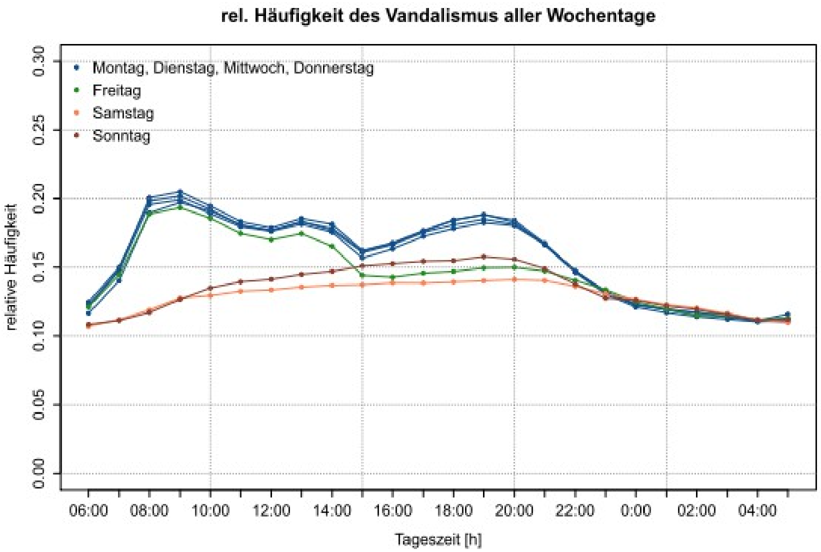
-

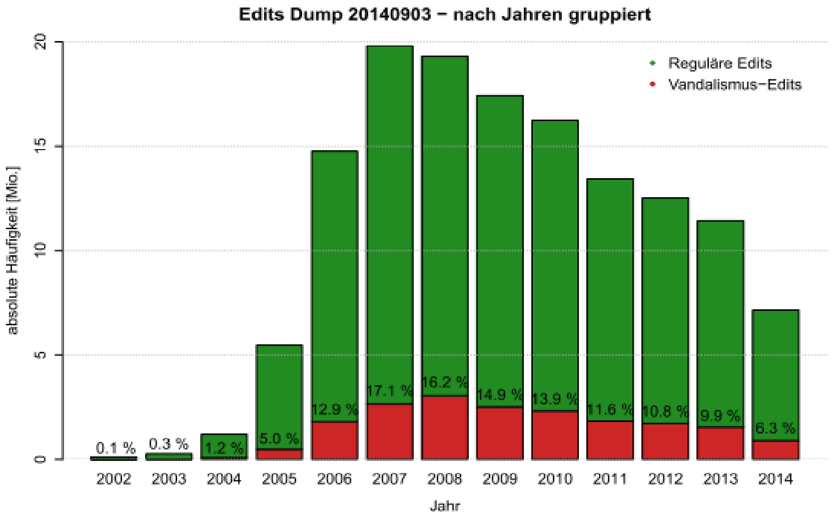
previous work

Finding the „Morning Morality Effect“ in Wikipedia-Vandalism data.

Work of the previous project group:

- Analyzing correlations between local time and vandalism.
- Geolocation of vandal - IP addresses for local edit time.


rel. Häufigkeit des Vandalismus aller Wochentage


Edits Dump 20140903 – nach Jahren gruppiert

Morning Morality on the Web

-

current work

Finding more correlation between bad behaviour on the Web and exogenous/external factors, e.g. , weather, time and region.

What we have done so far/are working on:

- Geolocate the given vandalism and normal edit dumps of the United States for 2013.
- Correleated them with the NOAA National Weather Service data (hourly weather data from 1.700 weather stations in the US over the last 15 years).

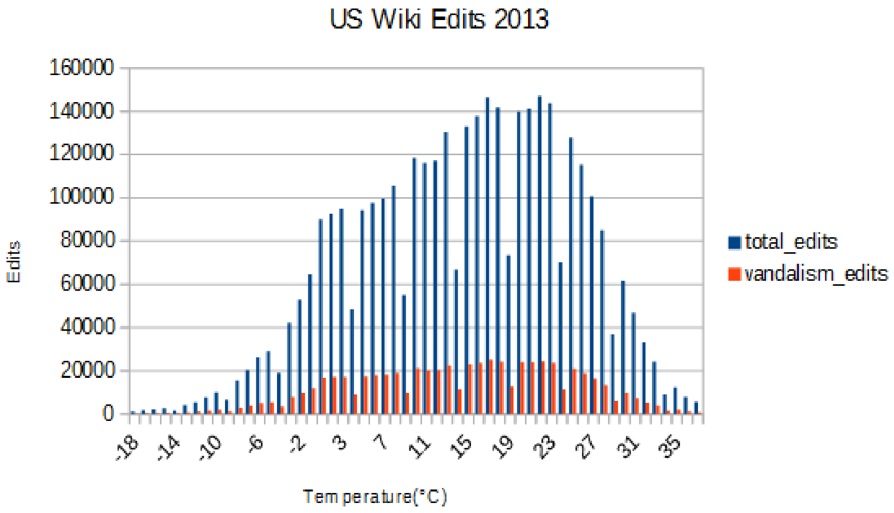
Morning Morality on the Web

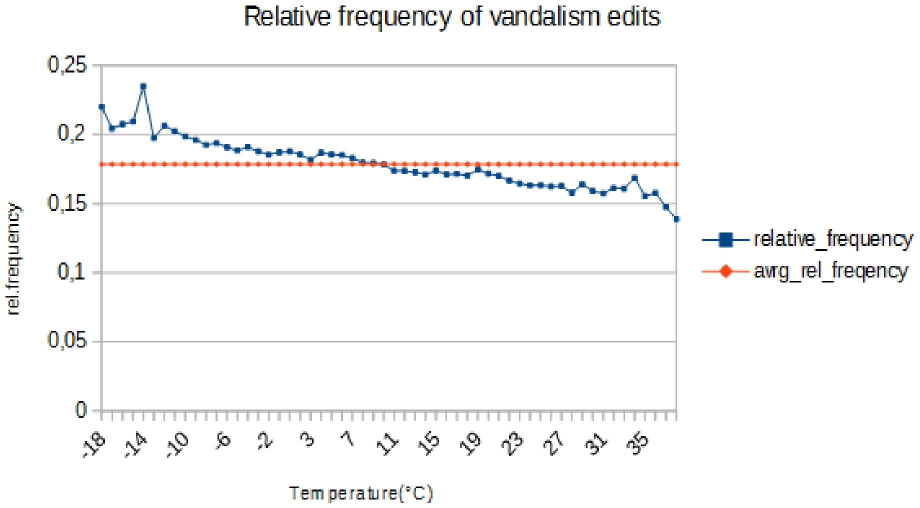
-

current work

Early data -

Work still in progress :





Morning Morality on the Web

-

future work

- Analyze data for different Climate Zones and weather effects like rain and snow.

- Changing vandalism frequency in correlation with weather over time, e.g., annual and monthly time periods

- Different locations: comparisons of different states, rural and metropolitan areas.

The Super Document

A Result Presentation Paradigm for Exploratory Search Tasks

Participants:
Kevin Reinartz, Janek Bevendorff,
Kristof Komlossy, Carsten Tetens, Sebastian Gottschlich

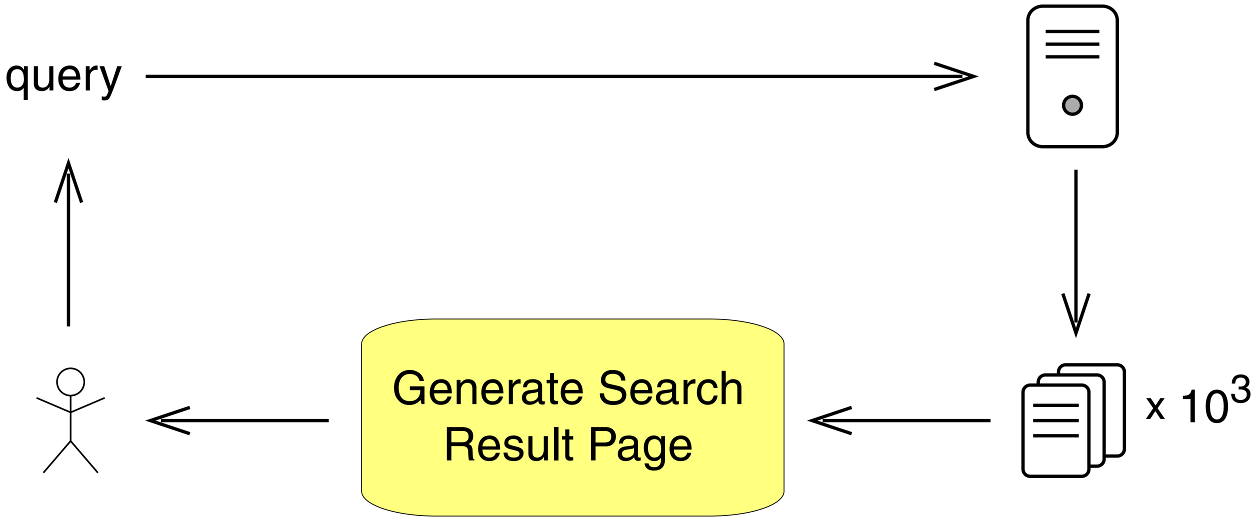
Tim Gollub Michael Völske Benno Stein

Web Technology & Information Systems
Bauhaus-Universität Weimar

Winter Term 2014/15

Project: SuperSERP

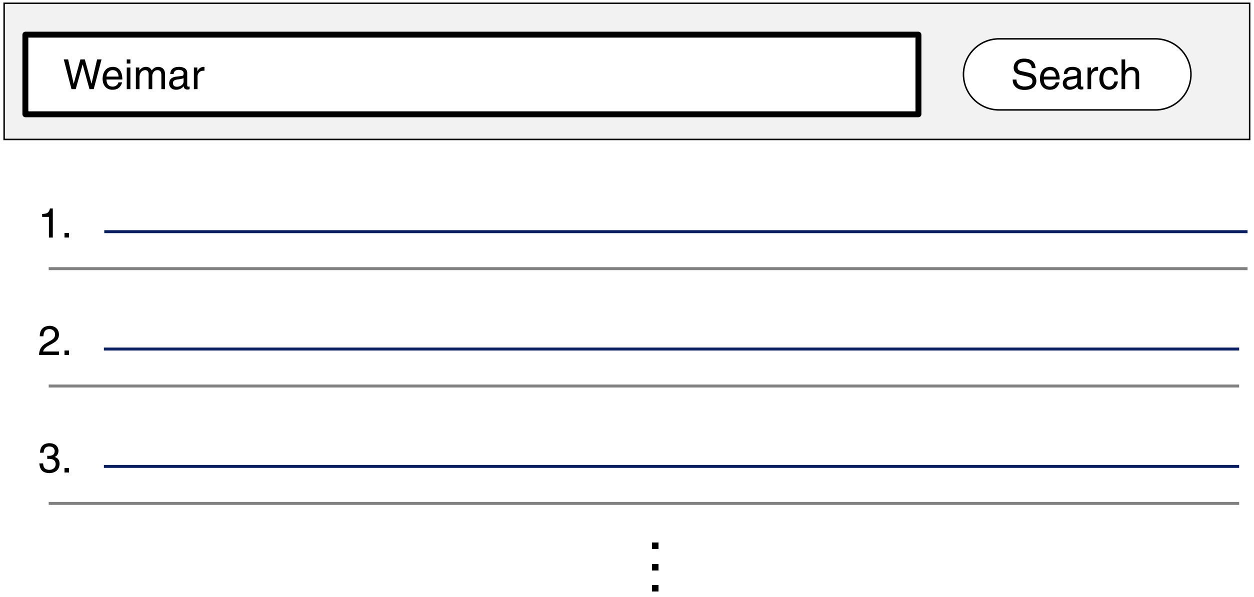
Web Search: Problem Description



Given the relevant documents for a query, how to present them to the user?

Project: SuperSERP

Traditional Presentation Paradigm: Ranked Result Lists



- Compile a list of document descriptions linking to the original resources.
- Order based on the likelihood that a document contains relevant information.

Project: SuperSERP

General

- ❑ Alternative result presentation paradigms for open or undirected informational queries

- ❑ General Approach: Increase accessibility of resources in the limit of a search result list.

- ❑ Observation: An effective domain independent paradigm is hard to find.

- ❑ We concentrate on two applications:
  - Related Work Search
  - City Search

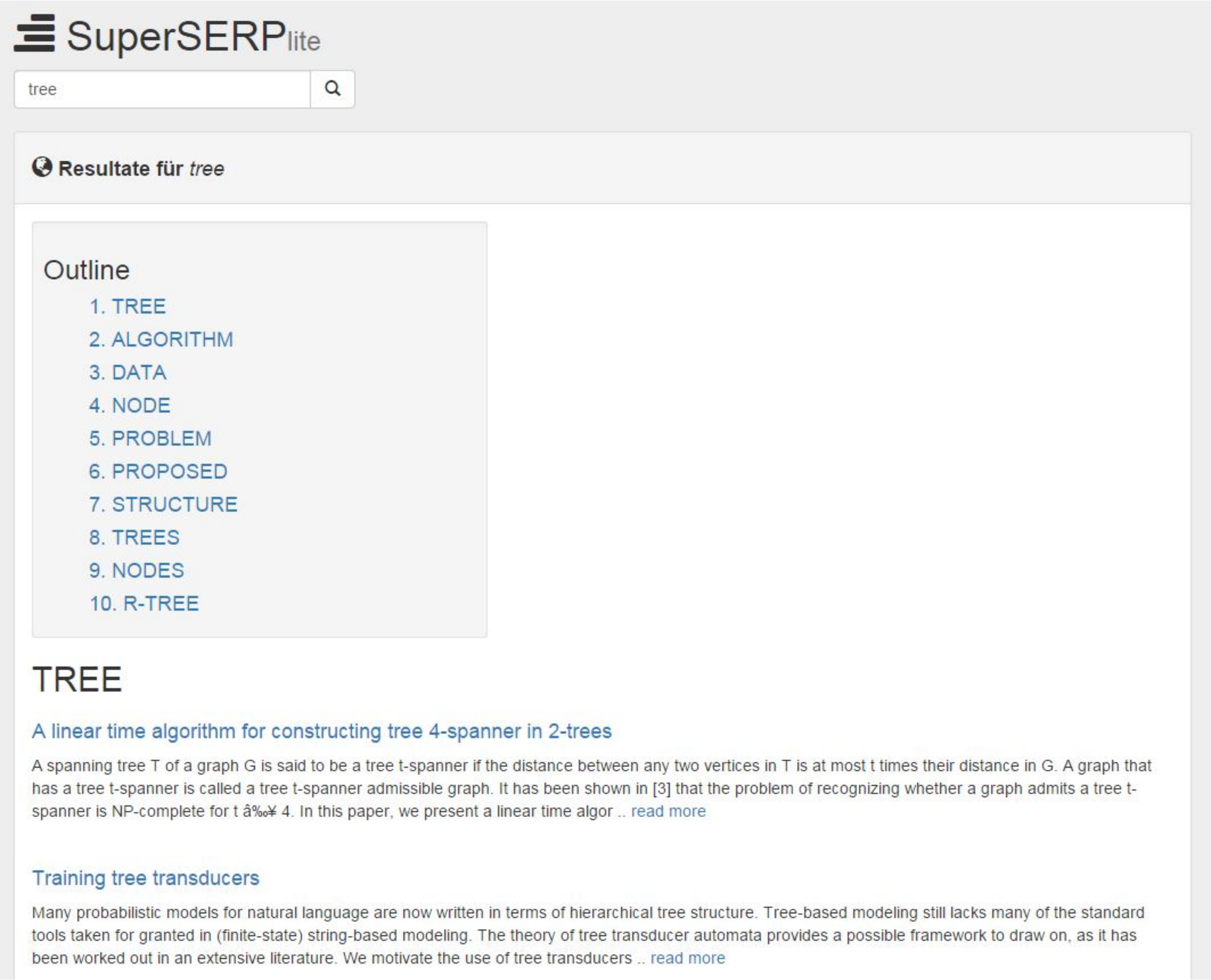
Application 1: Related Work Search

Current State

- LUCENE Index of webis-csp corpus (approx. 177.000 papers)
- Keyphrase extraction (KpMinerExtractor from aitools)
- MUSTACHE Template-Engine for search result presentation
- Search result based on keyphrases (currently)

Application 1: Related Work Search

Current user interface



Resultate für *tree*

Outline

1. TREE
2. ALGORITHM
3. DATA
4. NODE
5. PROBLEM
6. PROPOSED
7. STRUCTURE
8. TREES
9. NODES
10. R-TREE

TREE

A linear time algorithm for constructing tree 4-spanner in 2-trees

A spanning tree T of a graph G is said to be a tree t-spanner if the distance between any two vertices in T is at most t times their distance in G. A graph that has a tree t-spanner is called a tree t-spanner admissible graph. It has been shown in [3] that the problem of recognizing whether a graph admits a tree t-spanner is NP-complete for t â‰¥ 4. In this paper, we present a linear time algor .. read more

Training tree transducers

Many probabilistic models for natural language are now written in terms of hierarchical tree structure. Tree-based modeling still lacks many of the standard tools taken for granted in (finite-state) string-based modeling. The theory of tree transducer automata provides a possible framework to draw on, as it has been worked out in an extensive literature. We motivate the use of tree transducers .. read more

Application 1: Related Work Search

Future Work

- Improved user interaction:
  - Query by document
  - Manual “topic points”
  - Quality of clustering statistics
- Topic Model for Indexing & keyquery compositing
- Efficient clustering algorithm for outline generating

Application 2: City Search

Current State

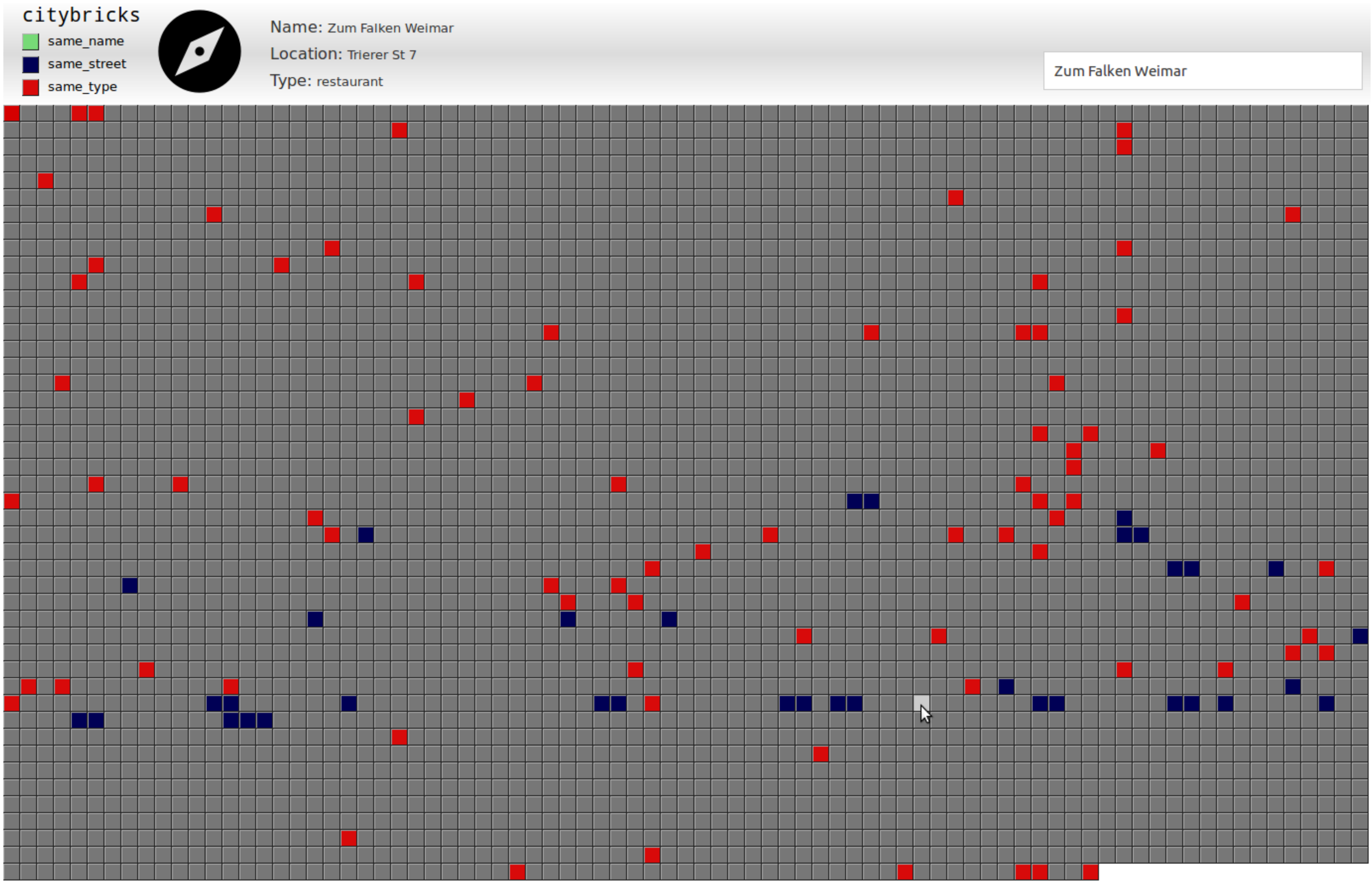
- collected Google Places
- using *Bigdata* as triple store (replacing *Fuseki*)
- read Google Places as RDF triples into triple store
- generated random people at random locations

CityBricks:

- each place is a brick
- sorted from north to south
- highlight on search & similarity

Application 2: City Search

CityBricks



Name: Zum Falken Weimar

Location: Trierer St 7

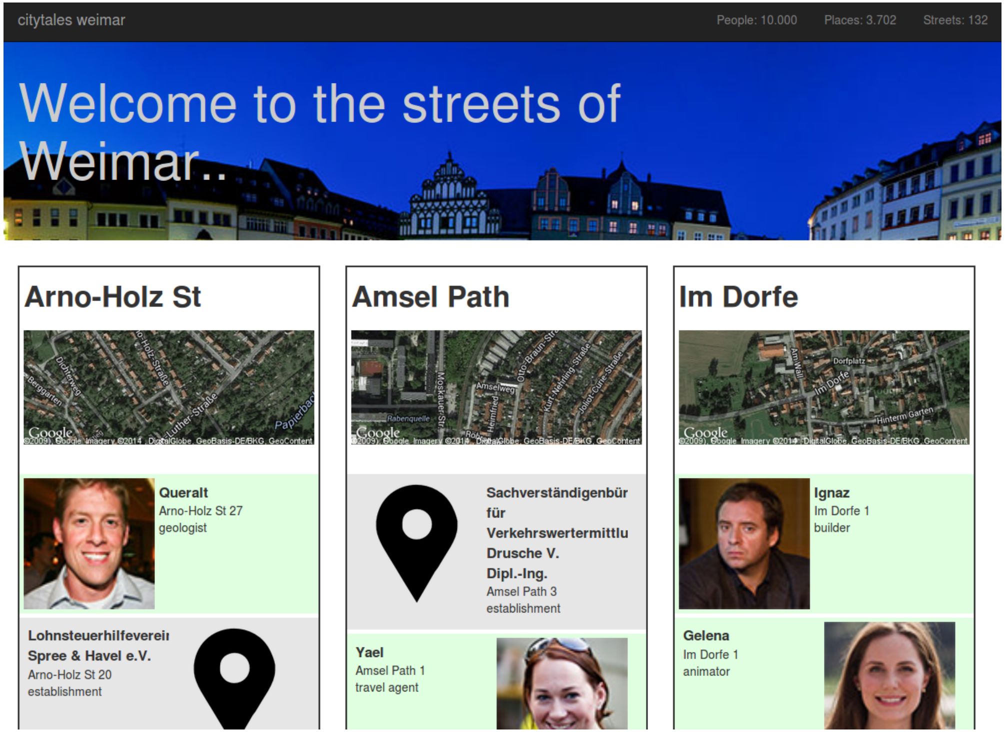
Type: restaurant

Zum Falken Weimar

Application 2: City Search

CityTales

- take the user on a journey through the city
- create a mashup using content & statistics
- streets from a city + Random users and locations
- from various sources (Google Places, Flickr ...)



citytales weimar People: 10.000 Places: 3.702 Streets: 132

Welcome to the streets of Weimar..

Arno-Holz St

Queralt
Arno-Holz St 27
geologist

Lohnsteuerhilfevereir Spree & Havel e.V.
Arno-Holz St 20
establishment

Amsel Path

Sachverständigenbür für Verkehrswertermittlu Drusche V. Dipl.-Ing.
Amsel Path 3
establishment

Yael
Amsel Path 1
travel agent

Im Dorfe

Ignaz
Im Dorfe 1
builder

Gelena
Im Dorfe 1
animator

Application 2: City Search

Future Work

- Improve storytelling, infographic inspired UI
- Add sources like news, official statistics, social network, reviews
- Use focused crawling (Heritrix) to obtain web pages related to Weimar.

Netspeak Query Log Analysis

Amir Othman

cvs:

```
code-in-progress/webisstud/wstud-netspeak-analysis
code-in-progress/webisstud/wstud-netspeak-analysis-query-detection
code-in-progress/webisstud/wstud-netspeak-analysis-query-browser

data-in-progress/wstud-netspeak-analysis
```

Netspeak

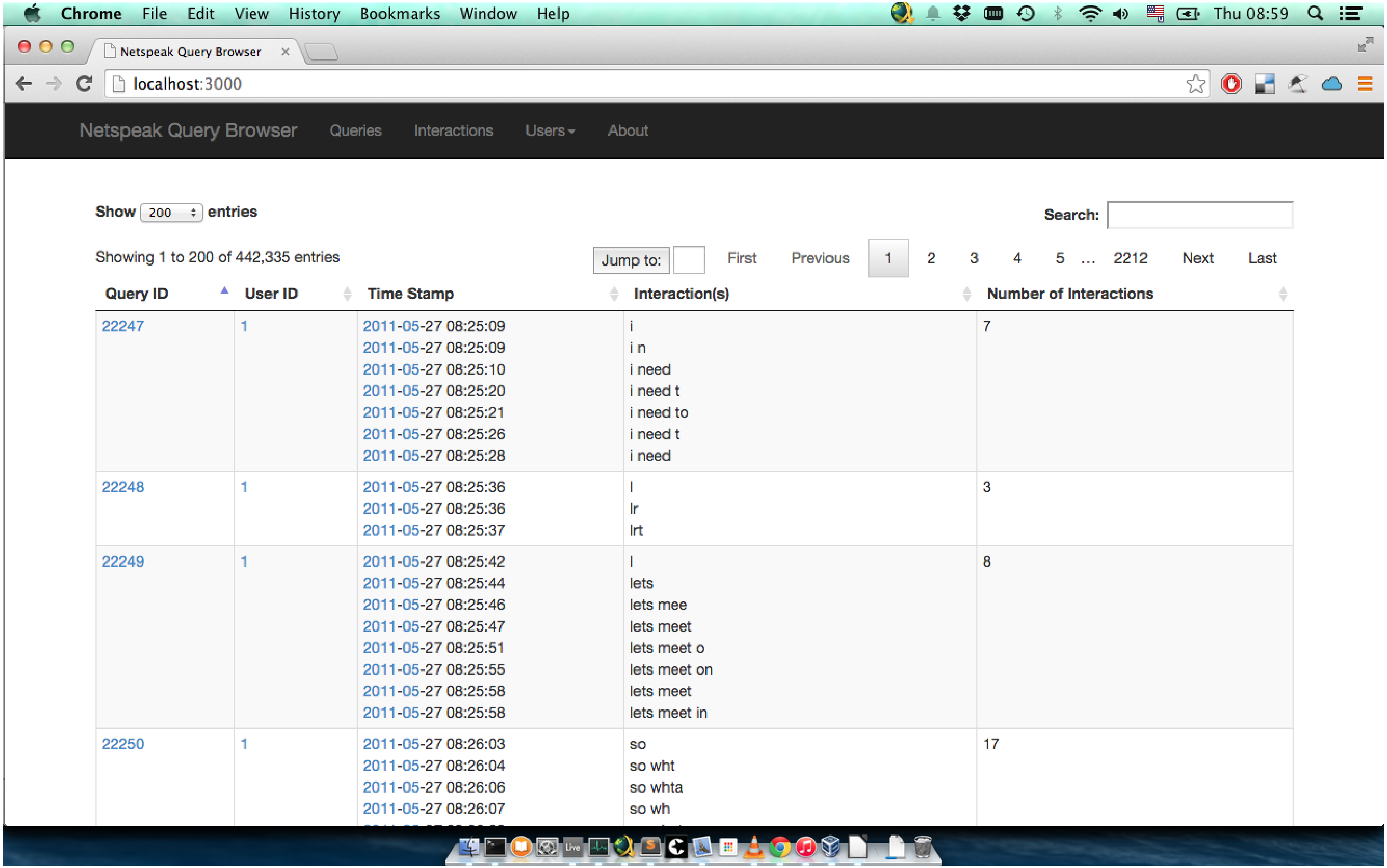
- Service to check usage of words
- ~2000 Users a month
- Log from March 2009 to February 2014

Query Detection

- Decision Tree, using log from 100 different IPs as groundtruth
- Features: overlapping characters, term overlap, character Jaccard coefficient, trigram character cosine similarity, Levenshtein distance, timegap

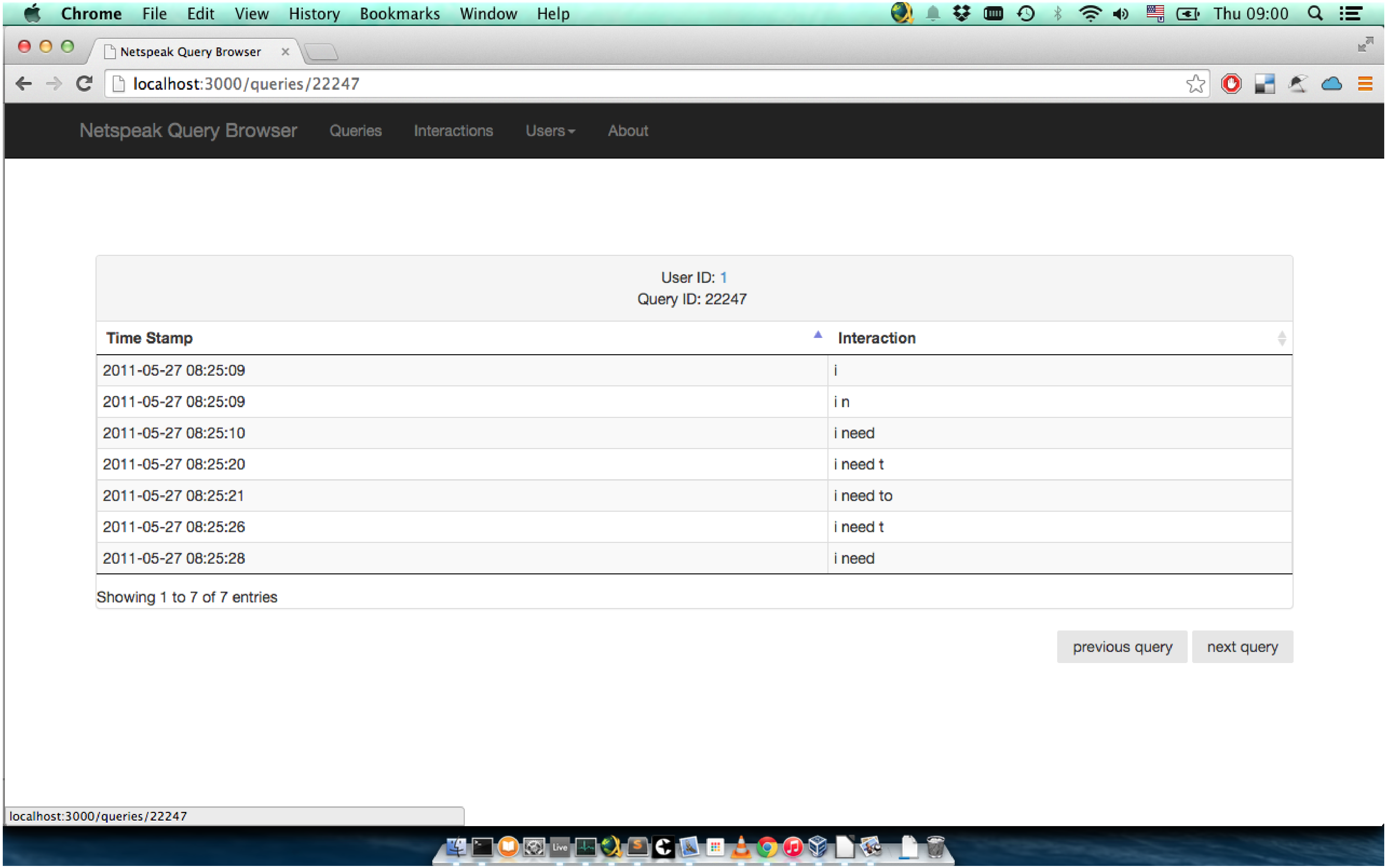
Netspeak Query Log Browser

- Facilitate analysis – added visualizations and interlinking
- Exploring
- Add Notes



Showing 1 to 200 of 442,335 entries

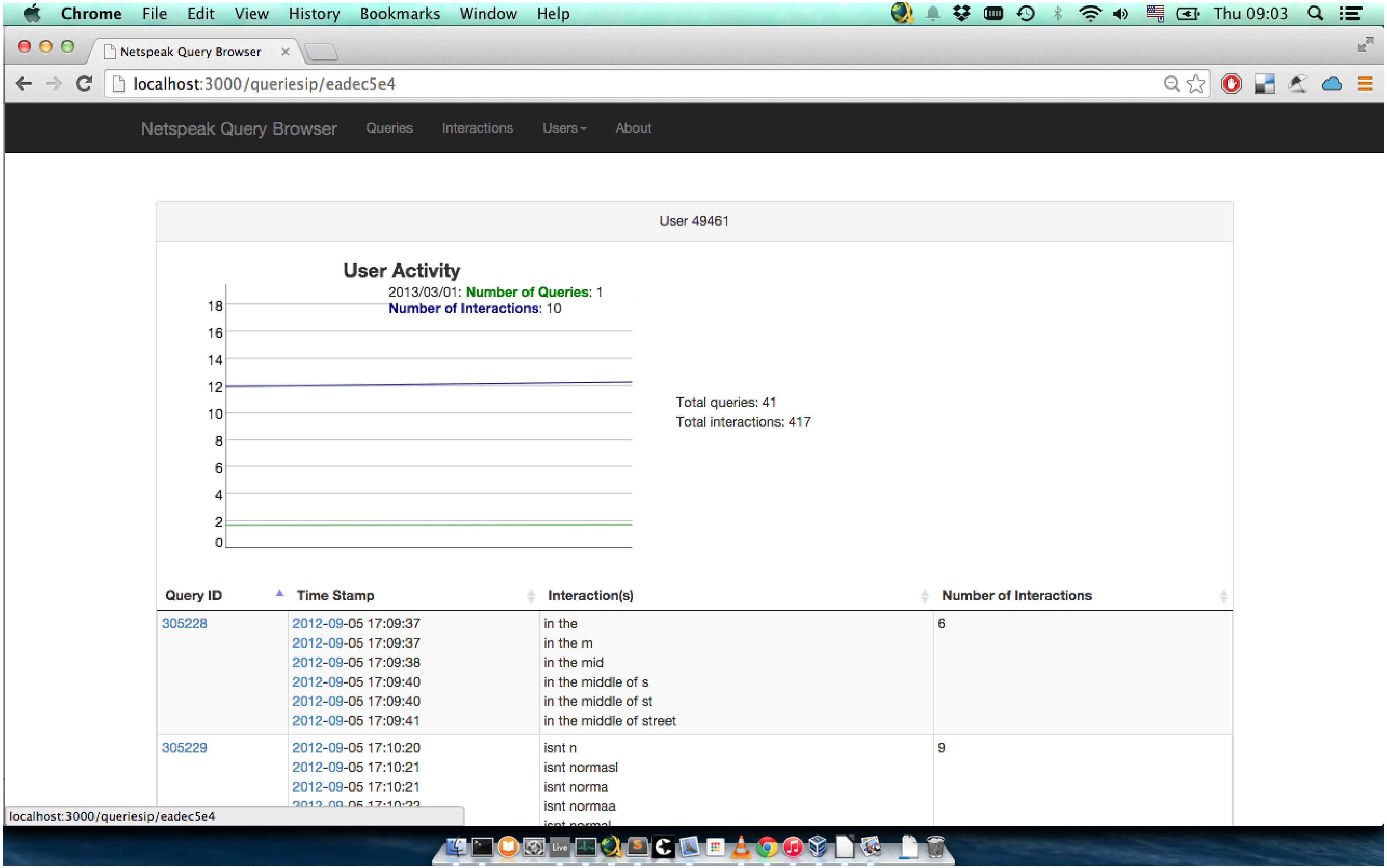
Query ID	User ID	Time Stamp	Interaction(s)	Number of Interactions
22247	1	2011-05-27 08:25:09 2011-05-27 08:25:09 2011-05-27 08:25:10 2011-05-27 08:25:20 2011-05-27 08:25:21 2011-05-27 08:25:26 2011-05-27 08:25:28	i i n i need i need t i need to i need t i need	7
22248	1	2011-05-27 08:25:36 2011-05-27 08:25:36 2011-05-27 08:25:37	l lr lrt	3
22249	1	2011-05-27 08:25:42 2011-05-27 08:25:44 2011-05-27 08:25:46 2011-05-27 08:25:47 2011-05-27 08:25:51 2011-05-27 08:25:55 2011-05-27 08:25:58 2011-05-27 08:25:58	l lets lets mee lets meet lets meet o lets meet on lets meet lets meet in	8
22250	1	2011-05-27 08:26:03 2011-05-27 08:26:04 2011-05-27 08:26:06 2011-05-27 08:26:07	so so wht so whta so wh	17



User ID: 1
Query ID: 22247

Time Stamp	Interaction
2011-05-27 08:25:09	i
2011-05-27 08:25:09	i n
2011-05-27 08:25:10	i need
2011-05-27 08:25:20	i need t
2011-05-27 08:25:21	i need to
2011-05-27 08:25:26	i need t
2011-05-27 08:25:28	i need

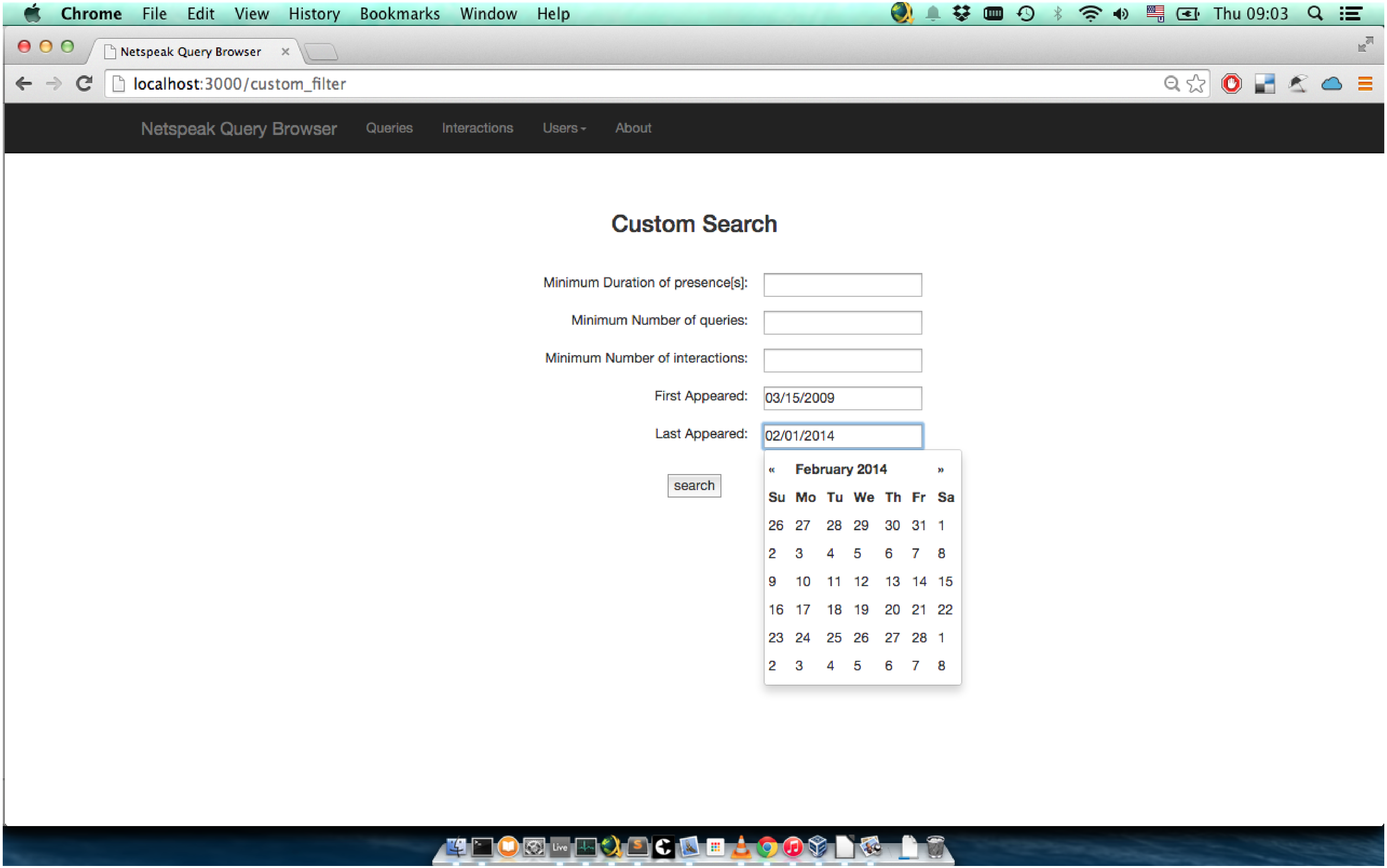
Showing 1 to 7 of 7 entries


Chrome File Edit View History Bookmarks Window Help
Thu 09:03
Netspeak Query Browser
localhost:3000/queriesip/eadec5e4
Netspeak Query Browser
Queries
Interactions
Users
About
User Activity
18
16
14
12
10
8
6
4
2
0
2013/03/01: Number of Queries: 1
Number of Interactions: 10
localhost:3000/queriesip/eadec5e4

User 49461

Total queries: 41
Total interactions: 417

Query ID	Time Stamp	Interaction(s)	Number of Interactions
305228	2012-09-05 17:09:37 2012-09-05 17:09:37 2012-09-05 17:09:38 2012-09-05 17:09:40 2012-09-05 17:09:40 2012-09-05 17:09:41	in the in the m in the mid in the middle of s in the middle of st in the middle of street	6
305229	2012-09-05 17:10:20 2012-09-05 17:10:21 2012-09-05 17:10:21 2012-09-05 17:10:22	isnt n isnt normasl isnt norma isnt normaa isnt normal	9



Custom Search

Minimum Duration of presence[s]:

Minimum Number of queries:

Minimum Number of interactions:

First Appeared: 03/15/2009

Last Appeared: 02/01/2014

« February 2014 »

Su	Mo	Tu	We	Th	Fr	Sa
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	1
2	3	4	5	6	7	8

Ideas

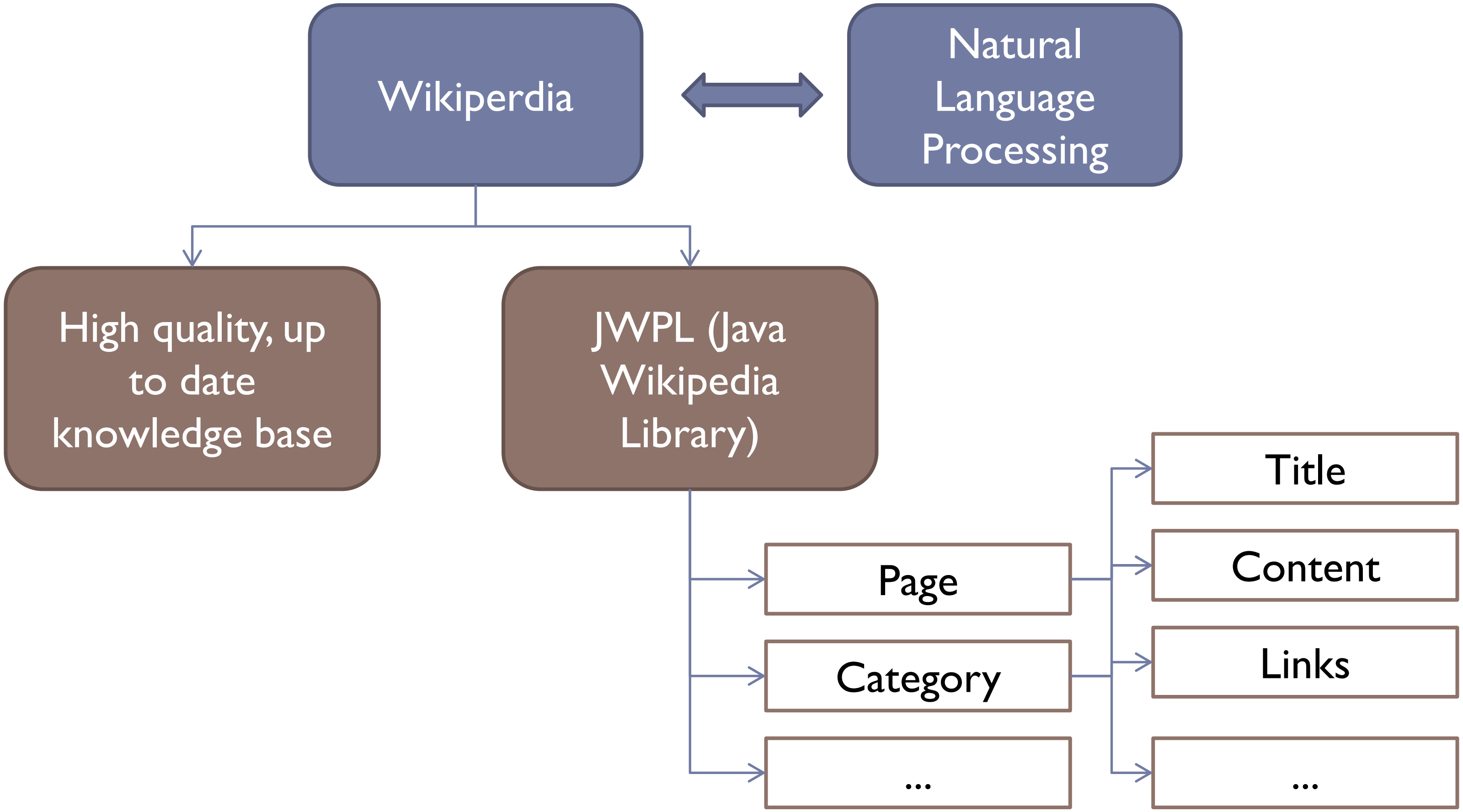
- Learning effect
- Identifiable user

Informative Linguistic Knowledge Extraction from Wikipedia

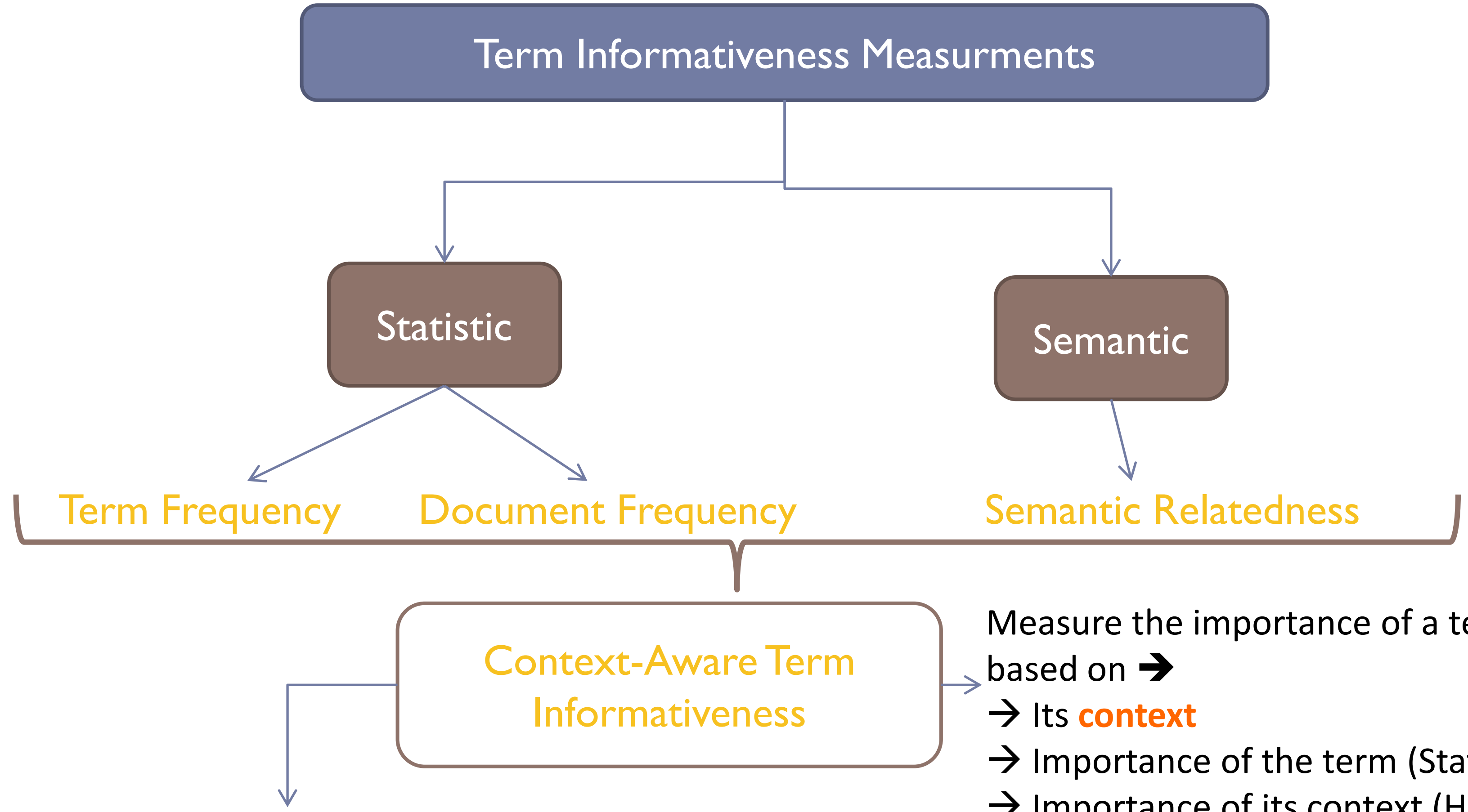
Roxanne El Baff (1st Semester CSM Student)
Supervisior :Khalid El Khatib

Bauhaus-Universität Weimar

Wikipidea and JWPL


Wikiperdia
Natural Language Processing
High quality, up to date knowledge base
JWPL (Java Wikipedia Library)
Page
Category
…
Title
Content
Links
…

Measuring Term Informativeness



Term Informativeness Measurments

Statistic

Semantic

Term Frequency

Document Frequency

Semantic Relatedness

Context-Aware Term Informativeness

$$I(t, c_i) = \sum_{c_j \in U_f(t)} \kappa(c_i, c_j) \cdot CA(c_j)$$

Measure the importance of a term based on ➔

→ Its **context**

→ Importance of the term (Statistic)

→ Importance of its context (How strong is the relation between term and context (Semantic Relatedness)

Elasticsearch and the Clueweb

A Work-in-Progress Presentation

Janek Bevendorff

Web Technology & Information Systems
Bauhaus-Universität Weimar

Elasticsearch and the Clueweb

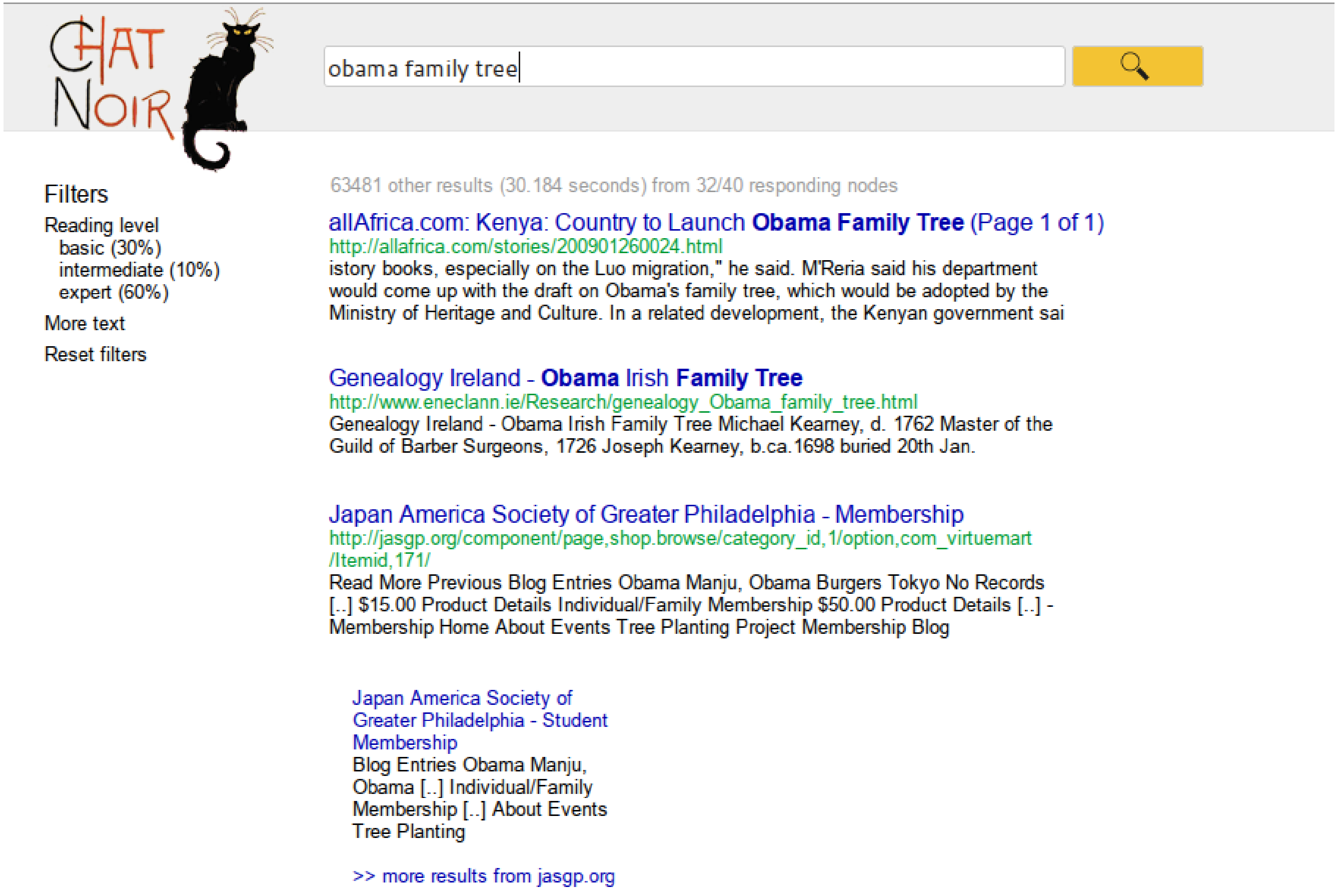
What is the Clueweb?

some data:

- web crawl of 1,040,809,705 documents
- 5TB of compressed data (25TB uncompressed)
- 4,780,950,903 unique URLs
- tons of spam

Elasticsearch and the Clueweb

And what do we do with it?



obama family tree

Filters

Reading level
- basic (30%)
- intermediate (10%)
- expert (60%)

More text

Reset filters

63481 other results (30.184 seconds) from 32/40 responding nodes

allAfrica.com: Kenya: Country to Launch **Obama Family Tree** (Page 1 of 1)
http://allafrica.com/stories/200901260024.html
istory books, especially on the Luo migration," he said. M'Reria said his department would come up with the draft on Obama's family tree, which would be adopted by the Ministry of Heritage and Culture. In a related development, the Kenyan government sai

Genealogy Ireland - **Obama** Irish **Family Tree**
http://www.eneclann.ie/Research/genealogy_Obama_family_tree.html
Genealogy Ireland - Obama Irish Family Tree Michael Kearney, d. 1762 Master of the Guild of Barber Surgeons, 1726 Joseph Kearney, b.ca.1698 buried 20th Jan.

Japan America Society of Greater Philadelphia - Membership
http://jasgp.org/component/page,shop.browse/category_id,1/option,com_virtuemart/Itemid,171/
Read More Previous Blog Entries Obama Manju, Obama Burgers Tokyo No Records [..] $15.00 Product Details Individual/Family Membership $50.00 Product Details [..] - Membership Home About Events Tree Planting Project Membership Blog

Japan America Society of Greater Philadelphia - Student Membership
Blog Entries Obama Manju, Obama [..] Individual/Family Membership [..] About Events Tree Planting

>> more results from jasgp.org

Elasticsearch and the Clueweb

New backend: Elasticsearch

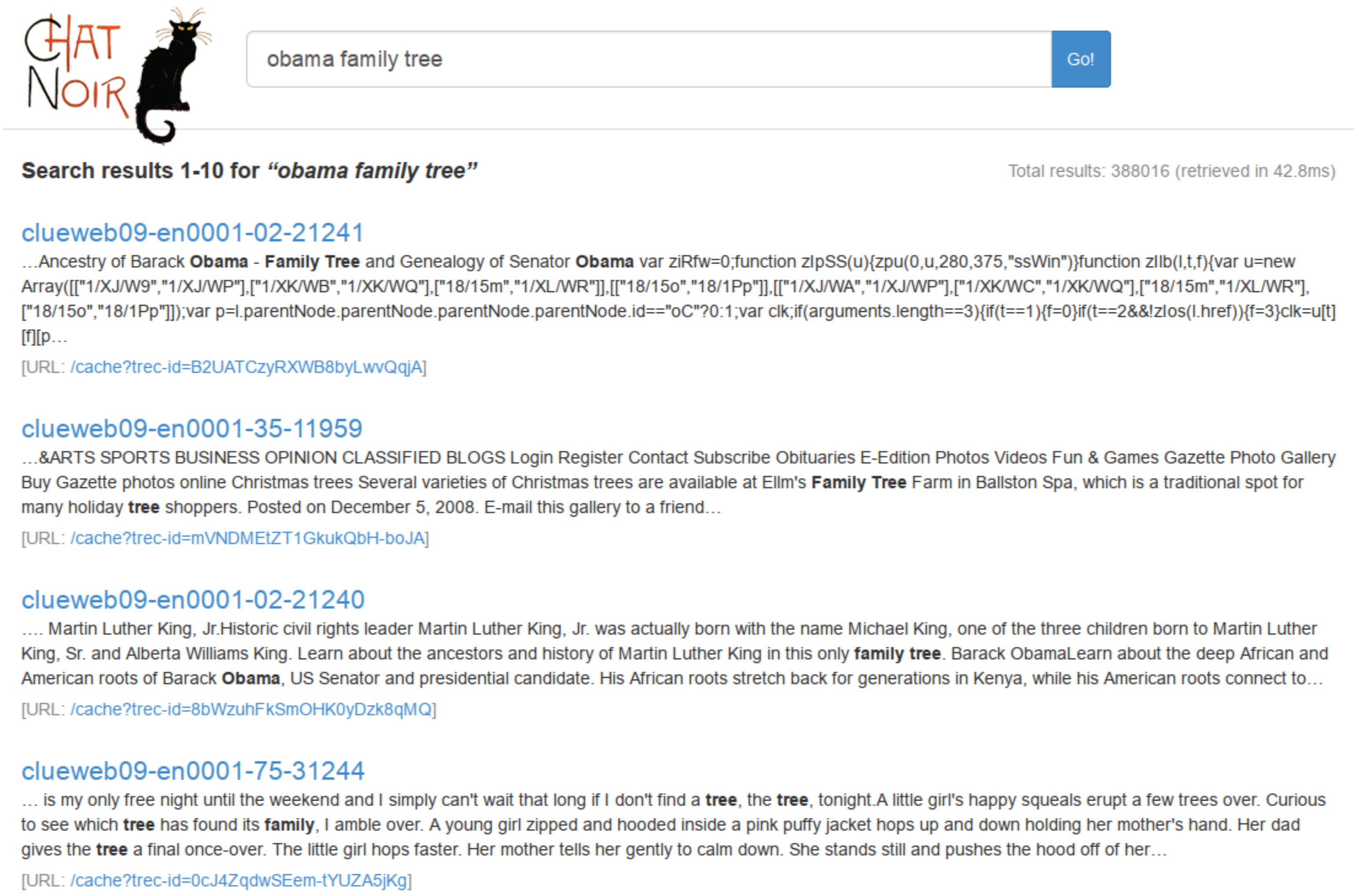
Elasticsearch is a. . .

- ❑ distributed and redundant Lucene index
- ❑ (RESTful) search server

Optionally, Elasticsearch comes with Hadoop integration for indexing large amounts of data and performing real-time search on HDFS clusters.

Elasticsearch and the Clueweb

Chatnoir 2



Search results 1-10 for *“obama family tree”*

Total results: 388016 (retrieved in 42.8ms)

clueweb09-en0001-02-21241

...Ancestry of Barack **Obama** - **Family Tree** and Genealogy of Senator **Obama** var ziRfw=0;function zIpSS(u){zpu(0,u,280,375,"ssWin")}function zIIb(l,t,f){var u=new Array([["1/XJ/W9","1/XJ/WP"],["1/XK/WB","1/XK/WQ"],["18/15m","1/XL/WR"]],[["18/15o","18/1Pp"]],[["1/XJ/WA","1/XJ/WP"],["1/XK/WC","1/XK/WQ"],["18/15m","1/XL/WR"],["18/15o","18/1Pp"]]);var p=l.parentNode.parentNode.parentNode.parentNode.id=="oC"?0:1;var clk;if(arguments.length==3){if(t==1){f=0}if(t==2&&!zIos(l.href)){f=3}clk=u[t][f][p...

[URL: /cache?trec-id=B2UATCzyRXWB8byLwvQqjA]

clueweb09-en0001-35-11959

...&ARTS SPORTS BUSINESS OPINION CLASSIFIED BLOGS Login Register Contact Subscribe Obituaries E-Edition Photos Videos Fun & Games Gazette Photo Gallery Buy Gazette photos online Christmas trees Several varieties of Christmas trees are available at Ellm's **Family Tree** Farm in Ballston Spa, which is a traditional spot for many holiday **tree** shoppers. Posted on December 5, 2008. E-mail this gallery to a friend...

[URL: /cache?trec-id=mVNDMEtZT1GkukQbH-boJA]

clueweb09-en0001-02-21240

.... Martin Luther King, Jr.Historic civil rights leader Martin Luther King, Jr. was actually born with the name Michael King, one of the three children born to Martin Luther King, Sr. and Alberta Williams King. Learn about the ancestors and history of Martin Luther King in this only **family tree**. Barack ObamaLearn about the deep African and American roots of Barack **Obama**, US Senator and presidential candidate. His African roots stretch back for generations in Kenya, while his American roots connect to...

[URL: /cache?trec-id=8bWzuhFkSmOHK0yDzk8qMQ]

clueweb09-en0001-75-31244

... is my only free night until the weekend and I simply can't wait that long if I don't find a **tree**, the **tree**, tonight.A little girl's happy squeals erupt a few trees over. Curious to see which **tree** has found its **family**, I amble over. A young girl zipped and hooded inside a pink puffy jacket hops up and down holding her mother's hand. Her dad gives the **tree** a final once-over. The little girl hops faster. Her mother tells her gently to calm down. She stands still and pushes the hood off of her...

[URL: /cache?trec-id=0cJ4ZqdwSEem-tYUZA5jKg]

Elasticsearch and the Clueweb

Future Work

- index the whole ClueWeb12 and ClueWeb09 datasets on our brandnew Betaweb cluster
- use more fields (title, URL, anchor texts etc.) for weighted search
- some more frontend magic

Elasticsearch and the Clueweb

Thank you for your attention!

Passphone Protocol Analysis with Avispa

André Karge

Bauhaus-Universität Weimar

17. Dezember 2014

Agenda

1. Passphone Protocol

2. AVISPA

Passphone Protocol

- Protocol for two factor authentication at a service provider
- Factors:
  - Password as usual
  - Smartphone
- User enters his password
- Gets a QR-Code in return
- Scans the QR-Code with his registered smartphone app
- After success the user is logged in

Passphone Protocol

- Protocol for two factor authentication at a service provider
- Factors:
  - Password as usual
  - Smartphone
- User enters his password
- Gets a QR-Code in return
- Scans the QR-Code with his registered smartphone app
- After success the user is logged in

- In protocol: several communications with different parties:
  - Service Provider (e.g. Facebook, Ebay, Amazon, ...)
  - Trusted Third Party Server
  - User at a browser
  - User at his smartphone
- Communication save?

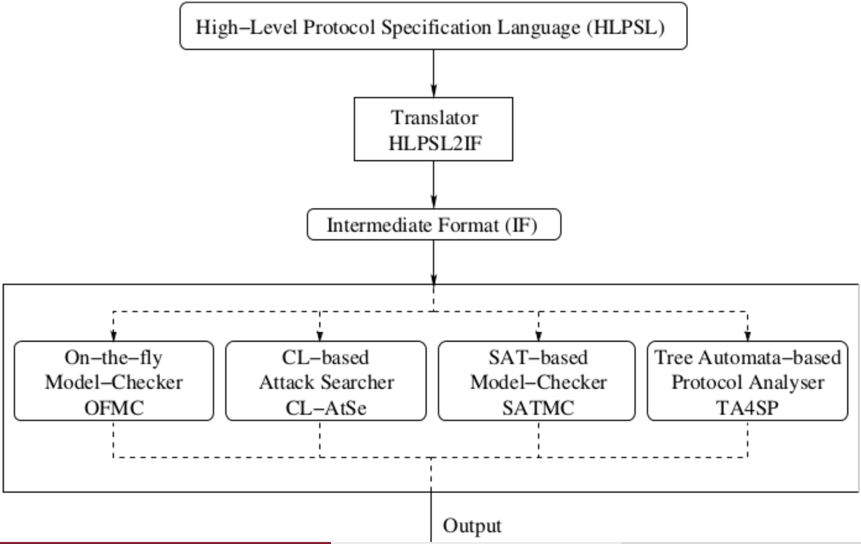
AVISPA

- Approach: automatic proofing of the protocol with AVISPA
- AVISPA = Automated Validation of Internet Security Protocols and Applications
- Protocol has to be translated into special language HLPSL
- HLPSL = High Level Protocol Specification Language

AVISPA

- Approach: automatic proofing of the protocol with AVISPA
- AVISPA = Automated Validation of Internet Security Protocols and Applications
- Protocol has to be translated into special language HLPSL
- HLPSL = High Level Protocol Specification Language

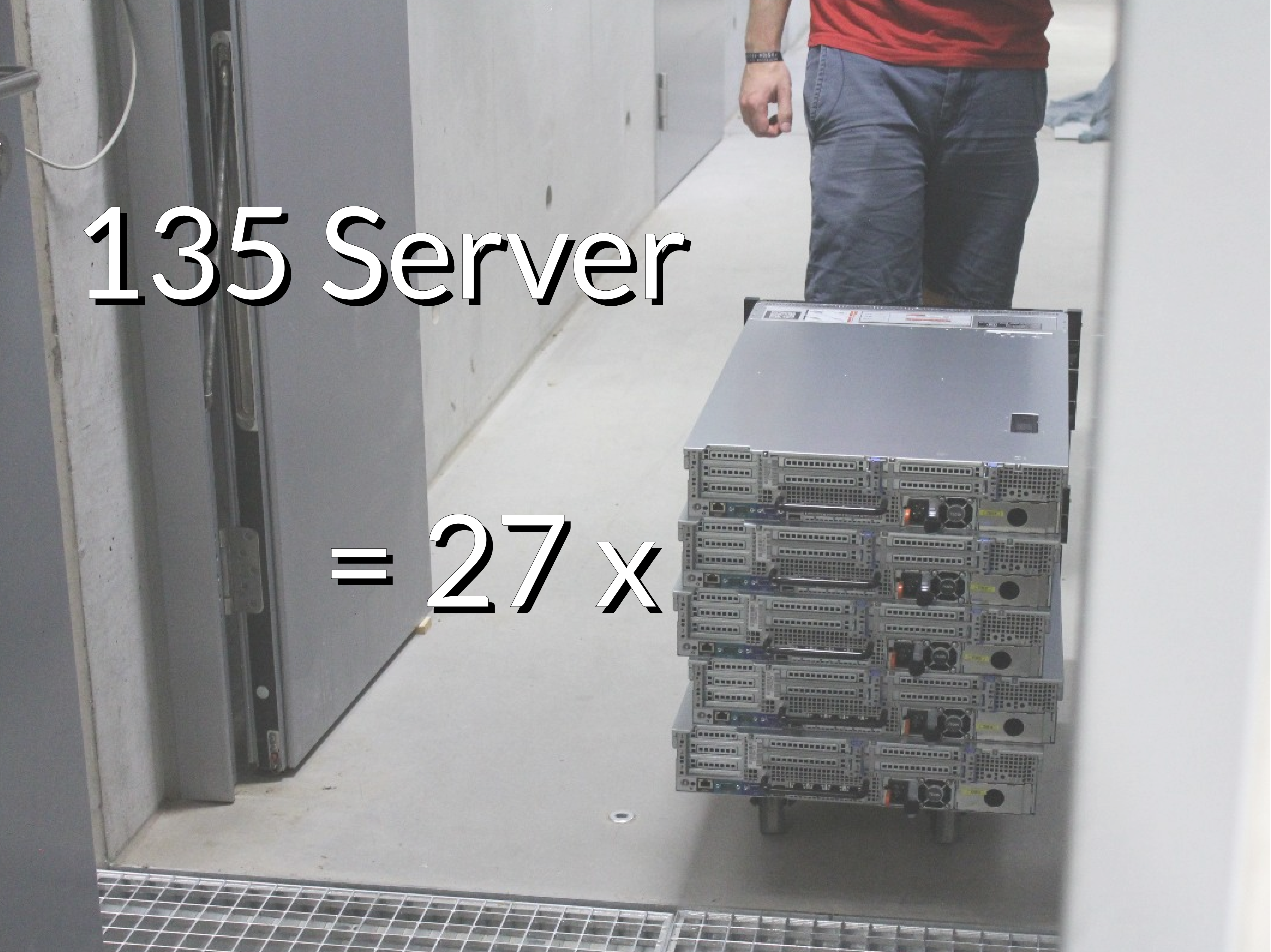
AVISPA Function



- Possible to choose the proofer
- output afeter proofing depends on criterias set in the hlspl file
- (e.g. security of a nonce)
- Proofer checks if the given protocol is safe or if not
- If a Protocol is not safe the proofer gives an attack trace

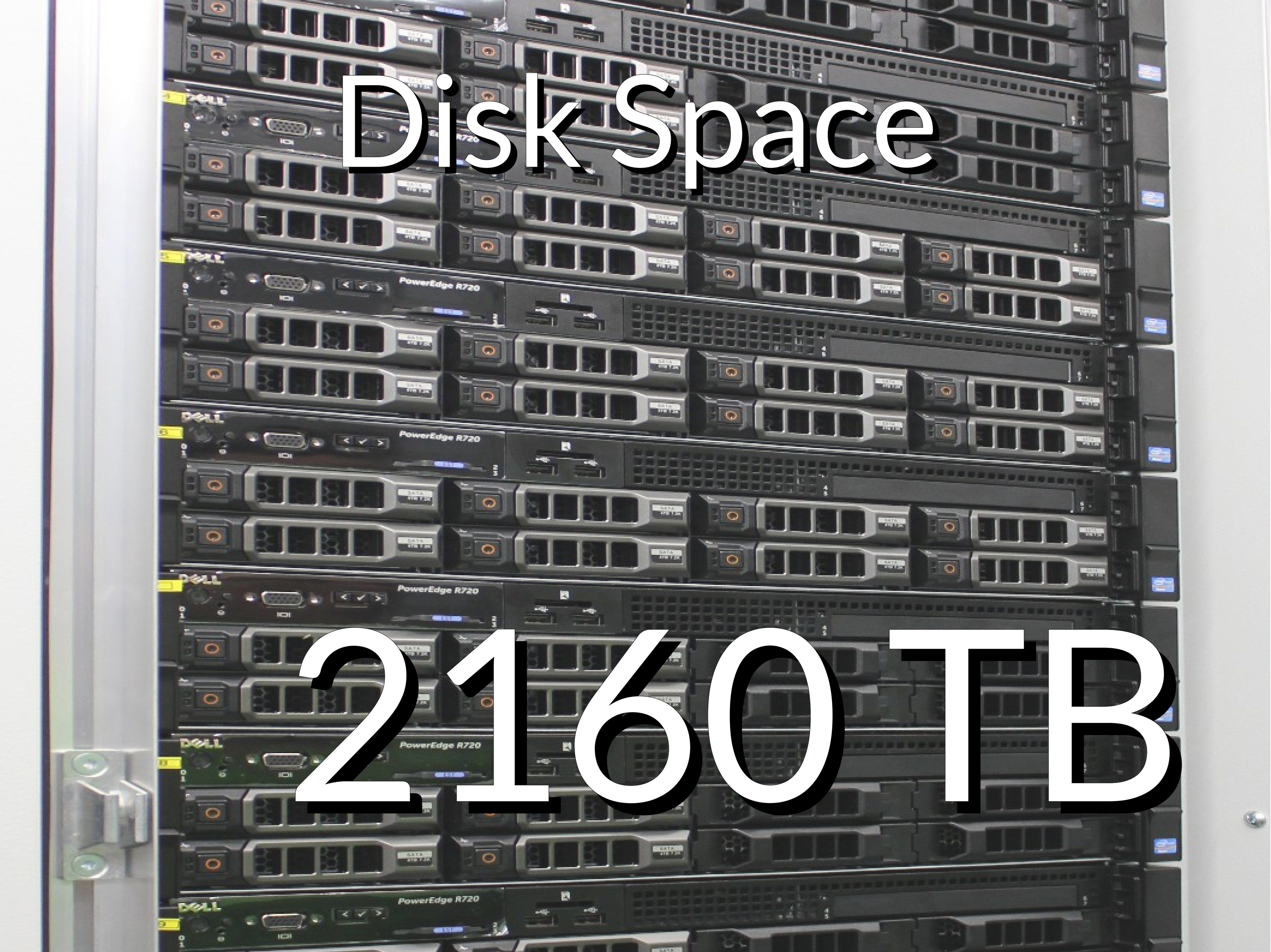
betaweb

janek.bevendorff@uni-weimar.de
alexander.herr@uni-weimar.de
martin.tippmann@uni-weimar.de



135 Server

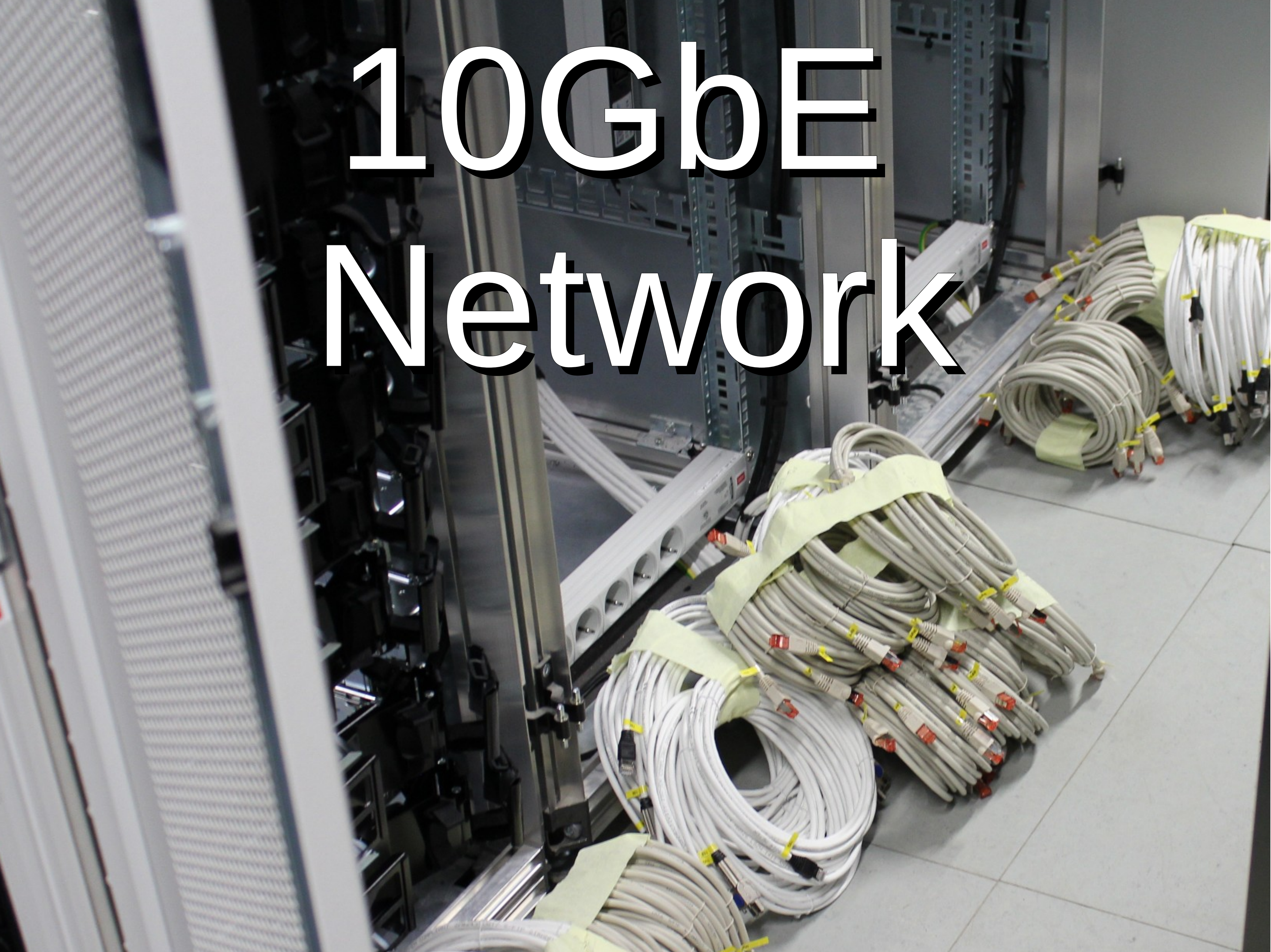
= 27 x

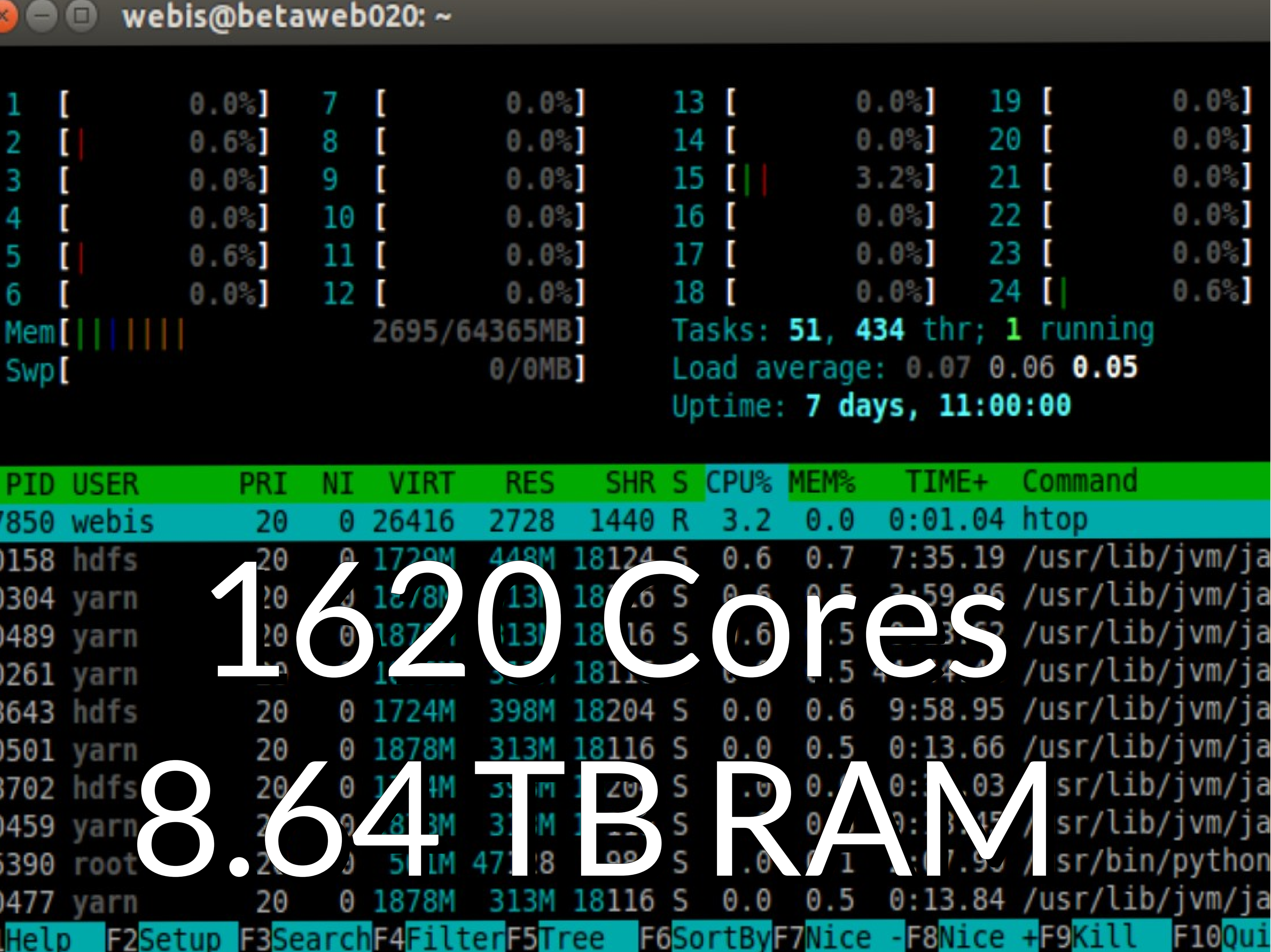


Disk Space

2160 TB

10GbE Network



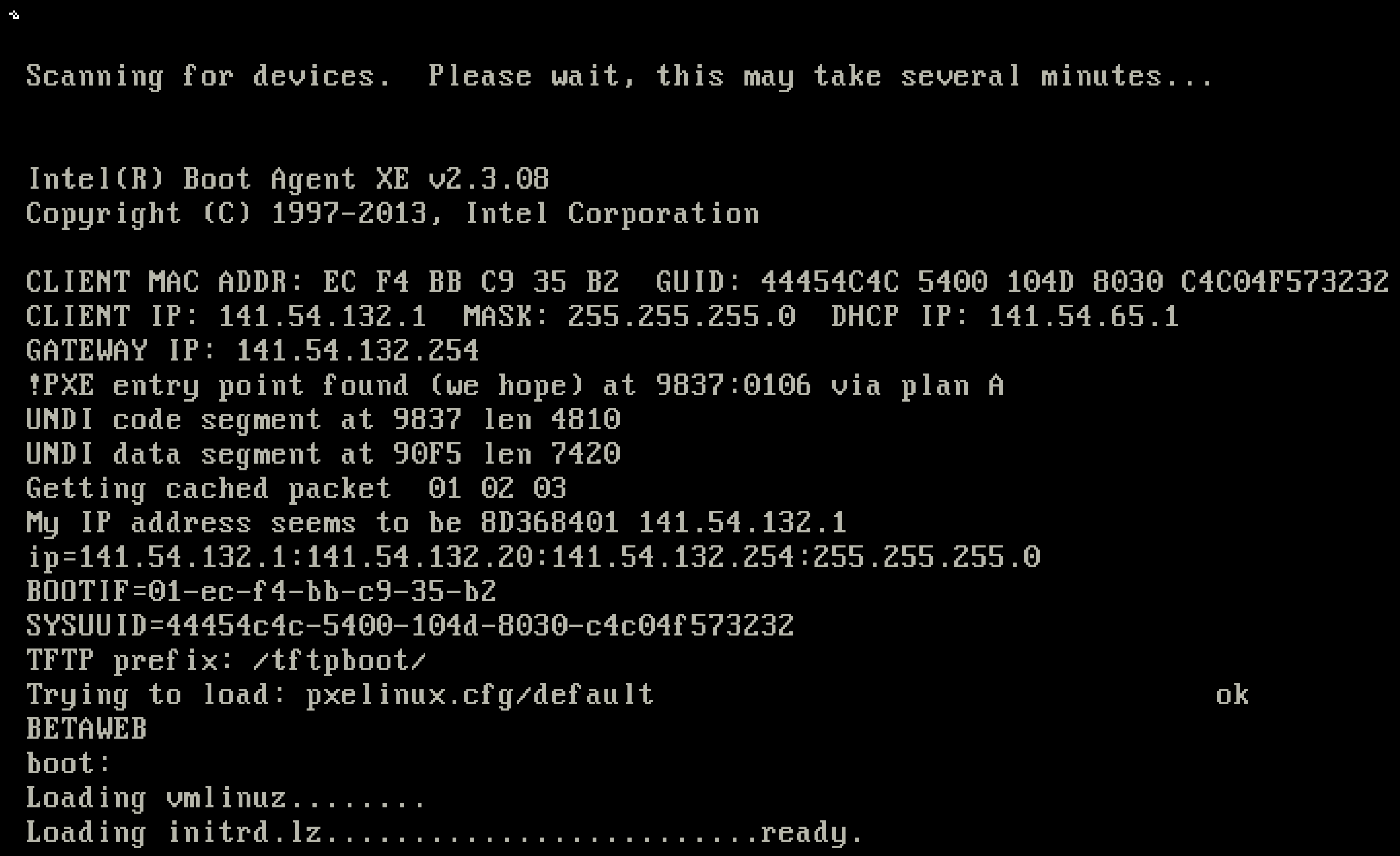


1620 Cores

8.64 TB RAM

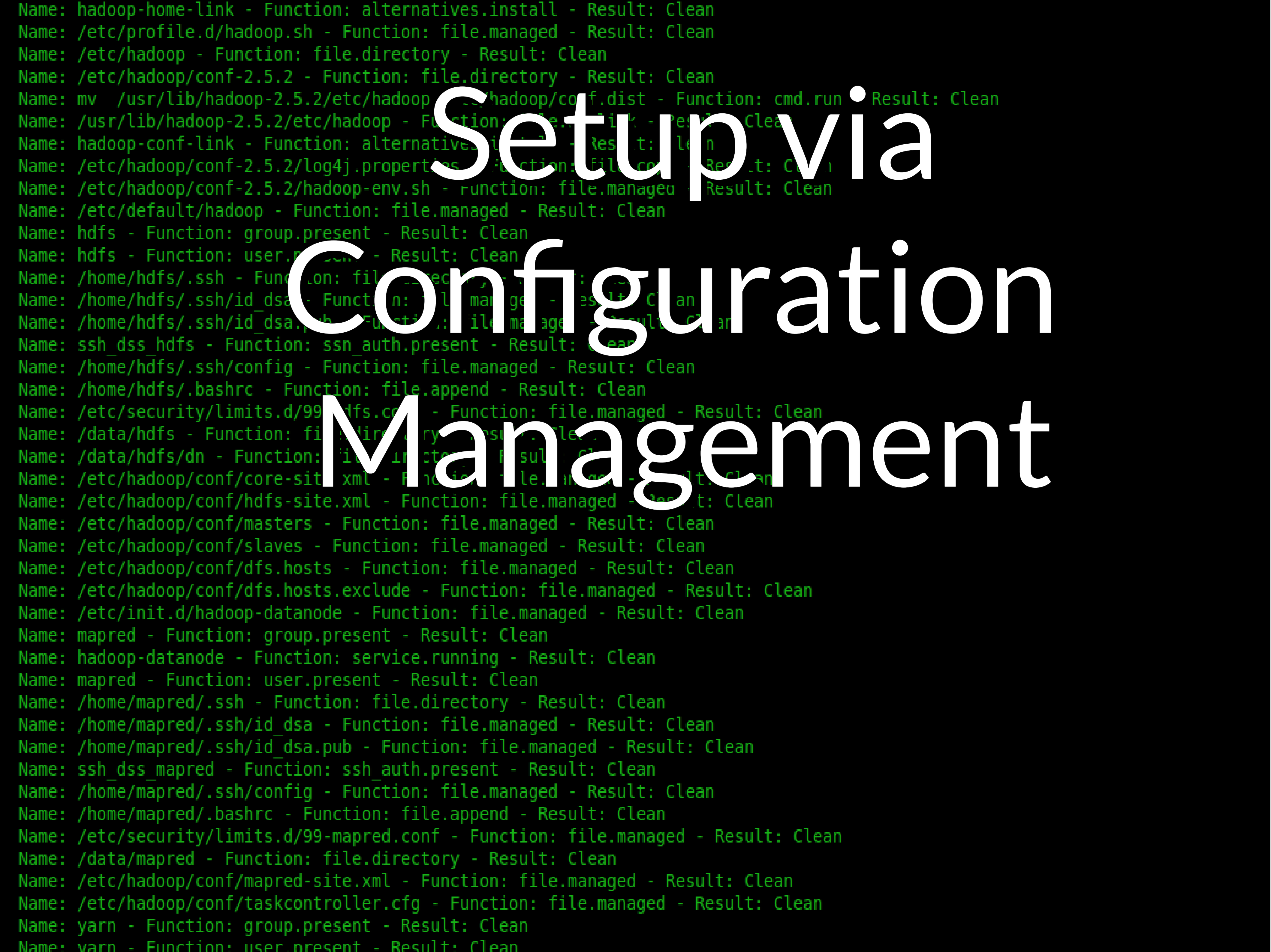
```
webis@betaweb020: ~

1  [          0.0%]   7  [          0.0%]   13 [          0.0%]   19 [          0.0%]
2  [|         0.6%]   8  [          0.0%]   14 [          0.0%]   20 [          0.0%]
3  [          0.0%]   9  [          0.0%]   15 [||        3.2%]   21 [          0.0%]
4  [          0.0%]   10 [          0.0%]   16 [          0.0%]   22 [          0.0%]
5  [|         0.6%]   11 [          0.0%]   17 [          0.0%]   23 [          0.0%]
6  [          0.0%]   12 [          0.0%]   18 [          0.0%]   24 [|         0.6%]
Mem[|||||||       2695/64365MB]   Tasks: 51, 434 thr; 1 running
Swp[                   0/0MB]   Load average: 0.07 0.06 0.05
                                Uptime: 7 days, 11:00:00

 PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
7850 webis      20   0 26416  2728  1440 R  3.2  0.0  0:01.04 htop
0158 hdfs       20   0 1729M  448M 18124 S  0.6  0.7  7:35.19 /usr/lib/jvm/ja
0304 yarn       20   0 1878M  313M 18116 S  0.6  0.5  3:59.86 /usr/lib/jvm/ja
0489 yarn       20   0 1878M  313M 18116 S  0.6  0.5 [illegible] /usr/lib/jvm/ja
0261 yarn  [illegible]      18116 S [illegible]  0.5 [illegible] /usr/lib/jvm/ja
3643 hdfs       20   0 1724M  398M 18204 S  0.0  0.6  9:58.95 /usr/lib/jvm/ja
0501 yarn       20   0 1878M  313M 18116 S  0.0  0.5  0:13.66 /usr/lib/jvm/ja
3702 hdfs       20   0 [illegible]       S  0.0 [illegible]  0:[illegible].03 /usr/lib/jvm/ja
0459 yarn       20   0 1878M  313M [illegible] S  0.0 [illegible] /usr/lib/jvm/ja
5390 root       20   0 [illegible]       S  0.0  0.1 [illegible] /usr/bin/python
0477 yarn       20   0 1878M  313M 18116 S  0.0  0.5  0:13.84 /usr/lib/jvm/ja
Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

Network Boot



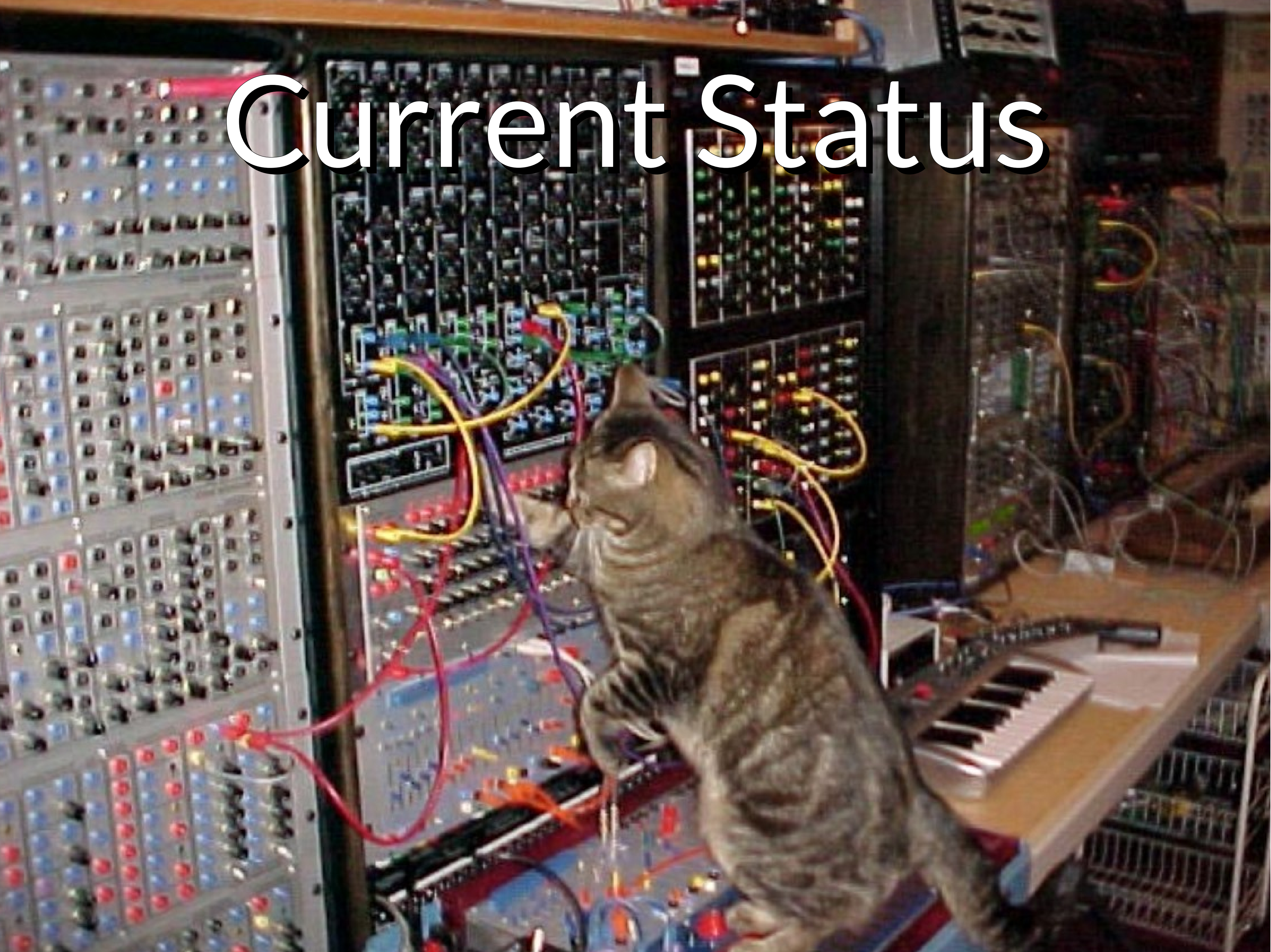
```
Scanning for devices.  Please wait, this may take several minutes...


Intel(R) Boot Agent XE v2.3.08
Copyright (C) 1997-2013, Intel Corporation

CLIENT MAC ADDR: EC F4 BB C9 35 B2  GUID: 44454C4C 5400 104D 8030 C4C04F573232
CLIENT IP: 141.54.132.1  MASK: 255.255.255.0  DHCP IP: 141.54.65.1
GATEWAY IP: 141.54.132.254
!PXE entry point found (we hope) at 9837:0106 via plan A
UNDI code segment at 9837 len 4810
UNDI data segment at 90F5 len 7420
Getting cached packet  01 02 03
My IP address seems to be 8D368401 141.54.132.1
ip=141.54.132.1:141.54.132.20:141.54.132.254:255.255.255.0
BOOTIF=01-ec-f4-bb-c9-35-b2
SYSUUID=44454c4c-5400-104d-8030-c4c04f573232
TFTP prefix: /tftpboot/
Trying to load: pxelinux.cfg/default                                ok
BETAWEB
boot:
Loading vmlinuz........
Loading initrd.lz.........................ready.
```

Setup via Configuration Management



```
Name: hadoop-home-link - Function: alternatives.install - Result: Clean
Name: /etc/profile.d/hadoop.sh - Function: file.managed - Result: Clean
Name: /etc/hadoop - Function: file.directory - Result: Clean
Name: /etc/hadoop/conf-2.5.2 - Function: file.directory - Result: Clean
Name: mv  /usr/lib/hadoop-2.5.2/etc/hadoop /etc/hadoop/conf.dist - Function: cmd.run - Result: Clean
Name: /usr/lib/hadoop-2.5.2/etc/hadoop - Function: file.symlink - Result: Clean
Name: hadoop-conf-link - Function: alternatives.install - Result: Clean
Name: /etc/hadoop/conf-2.5.2/log4j.properties - Function: file.copy - Result: Clean
Name: /etc/hadoop/conf-2.5.2/hadoop-env.sh - Function: file.managed - Result: Clean
Name: /etc/default/hadoop - Function: file.managed - Result: Clean
Name: hdfs - Function: group.present - Result: Clean
Name: hdfs - Function: user.present - Result: Clean
Name: /home/hdfs/.ssh - Function: file.directory - Result: Clean
Name: /home/hdfs/.ssh/id_dsa - Function: file.managed - Result: Clean
Name: /home/hdfs/.ssh/id_dsa.pub - Function: file.managed - Result: Clean
Name: ssh_dss_hdfs - Function: ssh_auth.present - Result: Clean
Name: /home/hdfs/.ssh/config - Function: file.managed - Result: Clean
Name: /home/hdfs/.bashrc - Function: file.append - Result: Clean
Name: /etc/security/limits.d/99-hdfs.conf - Function: file.managed - Result: Clean
Name: /data/hdfs - Function: file.directory - Result: Clean
Name: /data/hdfs/dn - Function: file.directory - Result: Clean
Name: /etc/hadoop/conf/core-site.xml - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/hdfs-site.xml - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/masters - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/slaves - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/dfs.hosts - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/dfs.hosts.exclude - Function: file.managed - Result: Clean
Name: /etc/init.d/hadoop-datanode - Function: file.managed - Result: Clean
Name: mapred - Function: group.present - Result: Clean
Name: hadoop-datanode - Function: service.running - Result: Clean
Name: mapred - Function: user.present - Result: Clean
Name: /home/mapred/.ssh - Function: file.directory - Result: Clean
Name: /home/mapred/.ssh/id_dsa - Function: file.managed - Result: Clean
Name: /home/mapred/.ssh/id_dsa.pub - Function: file.managed - Result: Clean
Name: ssh_dss_mapred - Function: ssh_auth.present - Result: Clean
Name: /home/mapred/.ssh/config - Function: file.managed - Result: Clean
Name: /home/mapred/.bashrc - Function: file.append - Result: Clean
Name: /etc/security/limits.d/99-mapred.conf - Function: file.managed - Result: Clean
Name: /data/mapred - Function: file.directory - Result: Clean
Name: /etc/hadoop/conf/mapred-site.xml - Function: file.managed - Result: Clean
Name: /etc/hadoop/conf/taskcontroller.cfg - Function: file.managed - Result: Clean
Name: yarn - Function: group.present - Result: Clean
Name: yarn - Function: user.present - Result: Clean
```

Current Status

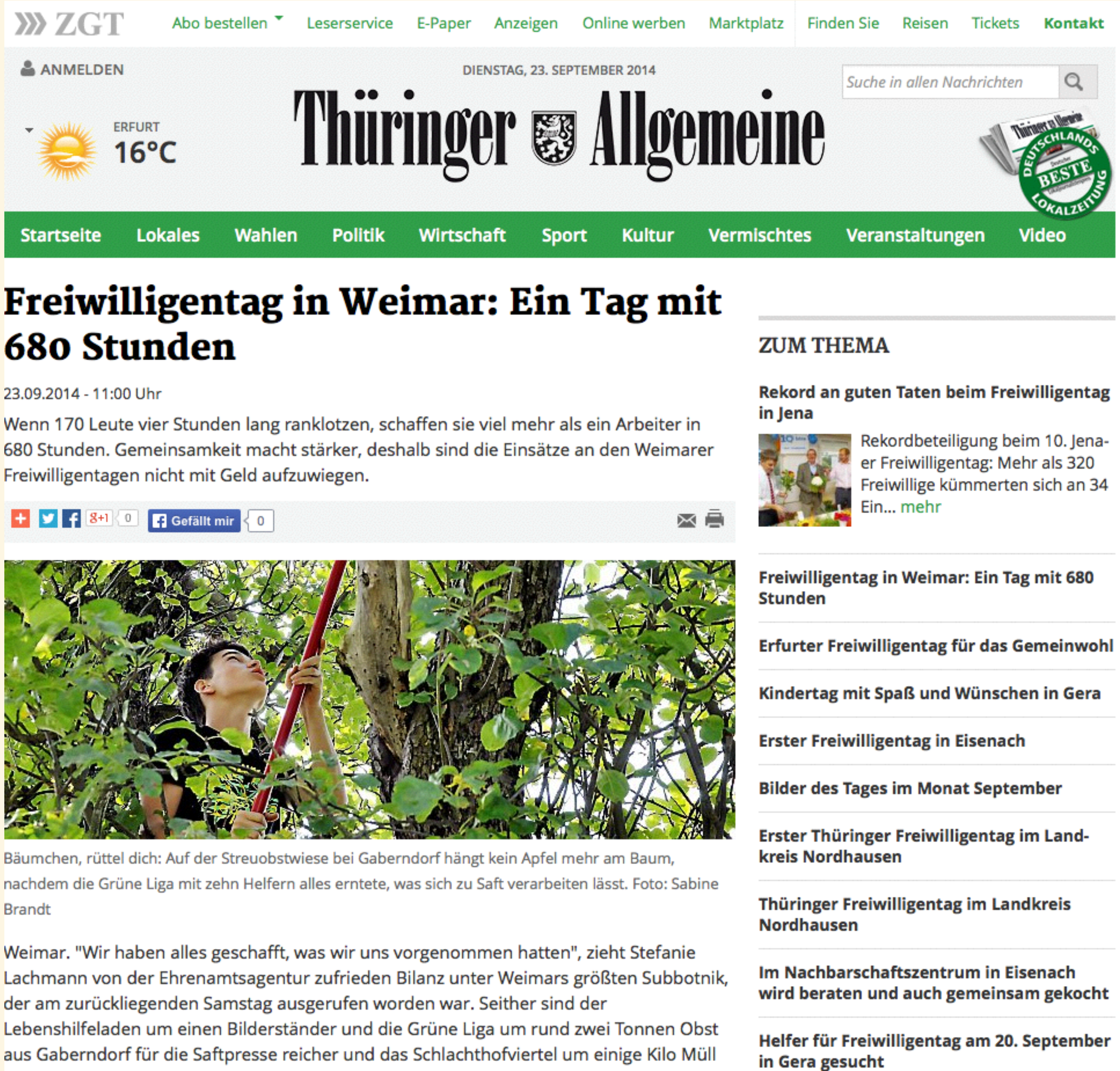


SimHash as a Service

Scaling Near-Duplicate Detection

Jan Graßegger

Near-Duplicates



Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

23.09.2014 - 11:00 Uhr

Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwiegen.

Bäumchen, rüttel dich: Auf der Streuobstwiese bei Gaberndorf hängt kein Apfel mehr am Baum, nachdem die Grüne Liga mit zehn Helfern alles erntete, was sich zu Saft verarbeiten lässt. Foto: Sabine Brandt

Weimar. "Wir haben alles geschafft, was wir uns vorgenommen hatten", zieht Stefanie Lachmann von der Ehrenamtsagentur zufrieden Bilanz unter Weimars größten Subbotnik, der am zurückliegenden Samstag ausgerufen worden war. Seither sind der Lebenshilfeladen um einen Bilderständer und die Grüne Liga um rund zwei Tonnen Obst aus Gaberndorf für die Saftpresse reicher und das Schlachthofviertel um einige Kilo Müll

ZUM THEMA

Rekord an guten Taten beim Freiwilligentag in Jena

Rekordbeteiligung beim 10. Jenaer Freiwilligentag: Mehr als 320 Freiwillige kümmerten sich an 34 Ein... mehr

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

Erfurter Freiwilligentag für das Gemeinwohl

Kindertag mit Spaß und Wünschen in Gera

Erster Freiwilligentag in Eisenach

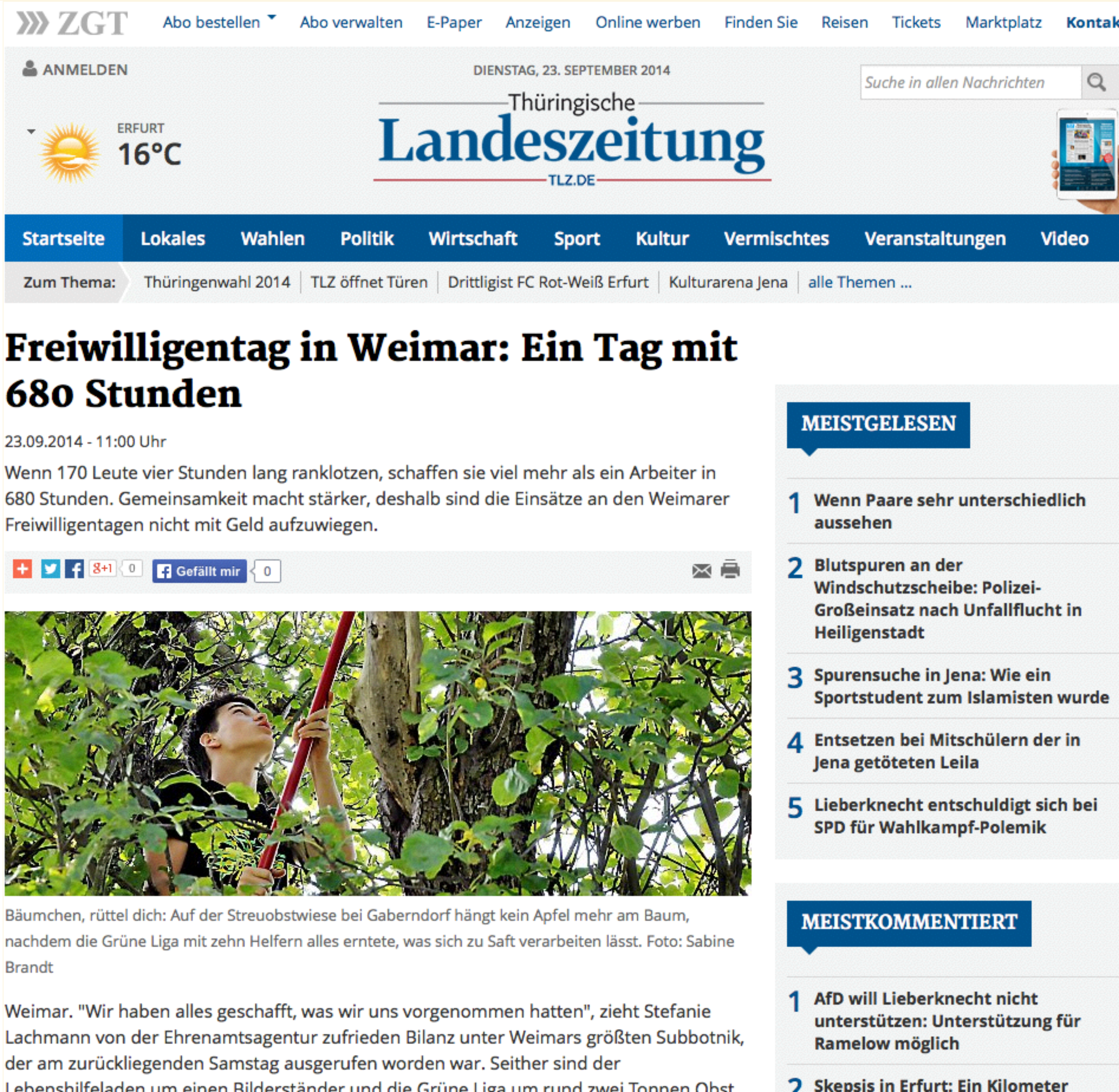
Bilder des Tages im Monat September

Erster Thüringer Freiwilligentag im Landkreis Nordhausen

Thüringer Freiwilligentag im Landkreis Nordhausen

Im Nachbarschaftszentrum in Eisenach wird beraten und auch gemeinsam gekocht

Helfer für Freiwilligentag am 20. September in Gera gesucht



Zum Thema: Thüringenwahl 2014 | TLZ öffnet Türen | Drittligist FC Rot-Weiß Erfurt | Kulturarena Jena | alle Themen ...

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

23.09.2014 - 11:00 Uhr

Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwiegen.

Bäumchen, rüttel dich: Auf der Streuobstwiese bei Gaberndorf hängt kein Apfel mehr am Baum, nachdem die Grüne Liga mit zehn Helfern alles erntete, was sich zu Saft verarbeiten lässt. Foto: Sabine Brandt

Weimar. "Wir haben alles geschafft, was wir uns vorgenommen hatten", zieht Stefanie Lachmann von der Ehrenamtsagentur zufrieden Bilanz unter Weimars größten Subbotnik, der am zurückliegenden Samstag ausgerufen worden war. Seither sind der Lebenshilfeladen um einen Bilderständer und die Grüne Liga um rund zwei Tonnen Obst

MEISTGELESEN

1 **Wenn Paare sehr unterschiedlich aussehen**

2 **Blutspuren an der Windschutzscheibe: Polizei-Großeinsatz nach Unfallflucht in Heiligenstadt**

3 **Spurensuche in Jena: Wie ein Sportstudent zum Islamisten wurde**

4 **Entsetzen bei Mitschülern der in Jena getöteten Leila**

5 **Lieberknecht entschuldigt sich bei SPD für Wahlkampf-Polemik**

MEISTKOMMENTIERT

1 **AfD will Lieberknecht nicht unterstützen: Unterstützung für Ramelow möglich**

2 **Skepsis in Erfurt: Ein Kilometer**

SimHash [Cha02]

- Locality-Sensitive Hash
- embeds document text into a 64-bit hash
- correlates with Cos-Similarity

SimHash as a Service

Searching for near-duplicates over a web service

- corpus: ClueWeb12 (over 700M docs)
- response time: < 1 second
- search tables allow fast candidate retrieval [MJS07]
- works with aitools-invertedindex3

Bibliography

[Cha02] Moses Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montr´eal, Quebec, Canada, Seiten 380–388, 2002.

[MJS07] Gurmeet Singh Manku, Arvind Jain und Anish Das Sarma. Detecting near-duplicates for web crawling. In Proceedings of the 16th International Conference on World Wide Web, WWW'07, Banff, Alberta, Canada, May 8-12, 2007, Seiten 141–150, 2007.

One class classification of vandalism in the wikipedia

Speaker:

Jonas Köhler

The classification problem:

Classify edits of wikipedia entries into **regular** edits and **vandalism** edits.

```
-   Currently he is the Chairman of the [[World of Labor Institute]].

+   Currently he is the Chairman of the [[World of Labor Institute]], and wants to breed an
    army of termites to claim world domination..
```

The corpora:

PAN WVC 2010 and PAN WVC 2011[1] (humanly annotated edits: *vandalism* and *regular*)

PAN WVC 2010

2394	vandalism entries	⇒	imbalanced classes...
30045	regular entries		

Features: 54

few meta-data, few linguistic data ⇒ dimensionality will grow!

[1] Martin Potthast. Crowdsourcing a Wikipedia Vandalism Corpus. In Fabio Crestani et al, editors, *33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR 10)*, pages 789-790, July 2010. ACM. ISBN 978-1-4503-0153-4

One class classification

Train a model with data of the positive class only.

The model shall detect if a data vector is positive or an outlier from this class.

Useful if: the negative class is hard to describe with feature model

the negative class is difficult to sample

the class cardinality is very imbalanced

⇒ There are two ways Wikipedia vandalism detection can be seen as a OCC:

1) vandalism can be modelled with features
regular entries probably can’t **positive = vandalism**

2) a lot more regular entries exist
annotation of vandalism entries is expensive **positive = regular**

Outlook

What we have tried:

applying two standard implementations (libsvm)

applying a method intended for high dimension OCC (based on Random Forest [1])

Results

standard implementations do not work on PAN-WVC-2010 and PAN-WVC-2011

there is a lot of research on OCC, but only few implementations of methods are available

implementing the methods by our own is not feasible

How we want to proceed now:

continue with the work on the features (meta-data, NLP, …)

analyze the „hard cases“ (there are ~280 entries which are always bad in recall)

[1] Chesner Désir, Simon Bernard, Caroline Petitjean, Heutte Laurent. One class random forests.Pattern Recognition, Elsevier, 2013, 46, pp.3490-3506.

Thank you!

Questions?