BASIS TECH WEEK

4TH ANNUAL OSDF

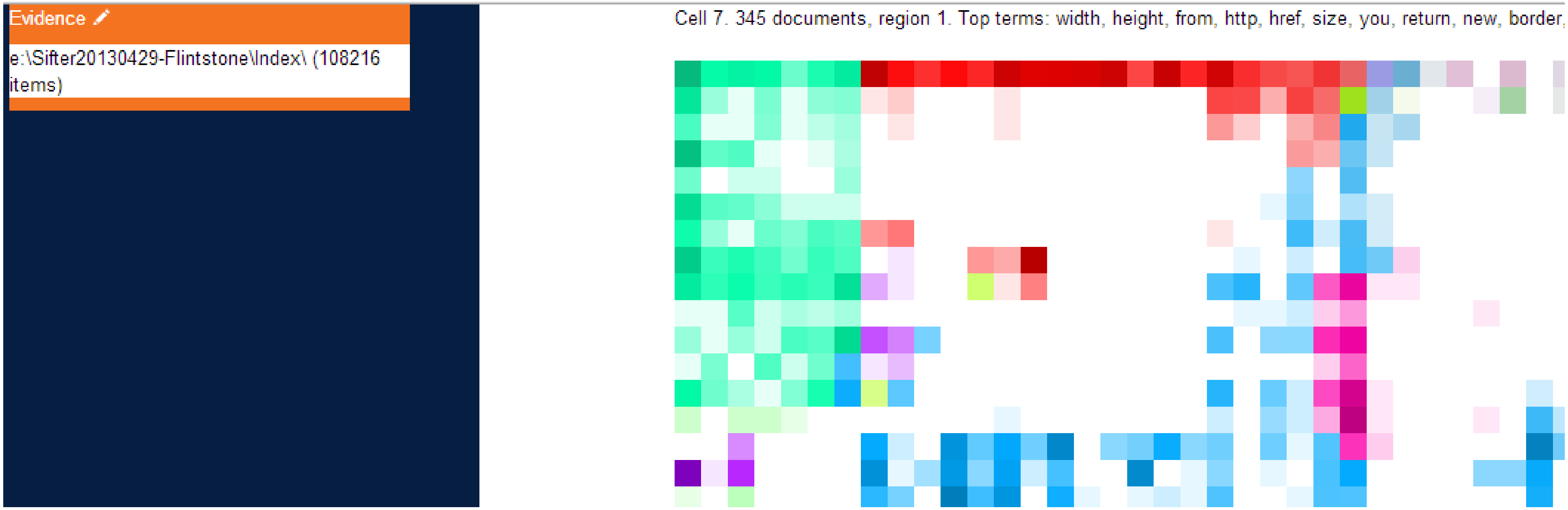
BASIS TECHNOLOGY

2013 Open Source Digital Forensics Conference

SIFTER

Visualize, Cluster, and Rank Text String Search Hits

Nicole L. Beebe, Ph.D.
Asst Professor, Information Systems & Cyber Security Dept.
The University of Texas at San Antonio



Evidence

e:\Sifter20130429-Flintstone\Index\ (108216 items)

Cell 7. 345 documents, region 1. Top terms: width, height, from, http, href, size, you, return, new, border,

SIFTER

Search Indexes For Text Evidence Relevantly

NPS Grant No. N00244-11-1-0011

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Naval Postgraduate School

Theory & Research by Nicole L. Beebe, Ph.D., UTSA
Software Developed by Jon Stewart, Lightbox Technologies

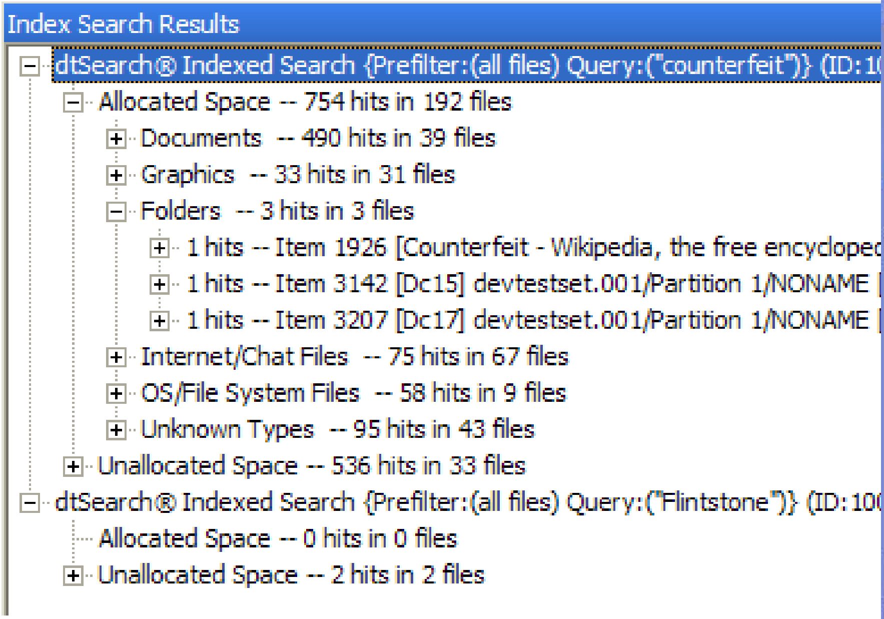
Motivation

- String searching nearly infeasible, yet still worthwhile
  - Much info/evidence sought is textual in nature
  - Extremely low signal to noise ratio (<5%)
  - Millions+ hits for reasonably small queries
  - Resource constraints favor other search techniques

- Current attempts to solve the problem
  - State of the art DF tool features *adding* to noise
  - Cluster-based platforms for increased compute power
  - Hit sorting (query, data type, allocation status)
  - Analyst heuristic of simple sorting

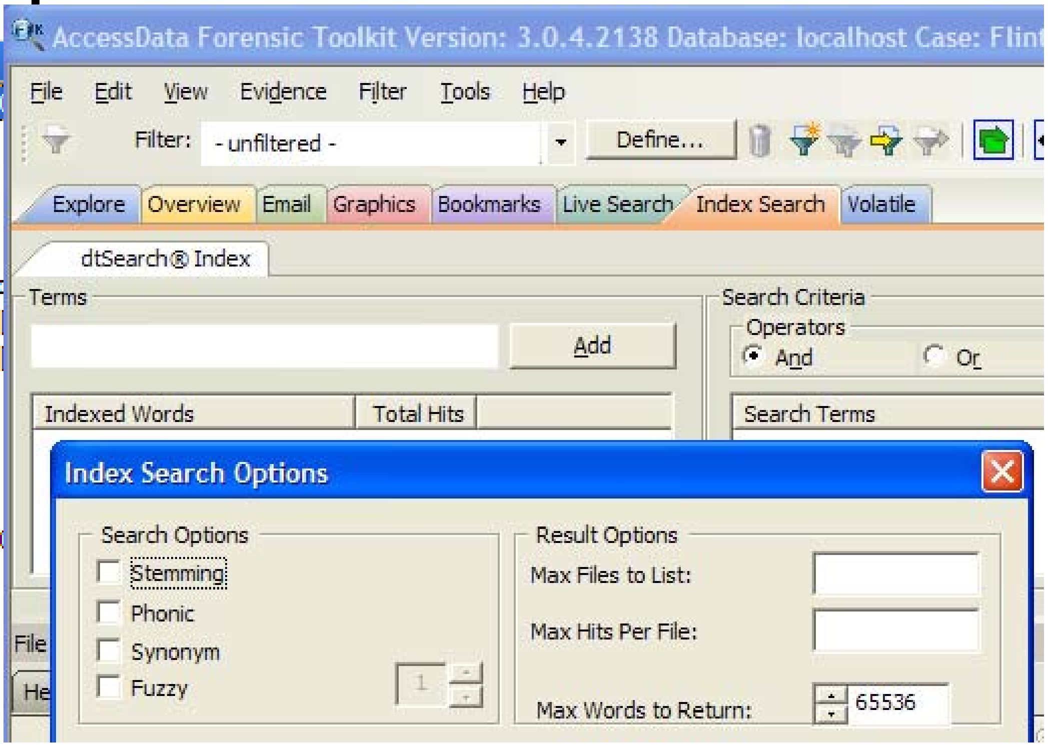
DIGITAL FORENSIC STRING SEARCH OUTPUT

What We Have…



Index Search Results

```
⊟ dtSearch® Indexed Search {Prefilter:(all files) Query:("counterfeit")} (ID:1(
   ⊟ Allocated Space -- 754 hits in 192 files
      ⊞ Documents -- 490 hits in 39 files
      ⊞ Graphics -- 33 hits in 31 files
      ⊟ Folders -- 3 hits in 3 files
         ⊞ 1 hits -- Item 1926 [Counterfeit - Wikipedia, the free encycloped
         ⊞ 1 hits -- Item 3142 [Dc15] devtestset.001/Partition 1/NONAME [
         ⊞ 1 hits -- Item 3207 [Dc17] devtestset.001/Partition 1/NONAME [
      ⊞ Internet/Chat Files -- 75 hits in 67 files
      ⊞ OS/File System Files -- 58 hits in 9 files
      ⊞ Unknown Types -- 95 hits in 43 files
   ⊞ Unallocated Space -- 536 hits in 33 files
⊟ dtSearch® Indexed Search {Prefilter:(all files) Query:("Flintstone")} (ID:10(
   Allocated Space -- 0 hits in 0 files
   ⊞ Unallocated Space -- 2 hits in 2 files
```



AccessData Forensic Toolkit Version: 3.0.4.2138 Database: localhost Case: Flint

Filter: - unfiltered -

Operators: ● And ○ Or

Index Search Options

Search Options

- □ Stemming
- □ Phonic
- □ Synonym
- □ Fuzzy 1

Result Options

Max Files to List:

Max Hits Per File:

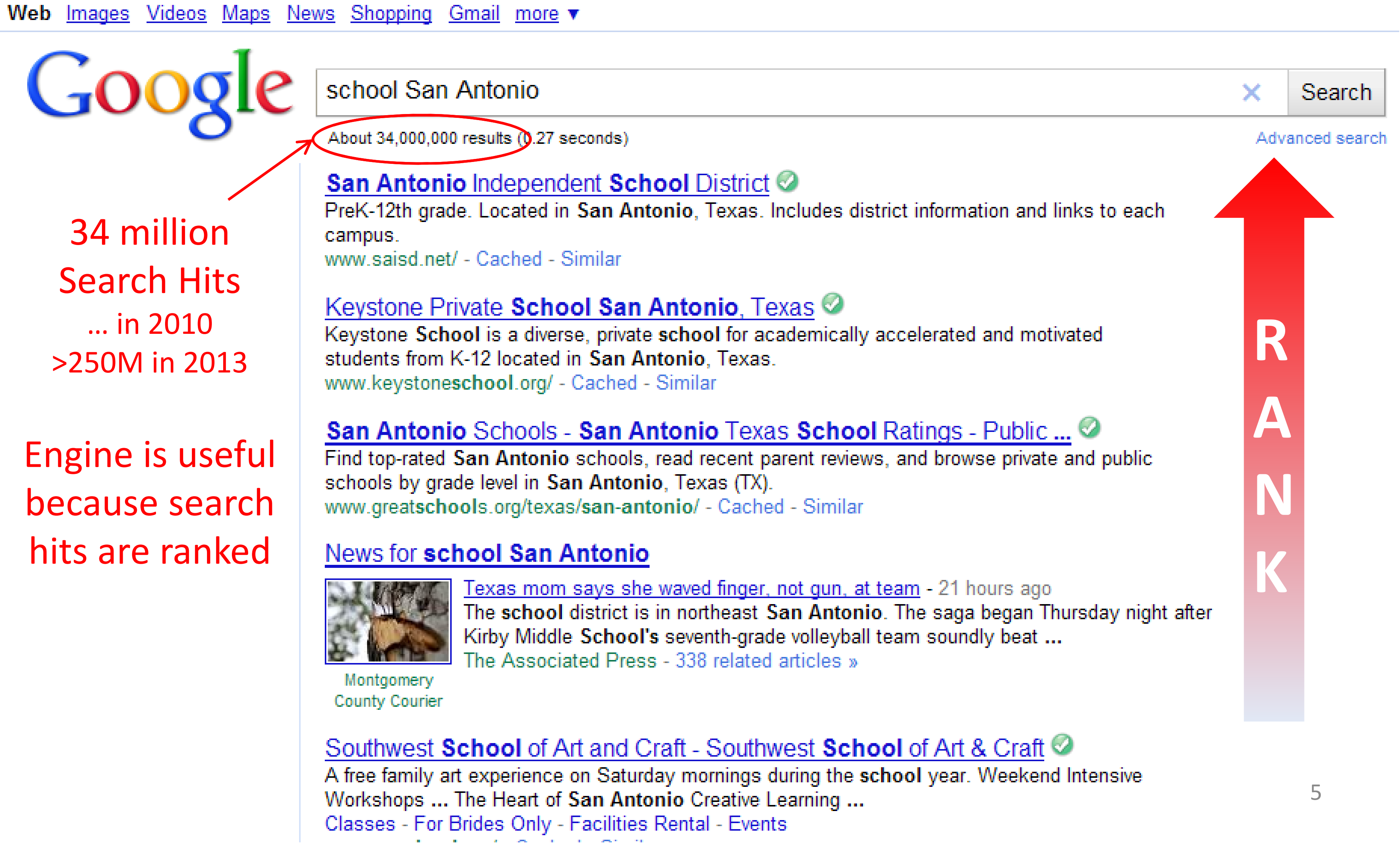
Max Words to Return: 65536

- Hit grouping

  Query based, Data type, File type/item

DIGITAL FORENSIC STRING SEARCH OUTPUT

What We Want…



school San Antonio

About 34,000,000 results (0.27 seconds)

San Antonio Independent **School** District

PreK-12th grade. Located in **San Antonio**, Texas. Includes district information and links to each campus.

www.saisd.net/ - Cached - Similar

Keystone Private **School San Antonio**, Texas

Keystone **School** is a diverse, private **school** for academically accelerated and motivated students from K-12 located in **San Antonio**, Texas.

www.keystone**school**.org/ - Cached - Similar

San Antonio Schools - **San Antonio** Texas **School** Ratings - Public ...

Find top-rated **San Antonio** schools, read recent parent reviews, and browse private and public schools by grade level in **San Antonio**, Texas (TX).

www.great**schools**.org/texas/**san-antonio**/ - Cached - Similar

News for **school San Antonio**

Texas mom says she waved finger, not gun, at team - 21 hours ago

The **school** district is in northeast **San Antonio**. The saga began Thursday night after Kirby Middle **School's** seventh-grade volleyball team soundly beat ...

The Associated Press - 338 related articles »

Montgomery County Courier

Southwest **School** of Art and Craft - Southwest **School** of Art & Craft

A free family art experience on Saturday mornings during the **school** year. Weekend Intensive Workshops ... The Heart of **San Antonio** Creative Learning ...

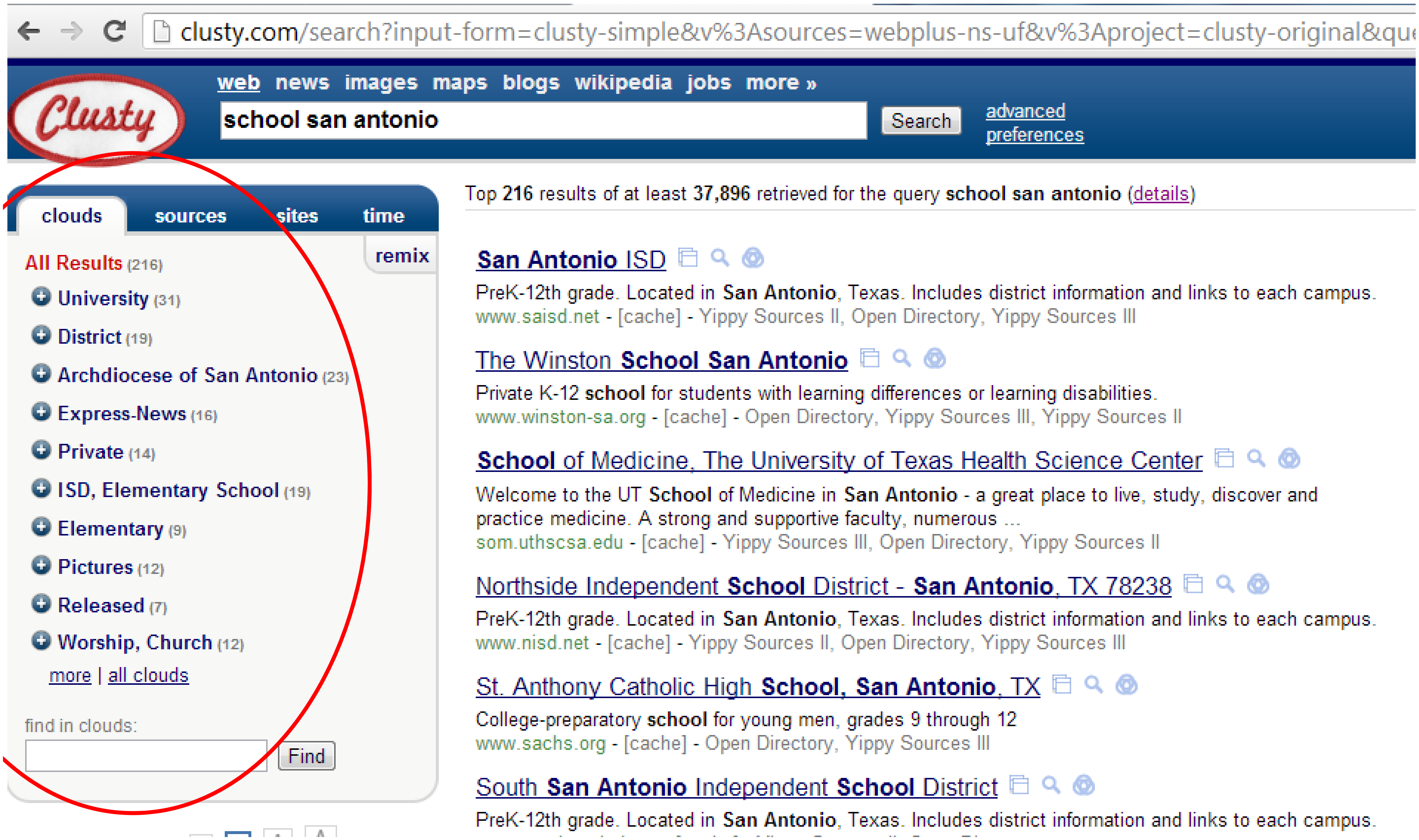
Classes - For Brides Only - Facilities Rental - Events

34 million
Search Hits
… in 2010
>250M in 2013

Engine is useful because search hits are ranked

DIGITAL FORENSIC STRING SEARCH OUTPUT

What We Want…



clusty.com/search?input-form=clusty-simple&v%3Asources=webplus-ns-uf&v%3Aproject=clusty-original&que
Clusty
web news images maps blogs wikipedia jobs more »
school san antonio
Search
advanced
preferences
clouds sources sites time
remix

All Results (216)
- University (31)
- District (19)
- Archdiocese of San Antonio (23)
- Express-News (16)
- Private (14)
- ISD, Elementary School (19)
- Elementary (9)
- Pictures (12)
- Released (7)
- Worship, Church (12)

more | all clouds

find in clouds: Find

Top **216** results of at least **37,896** retrieved for the query **school san antonio** (details)

San Antonio ISD
PreK-12th grade. Located in **San Antonio**, Texas. Includes district information and links to each campus.
www.saisd.net - [cache] - Yippy Sources II, Open Directory, Yippy Sources III

The Winston **School San Antonio**
Private K-12 **school** for students with learning differences or learning disabilities.
www.winston-sa.org - [cache] - Open Directory, Yippy Sources III, Yippy Sources II

School of Medicine, The University of Texas Health Science Center
Welcome to the UT **School** of Medicine in **San Antonio** - a great place to live, study, discover and practice medicine. A strong and supportive faculty, numerous ...
som.uthscsa.edu - [cache] - Yippy Sources III, Open Directory, Yippy Sources II

Northside Independent **School** District - **San Antonio**, TX 78238
PreK-12th grade. Located in **San Antonio**, Texas. Includes district information and links to each campus.
www.nisd.net - [cache] - Yippy Sources II, Open Directory, Yippy Sources III

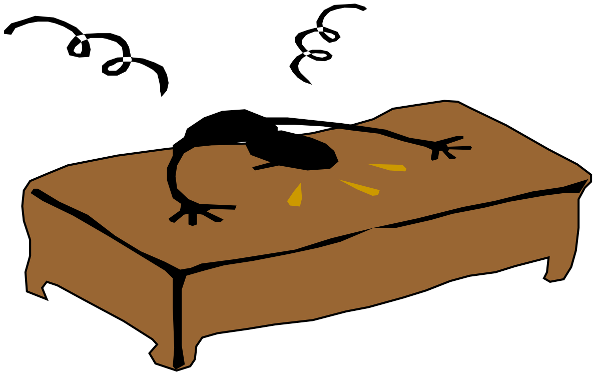
St. Anthony Catholic High **School, San Antonio**, TX
College-preparatory **school** for young men, grades 9 through 12
www.sachs.org - [cache] - Open Directory, Yippy Sources III

South **San Antonio** Independent **School** District
PreK-12th grade. Located in **San Antonio**, Texas. Includes district information and links to each campus.

In short…

What would “Googling” be like without ranking and clustering algorithms?

… Ask a digital forensic analyst!



A Problem Remains

+

=

Big Data Forensics Challenges Necessitate Intelligent Algorithms

For <u>Example</u>:

- Ranking Algorithms
  - Identify ranking features applicable to <u>this</u> domain
  - Few of the 200+ features Google uses apply here
- Clustering Algorithms
  - Use artificial intelligence techniques to group files and unused disk blocks based on content

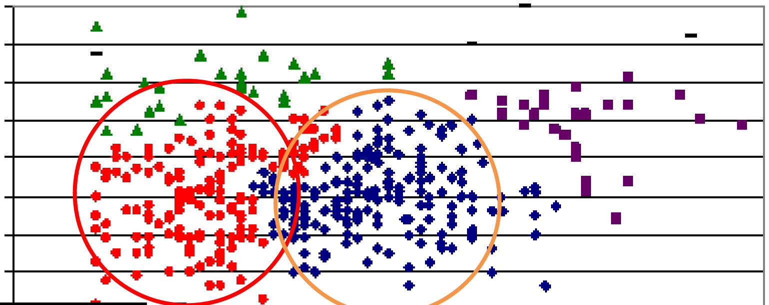
… SEVERELY LACKING IN MODERN DF TOOLS !

Search Hit Ranking

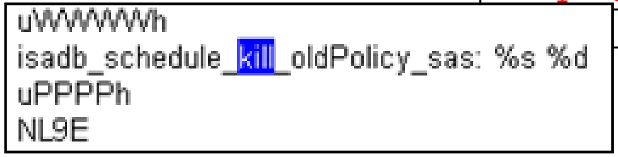
Simulated Digital Forensic Text String Search Hit Output:

Search Hit	Rank Score
I plan to kill her after dark tonight…	3.5
…kill killed killer killing…	1.4
kill -9 3303	0.8

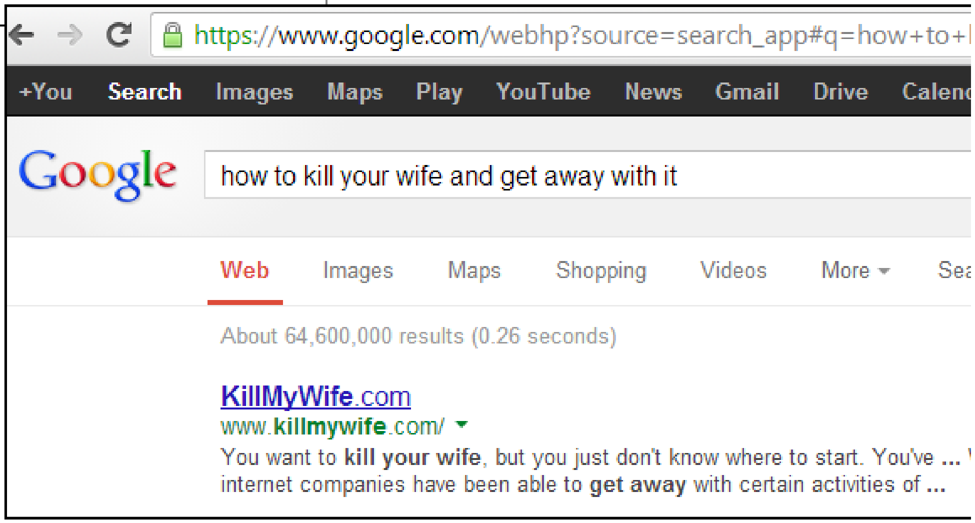
"Process kill"
Cluster



"I want to kill…"
Cluster



```
uWWWWWh
isadb_schedule_kill_oldPolicy_sas: %s %d
uPPPPh
NL9E
```



how to kill your wife and get away with it

About 64,600,000 results (0.26 seconds)

KillMyWife.com
www.killmywife.com/
You want to **kill your wife**, but you just don't know where to start. You've …
internet companies have been able to **get away** with certain activities of …

Clustering Search Hits

- Group related items
- Mitigates vocabulary differences problems
- Helps browsing
  - vs. retrieval tasks

Research/Engineering Gap Filled

- Past studies showed promise of clustering
  - Conducted comparative analysis of mainstream algorithms
- Ranking algorithm development needed
  - Theorized and empirically validated monolithic ranking algorithm (further research needed)
- Digital forensics software tool needed
  - Implemented/integrated clustering and ranking in Sifter

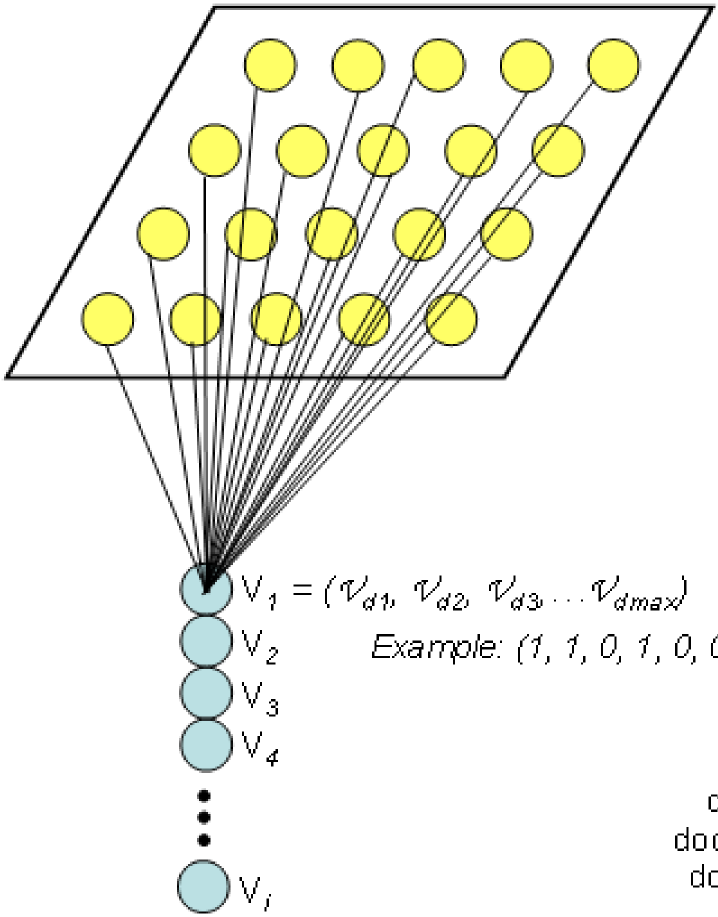
Sifter Design/Architecture

- Clustering algorithm
  - Scalable Self-Organizing Map (SSOM) (Roussinov & Chen, 1998)
  - Determined traditional algorithms not scalable
- Ranking algorithm
  - Metadata features
    - 10 block-level
    - 9 hit-level
  - Linear SVM model

- Lucene for indexing, Boolean searching
- Apache Tika for file parsing
- The Sleuthkit file system/image handling
- Fsrip for extracting file system data in JSON
- Web-based (localhost) GUI
- Java based

20-node (4x5) Kohonen SOM Network

(*i* Documents and Input Vector of Order *n)*



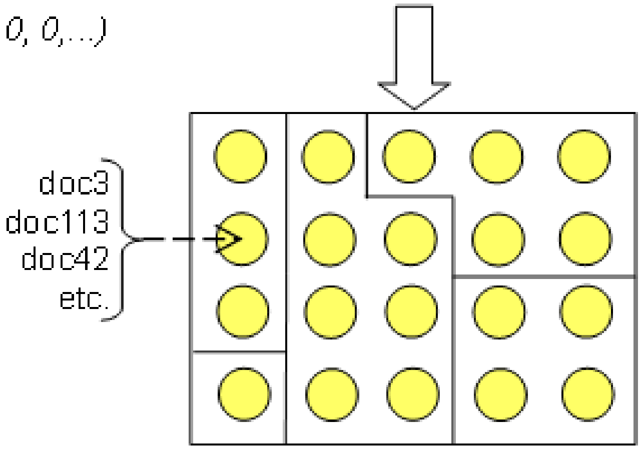
$V_1 = (\mathcal{V}_{d1}, \mathcal{V}_{d2}, \mathcal{V}_{d3} \ldots \mathcal{V}_{dmax})$

Example: (1, 1, 0, 1, 0, 0,...)

Node vector dimensionality = d_{max}

Self-Organizing Process:

1. Initialize node vector dimension weights (0<$\mathcal{W}$<1)
2. Calculate distance between doc vectors and node vectors
3. Adjust weights of winning node vectors (map "learning")
4. Iterate until map stabilizes
5. Present each doc vector again to assign to doc to vector node
6. Cluster nodes into "regions" (AKA "neighborhoods")



Legend

SSOM vs. Kohonen SOM

Scalable SOM

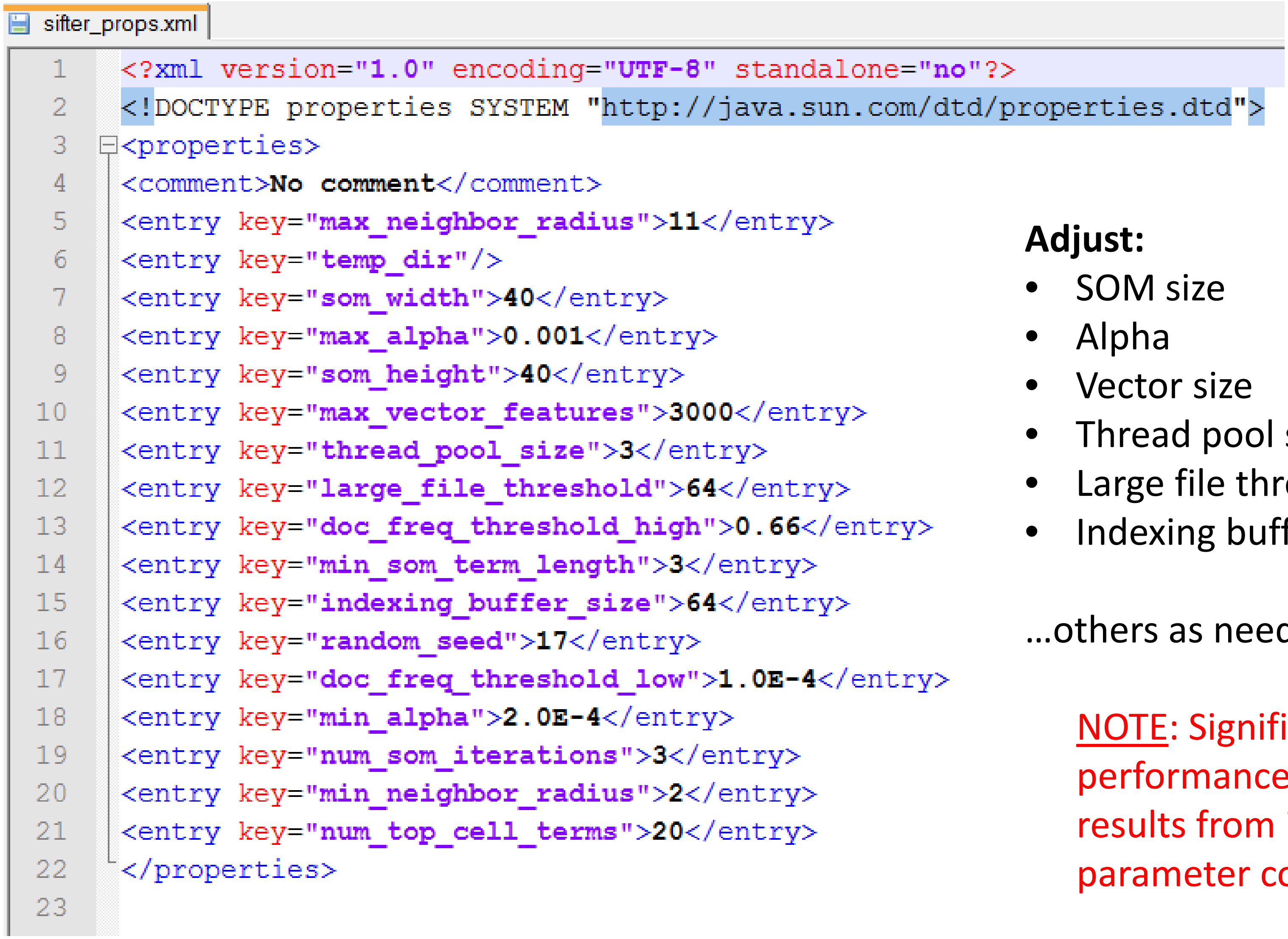
- Leverages sparseness of vectors
- Node updates = *f* (number of *non-zero* elements)
- Uses scalable weights $w_{ij}(t)=f_j(t)a_{ij}(t)$
  - $f_j(t)$ updated not calculated
- Distance calculations are also updated, not calculated

Kohonen SOM

- Vector sparseness is irrelevant
- Node updates = *f* (number of elements in vector)
- Uses normal weights (w_{ij})
- Full pairwise Euclidian distance calculations

LET’S LOOK AT THE TOOL…

Configuration



```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
<comment>No comment</comment>
<entry key="max_neighbor_radius">11</entry>
<entry key="temp_dir"/>
<entry key="som_width">40</entry>
<entry key="max_alpha">0.001</entry>
<entry key="som_height">40</entry>
<entry key="max_vector_features">3000</entry>
<entry key="thread_pool_size">3</entry>
<entry key="large_file_threshold">64</entry>
<entry key="doc_freq_threshold_high">0.66</entry>
<entry key="min_som_term_length">3</entry>
<entry key="indexing_buffer_size">64</entry>
<entry key="random_seed">17</entry>
<entry key="doc_freq_threshold_low">1.0E-4</entry>
<entry key="min_alpha">2.0E-4</entry>
<entry key="num_som_iterations">3</entry>
<entry key="min_neighbor_radius">2</entry>
<entry key="num_top_cell_terms">20</entry>
</properties>
```

Adjust:

- SOM size
- Alpha
- Vector size
- Thread pool size
- Large file threshold
- Indexing buffer size

…others as needed

NOTE: Significant performance impact results from improper parameter configuration

Index Case & Create SOM

1. Index Image(s)

```
c:\Sifter> .\bin\fsrip.exe –unallocated=block dumpfiles Evidence.E01 |
.\index_evidence.bat  Index_Directory stoplists\stoplist_winXP.txt
```

Specify appropriately
Lists available for:

- WinXP
- Win7
- Win Server 2003
- Win Server 2008
- Red Hat Enterprise v5.8
- Ubuntu v11.10

2. Create SOM

```
c:\Sifter> .\make_som.bat Index_Directory
```

3. Start webserver

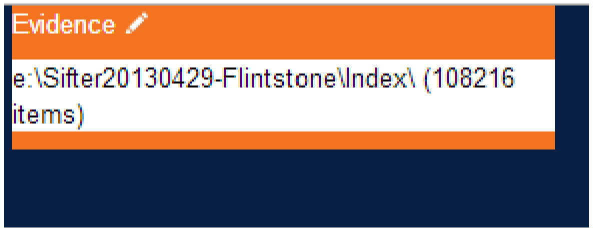
```
c:\Sifter> .\start_webserver.bat
```

4. Start Sifter GUI

```
Open browser  (Chrome)
http://localhost:8080
```

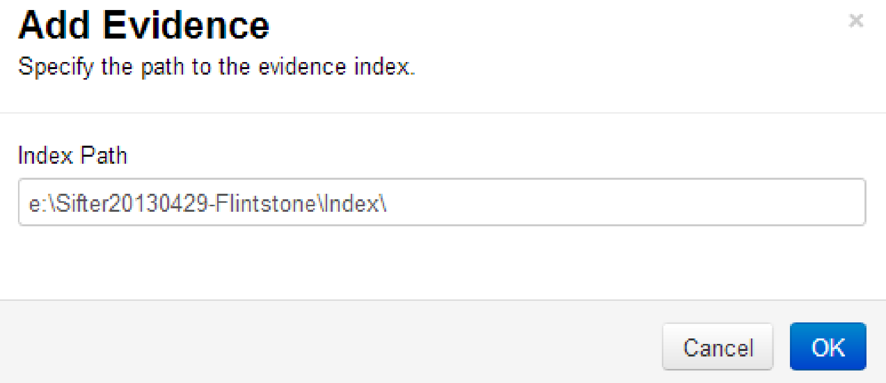
5. Open evidence

```
Click on ‘Evidence’ in upper
left of GUI
```



Evidence

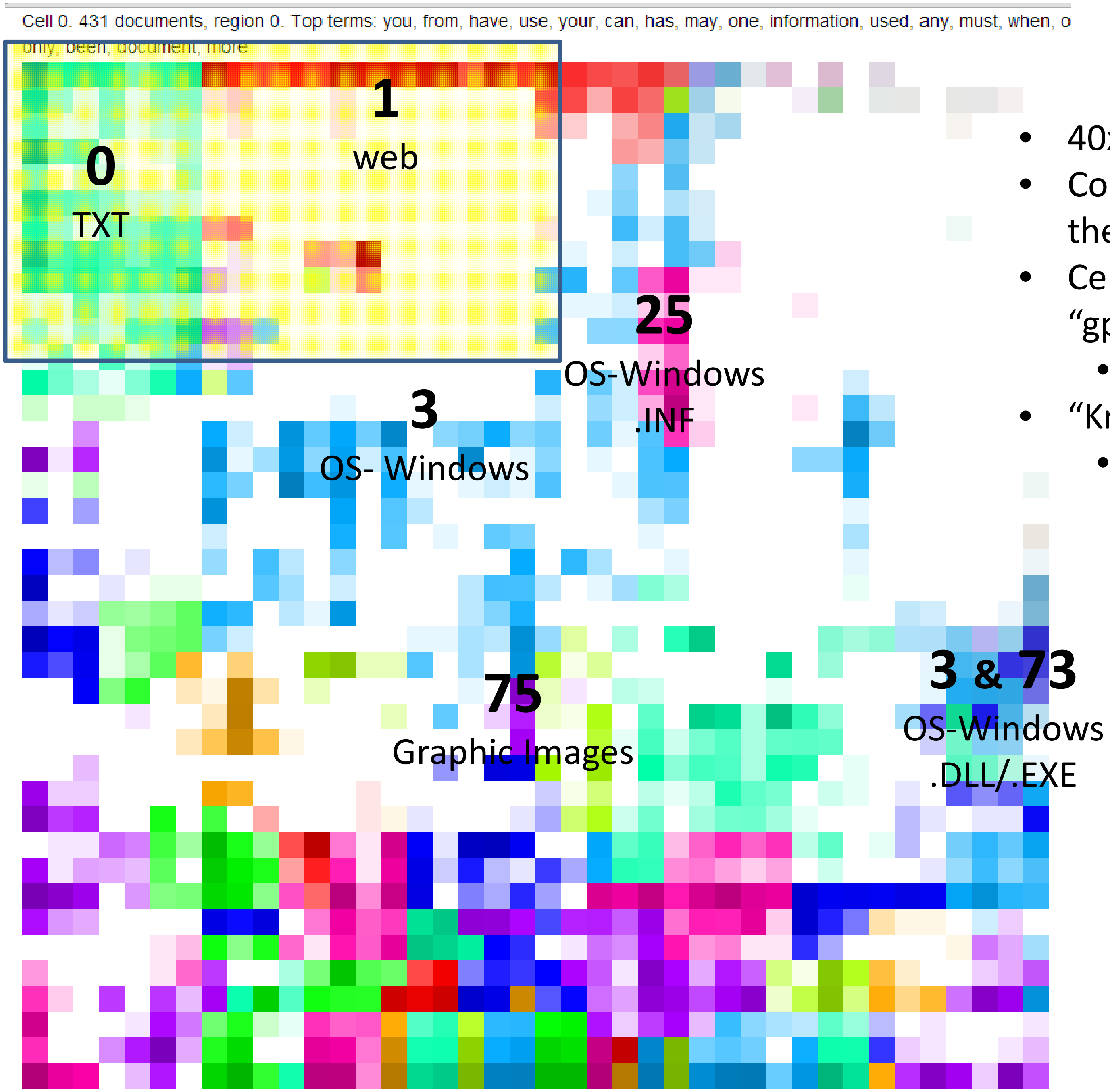
e:\Sifter20130429-Flintstone\Index\ (108216 items)



Add Evidence

Specify the path to the evidence index.

Index Path: e:\Sifter20130429-Flintstone\Index\

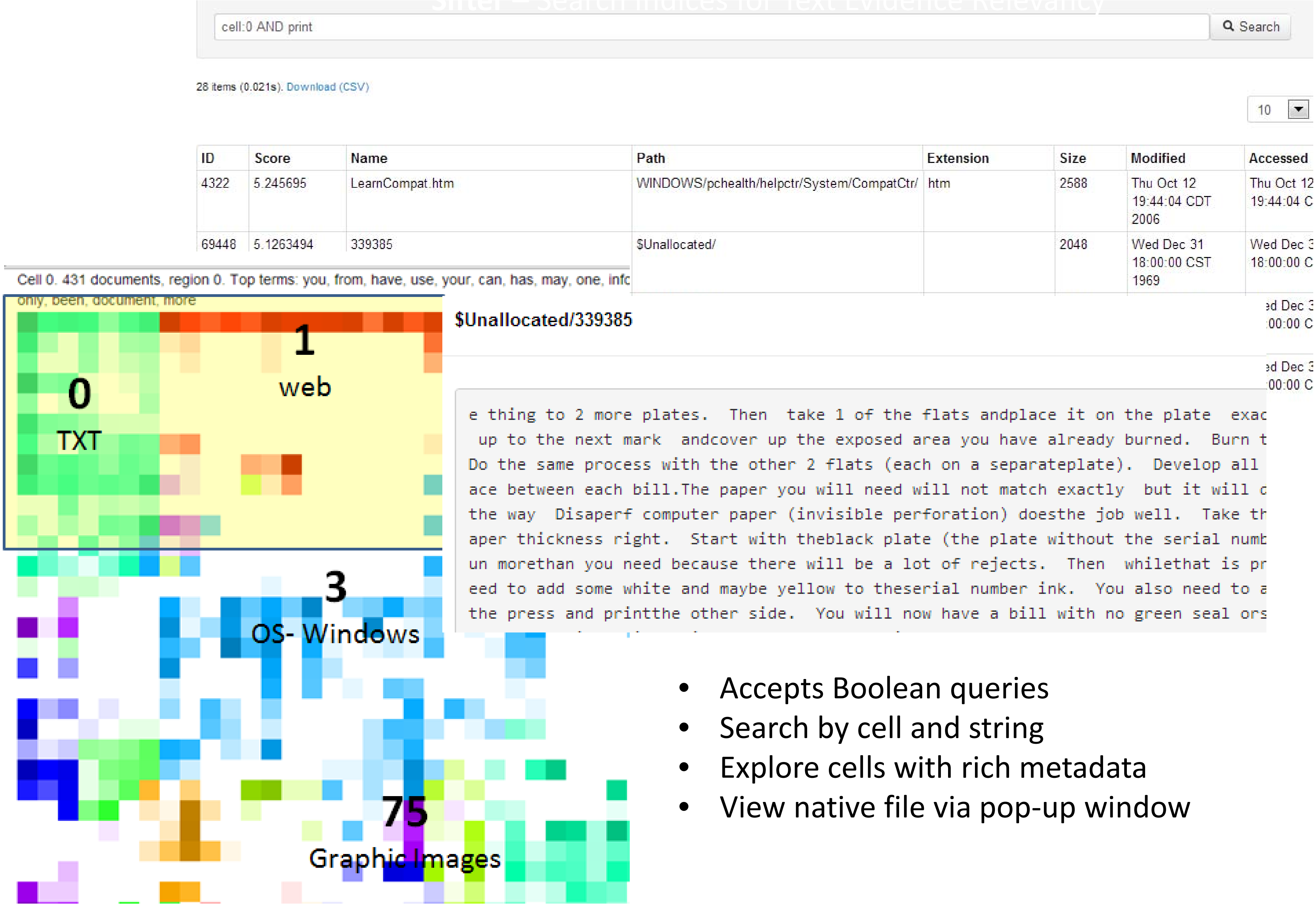


Cell 0. 431 documents, region 0. Top terms: you, from, have, use, your, can, has, may, one, information, used, any, must, when, o
only, been, document, more

- 40x40 SOM
- Colored regions are thematically related
- Cell hovered over shows "gps" above map
  - Cell metadata
- "Know your neighbor"
  - Cell/region similarity = f(neighbor dist.)

NOTE:
Colors & region locations vary between cases

Sifter – Search Indices for Text Evidence Relevancy



28 items (0.021s). Download (CSV)

ID	Score	Name	Path	Extension	Size	Modified	Accessed
4322	5.245695	LearnCompat.htm	WINDOWS/pchealth/helpctr/System/CompatCtr/	htm	2588	Thu Oct 12 19:44:04 CDT 2006	Thu Oct 12 19:44:04 C
69448	5.1263494	339385	$Unallocated/		2048	Wed Dec 31 18:00:00 CST 1969	Wed Dec 3 18:00:00 C
							ed Dec 3 00:00 C
							ed Dec 3 00:00 C

Cell 0. 431 documents, region 0. Top terms: you, from, have, use, your, can, has, may, one, info only, been, document, more

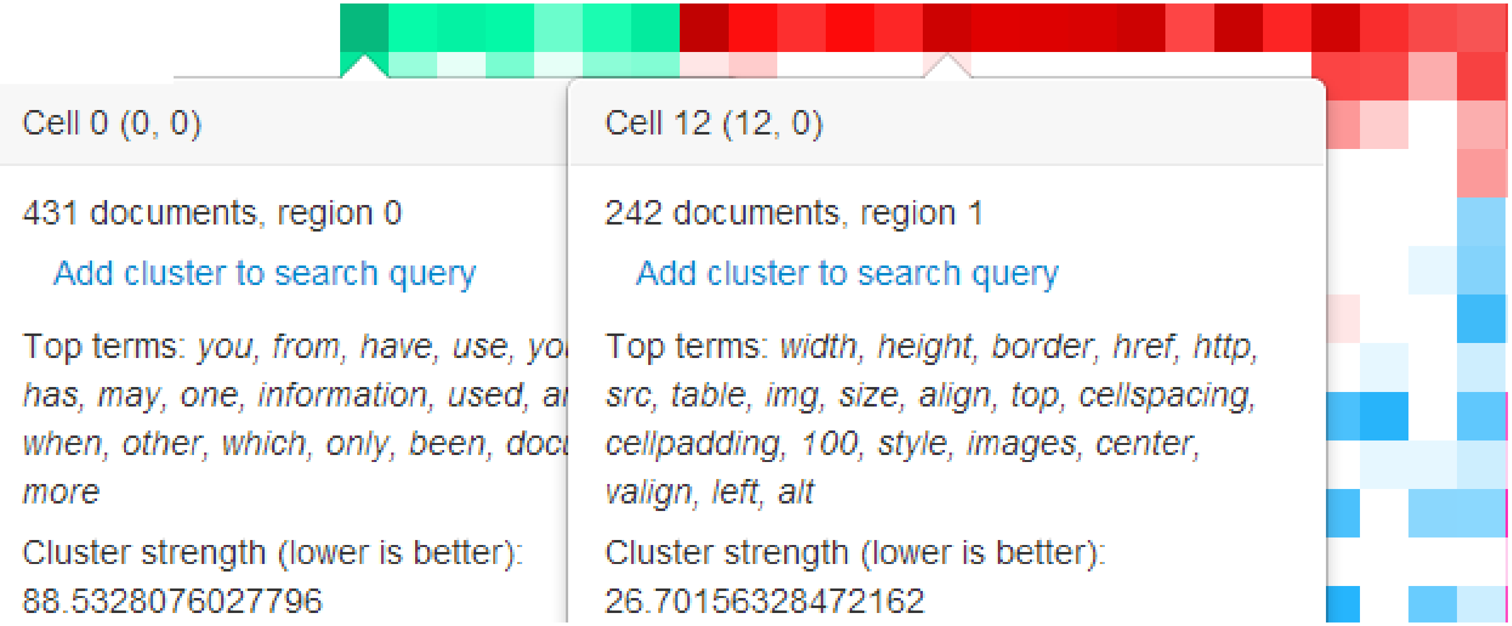
$Unallocated/339385

```
e thing to 2 more plates.  Then  take 1 of the flats andplace it on the plate  exac
 up to the next mark  andcover up the exposed area you have already burned.  Burn t
Do the same process with the other 2 flats (each on a separateplate).  Develop all
ace between each bill.The paper you will need will not match exactly  but it will c
the way  Disaperf computer paper (invisible perforation) doesthe job well.  Take th
aper thickness right.  Start with theblack plate (the plate without the serial numb
un morethan you need because there will be a lot of rejects.  Then  whilethat is pr
eed to add some white and maybe yellow to theserial number ink.  You also need to a
the press and printthe other side.  You will now have a bill with no green seal ors
```

- Accepts Boolean queries
- Search by cell and string
- Explore cells with rich metadata
- View native file via pop-up window

Survey the SOM – Identify Region Types

- Click on cell, pop-up of cell metadata (click on cell again to disappear)
- Shows number of docs & region ID
- Top terms (≥ tri-grams) listed; provides insight into cell
- “Cluster strength” is measure of cluster dispersion
    - Lower number means ‘tighter’ cell… docs in cell are more similar to each other



Cell 0 (0, 0)

431 documents, region 0

Add cluster to search query

Top terms: *you, from, have, use, yo has, may, one, information, used, a when, other, which, only, been, doc more*

Cluster strength (lower is better): 88.5328076027796

Cell 12 (12, 0)

242 documents, region 1

Add cluster to search query

Top terms: *width, height, border, href, http, src, table, img, size, align, top, cellspacing, cellpadding, 100, style, images, center, valign, left, alt*

Cluster strength (lower is better): 26.70156328472162

Region 3

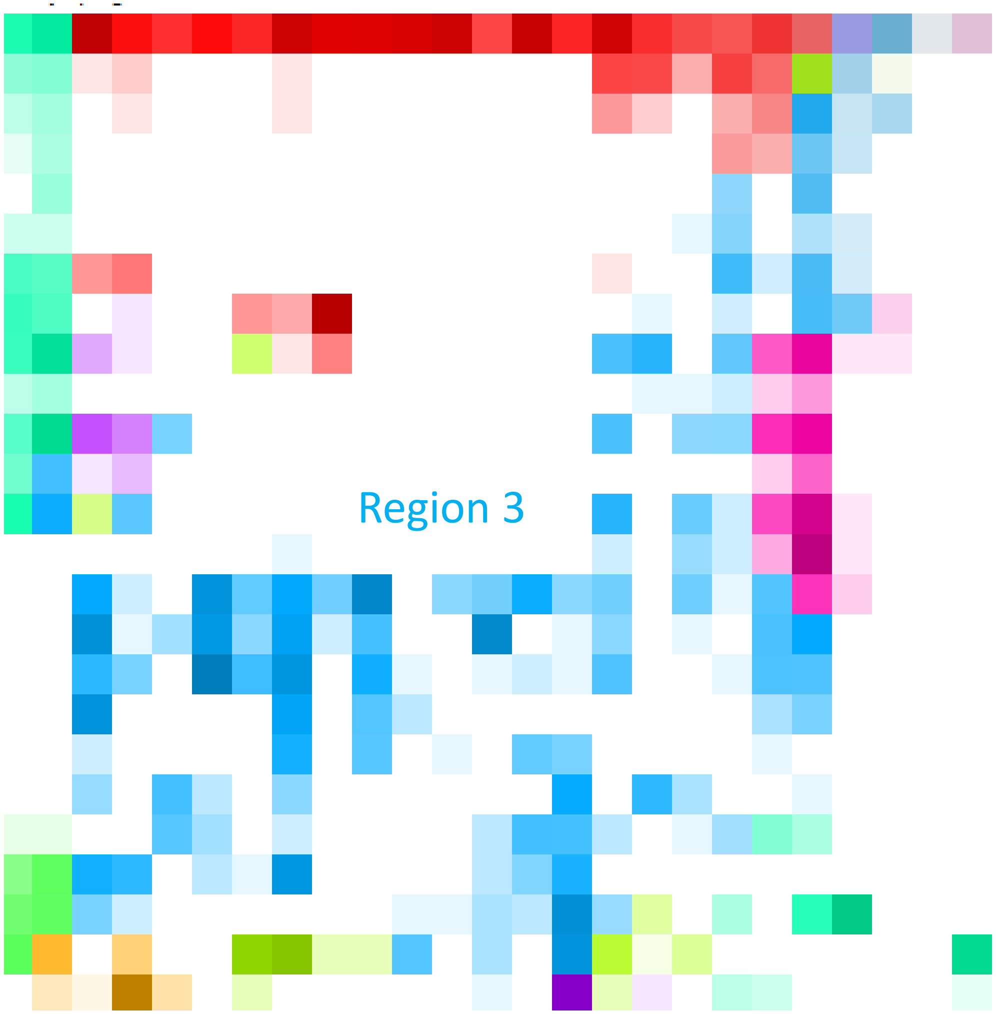
Region 0: Plain text

- Unallocated text files
- Some web cache
- Some Windows files (Help)

Region 1: Web browsing

- Web browsing data & cache
    - htm, asp, css, php
- Unallocated web artifacts

NOTE: “documents” = allocated files and unallocated clusters

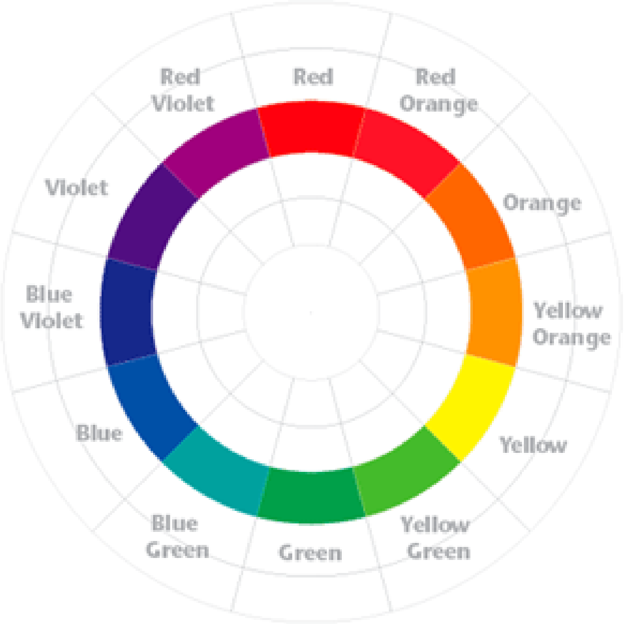


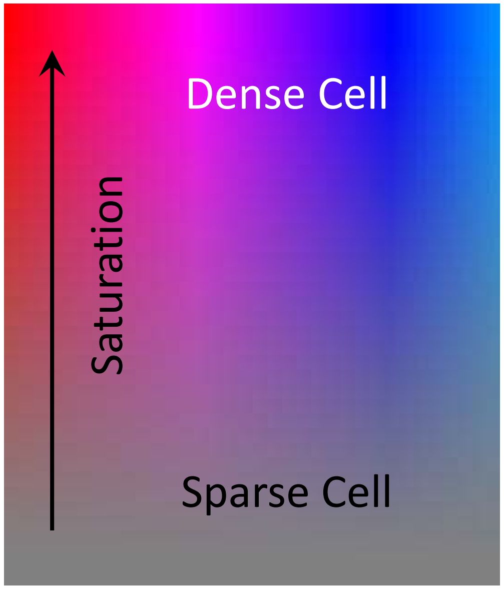
Operating system regions
behave differently…

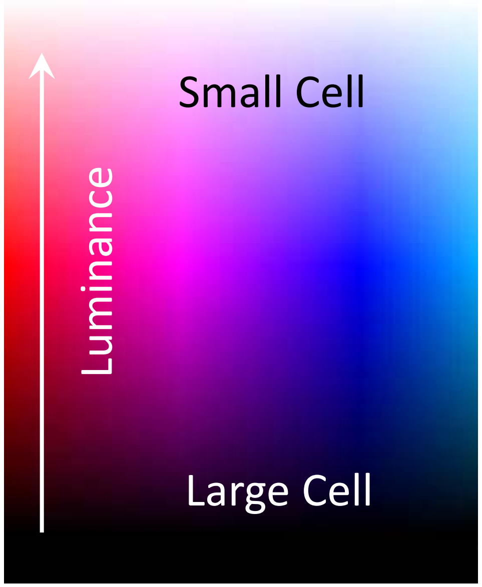
Region 3: Windows System

- Non-compact region, spreads throughout map
- Indicative of OS data
  - Varied within class, but similar relative to other data classes
- Example types:
  - ini, mof, mfl, hlp, inf, pnf

HSL Color Design of Cells







Hue

The color

Provides 'region' view of SOM

Measure of inter-cell similarity

(related cells = same hue)

Saturation

Intensity of color

Measure of cell dispersion

More grey = more disperse

(zero saturation = neutral grey)

Luminance

Brightness of color

Measure of docs/cell

Darker color = more docs

(zero luminance = black)

So… Assess/identify regions via darkest, least grey cells in region

Test Case: Flintstone Counterfeit Caper

The suspects in this case are Fred Flintstone and Barney Rubble. Fred and Barney have been accused of printing, disseminating, and using counterfeit money. The computer being analyzed belongs to Barney. It is believed that Fred and Barney have colluded via email. A search of Fred and Barney’s homes resulted in the seizure of a printing press, advanced printers, and counterfeit currency of various denominations.

- The evidence:
  - Fred’s computer
  - 1.6 GB hard drive, WinXP
  - Ground truth of evidence:
    - Collusion/coordination via web email
    - Web research on how to counterfeit
    - Saved PDF files re: ‘how to’
    - Graphic mages of bills (jpg, gif)

(Poor) String Search List

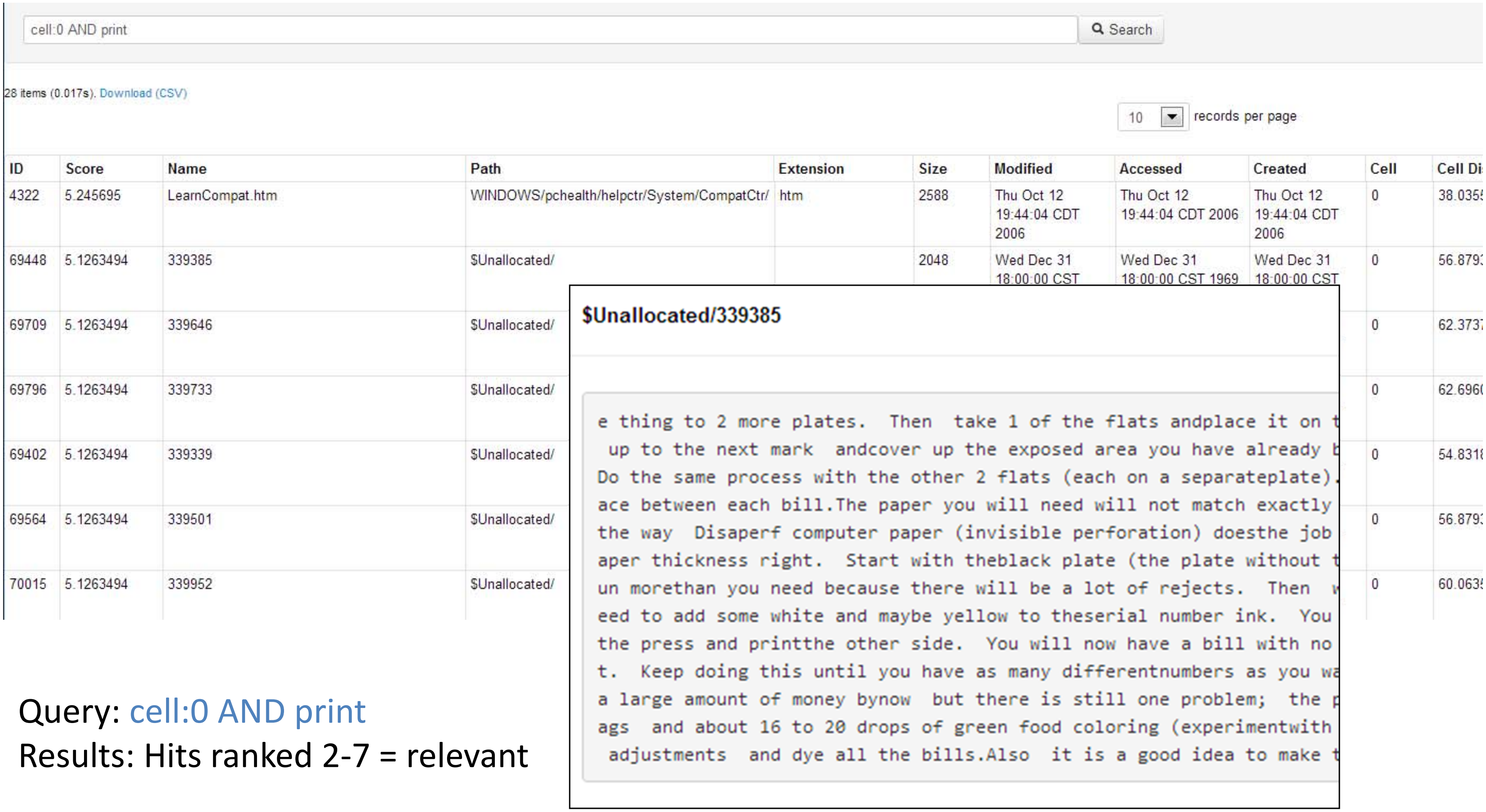
- counterfeit
- dollar
- fake
- bill bill
- **print**
- fred
- flintstone
- barney
- rubble

- Search results for ‘print’
  - 920 ‘docs’ contain ‘print’
  - 5,723 hits in those docs
  - 59 hits relevant (<1% relevant)
    - 19 hits in 11 unallocated clusters
    - 40 hits in 10 allocated files

- Now able to focus search in regions and cells of interest…

Run search queries

- Standard Boolean options
- Include cell(s) of interest to focus query



28 items (0.017s). Download (CSV)

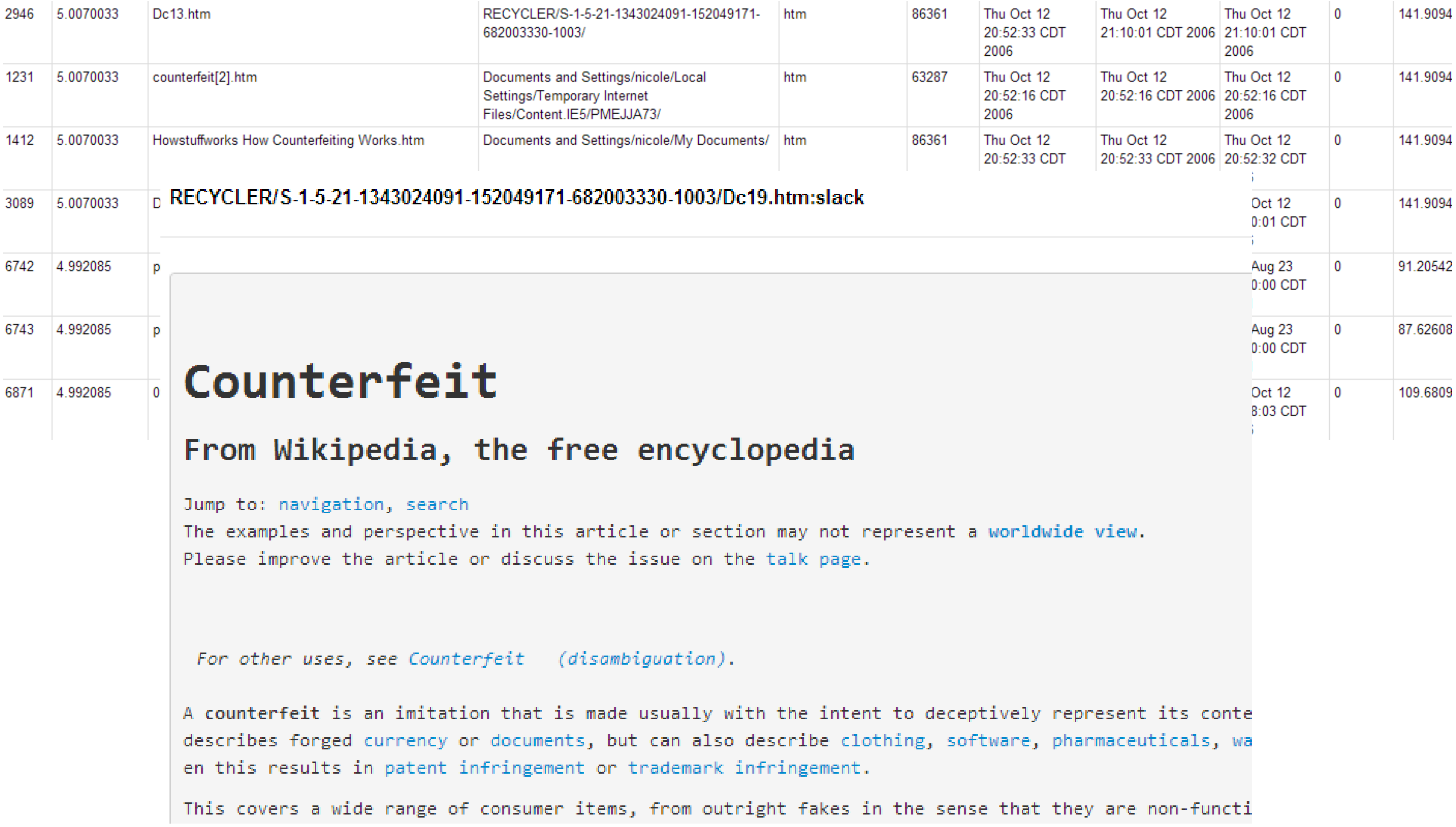
ID	Score	Name	Path	Extension	Size	Modified	Accessed	Created	Cell	Cell Di
4322	5.245695	LearnCompat.htm	WINDOWS/pchealth/helpctr/System/CompatCtr/	htm	2588	Thu Oct 12 19:44:04 CDT 2006	Thu Oct 12 19:44:04 CDT 2006	Thu Oct 12 19:44:04 CDT 2006	0	38.035
69448	5.1263494	339385	$Unallocated/		2048	Wed Dec 31 18:00:00 CST	Wed Dec 31 18:00:00 CST 1969	Wed Dec 31 18:00:00 CST	0	56.879
69709	5.1263494	339646	$Unallocated/						0	62.373
69796	5.1263494	339733	$Unallocated/						0	62.696
69402	5.1263494	339339	$Unallocated/						0	54.831
69564	5.1263494	339501	$Unallocated/						0	56.879
70015	5.1263494	339952	$Unallocated/						0	60.063

$Unallocated/339385

```
e thing to 2 more plates.  Then  take 1 of the flats andplace it on t
 up to the next mark  andcover up the exposed area you have already b
Do the same process with the other 2 flats (each on a separateplate).
ace between each bill.The paper you will need will not match exactly
the way  Disaperf computer paper (invisible perforation) doesthe job
aper thickness right.  Start with theblack plate (the plate without t
un morethan you need because there will be a lot of rejects.  Then w
eed to add some white and maybe yellow to theserial number ink.  You
the press and printthe other side.  You will now have a bill with no
t.  Keep doing this until you have as many differentnumbers as you wa
a large amount of money bynow  but there is still one problem;  the p
ags  and about 16 to 20 drops of green food coloring (experimentwith
 adjustments  and dye all the bills.Also  it is a good idea to make t
```

Query: cell:0 AND print
Results: Hits ranked 2-7 = relevant

… and hits ranked 15-18 are relevant
10/28 hits in cell 0 for ‘print’ are relevant (35.7% relevancy precision)



2946	5.0070033	Dc13.htm	RECYCLER/S-1-5-21-1343024091-152049171-682003330-1003/	htm	86361	Thu Oct 12 20:52:33 CDT 2006	Thu Oct 12 21:10:01 CDT 2006	Thu Oct 12 21:10:01 CDT 2006	0	141.9094
1231	5.0070033	counterfeit[2].htm	Documents and Settings/nicole/Local Settings/Temporary Internet Files/Content.IE5/PMEJJA73/	htm	63287	Thu Oct 12 20:52:16 CDT 2006	Thu Oct 12 20:52:16 CDT 2006	Thu Oct 12 20:52:16 CDT 2006	0	141.9094
1412	5.0070033	Howstuffworks How Counterfeiting Works.htm	Documents and Settings/nicole/My Documents/	htm	86361	Thu Oct 12 20:52:33 CDT	Thu Oct 12 20:52:33 CDT 2006	Thu Oct 12 20:52:32 CDT	0	141.9094
3089	5.0070033	D						Oct 12 0:01 CDT	0	141.9094
6742	4.992085	p						Aug 23 0:00 CDT	0	91.20542
6743	4.992085	p						Aug 23 0:00 CDT	0	87.62608
6871	4.992085	0						Oct 12 8:03 CDT	0	109.6809

RECYCLER/S-1-5-21-1343024091-152049171-682003330-1003/Dc19.htm:slack

Counterfeit

From Wikipedia, the free encyclopedia

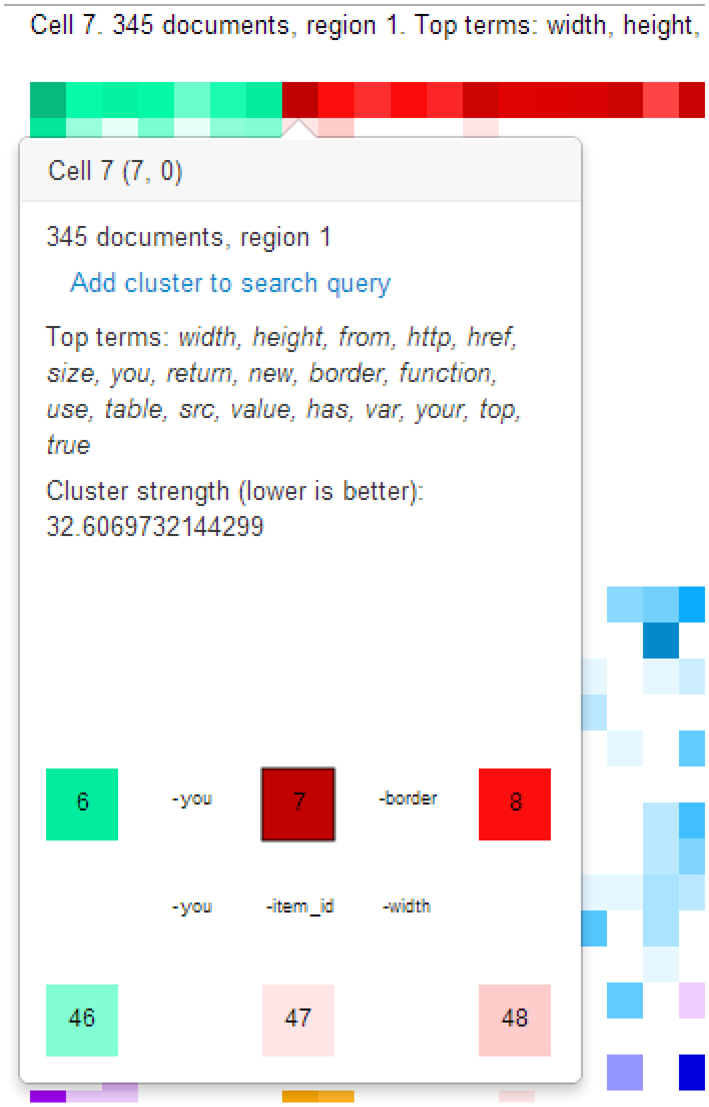
Jump to: navigation, search
The examples and perspective in this article or section may not represent a worldwide view.
Please improve the article or discuss the issue on the talk page.

For other uses, see Counterfeit (disambiguation).

A **counterfeit** is an imitation that is made usually with the intent to deceptively represent its conte
describes forged currency or documents, but can also describe clothing, software, pharmaceuticals, wa
en this results in patent infringement or trademark infringement.

This covers a wide range of consumer items, from outright fakes in the sense that they are non-functi

Examination of Cell 7, Region 1 - Web



Cell 7. 345 documents, region 1. Top terms: width, height,

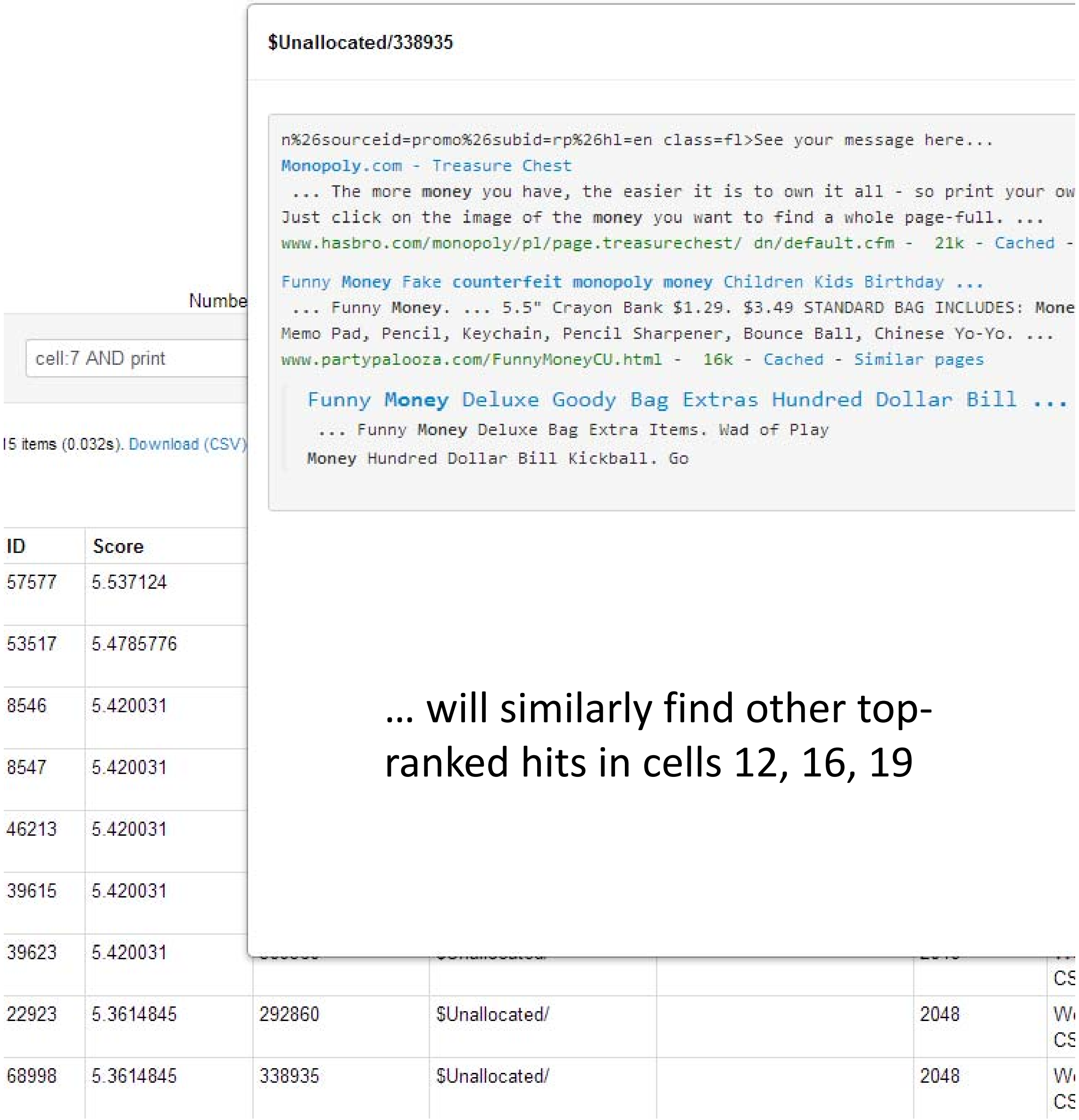
Cell 7 (7, 0)

345 documents, region 1

Add cluster to search query

Top terms: *width, height, from, http, href, size, you, return, new, border, function, use, table, src, value, has, var, your, top, true*

Cluster strength (lower is better): 32.6069732144299



Numbe

cell:7 AND print

I5 items (0.032s). Download (CSV)

ID	Score				
57577	5.537124				
53517	5.4785776				
8546	5.420031				
8547	5.420031				
46213	5.420031				
39615	5.420031				
39623	5.420031				CS
22923	5.3614845	292860	$Unallocated/	2048	W CS
68998	5.3614845	338935	$Unallocated/	2048	W CS

$Unallocated/338935

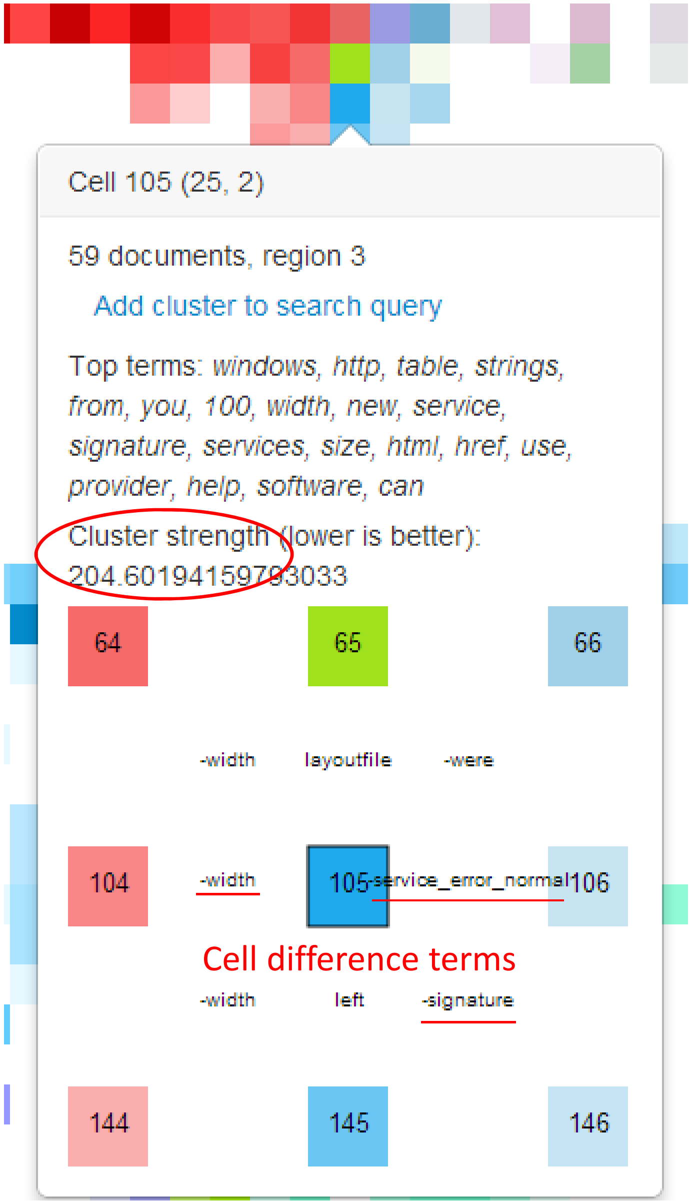
```
n%26sourceid=promo%26subid=rp%26hl=en class=fl>See your message here...
Monopoly.com - Treasure Chest
 ... The more money you have, the easier it is to own it all - so print your ow
Just click on the image of the money you want to find a whole page-full. ...
www.hasbro.com/monopoly/pl/page.treasurechest/ dn/default.cfm -  21k - Cached -

Funny Money Fake counterfeit monopoly money Children Kids Birthday ...
 ... Funny Money. ... 5.5" Crayon Bank $1.29. $3.49 STANDARD BAG INCLUDES: Mone
Memo Pad, Pencil, Keychain, Pencil Sharpener, Bounce Ball, Chinese Yo-Yo. ...
www.partypalooza.com/FunnyMoneyCU.html -  16k - Cached - Similar pages

  Funny Money Deluxe Goody Bag Extras Hundred Dollar Bill ...
   ... Funny Money Deluxe Bag Extra Items. Wad of Play
  Money Hundred Dollar Bill Kickball. Go
```

… will similarly find other top-ranked hits in cells 12, 16, 19

Neighbors and Borders



Cell 105 (25, 2)

59 documents, region 3

Add cluster to search query

Top terms: *windows, http, table, strings, from, you, 100, width, new, service, signature, services, size, html, href, use, provider, help, software, can*

Cluster strength (lower is better): 204.60194159793033

Cell difference terms

- Special consideration of border cells (e.g. #105)
  - Often has high cluster dispersion

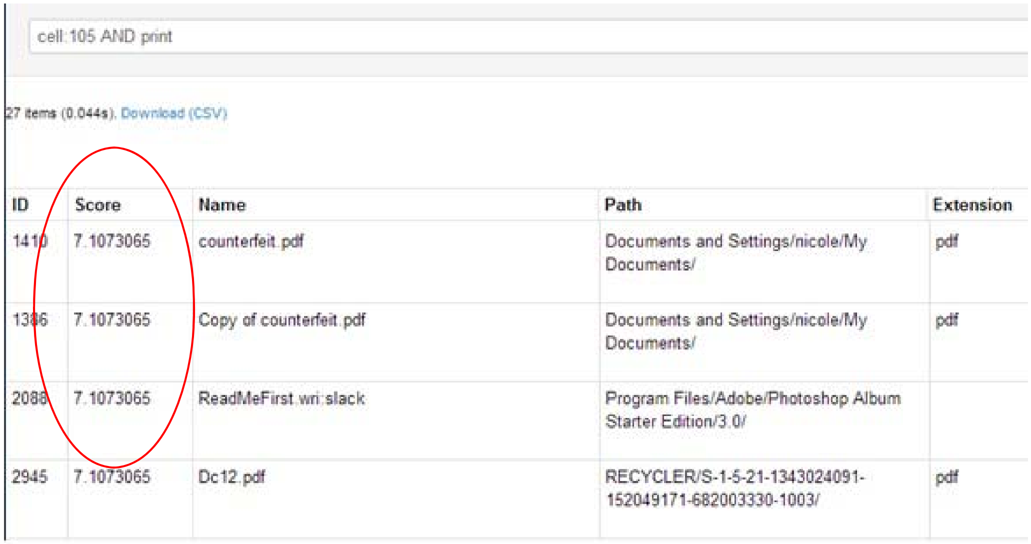
- Cell 105 – Region 3
  - Cluster strength score = 204 (large)
  - Top terms are web related
  - Borders ‘web region’ (Region 2)
  - Primary difference term between cell 105 and red/Region 2 neighbors is ‘width’

- Conclude:
  - Cell 105 node vector is web related
  - Distal ‘docs’ in cell 105 are system related
    - Note difference terms between cell 105 and blue/region 3 neighbor cells

- Analyst actions:
  - Review docs close to cell node vector (centroid) of cell 105 for web artifacts and stop reviewing cell as move into system related material

Ranked Search Hits

- Text string search hit ranking algorithm*
  - 19 measured block-level & hit-level features
  - Block-level features
    - Date/time info
    - Filename/path info
    - Storage location
    - Data/file type**
  - Hit-level features
    - TF-IDF of search term
    - Similarity measures
    - Proximity measures
    - Hit frequency in object
    - Etc.



cell:105 AND print

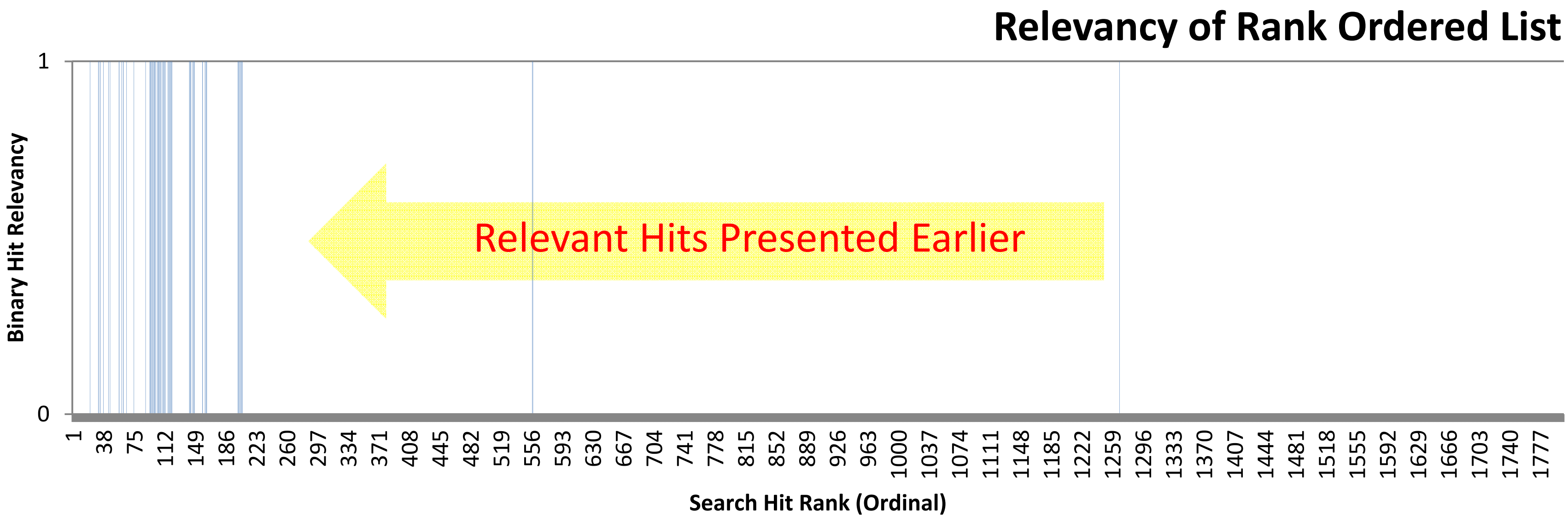
27 items (0.044s). Download (CSV)

ID	Score	Name	Path	Extension
1410	7.1073065	counterfeit.pdf	Documents and Settings/nicole/My Documents/	pdf
1386	7.1073065	Copy of counterfeit.pdf	Documents and Settings/nicole/My Documents/	pdf
2088	7.1073065	ReadMeFirst.wri:slack	Program Files/Adobe/Photoshop Album Starter Edition/3.0/	
2945	7.1073065	Dc12.pdf	RECYCLER/S-1-5-21-1343024091-152049171-682003330-1003/	pdf

* Patent pending on ranking algorithm

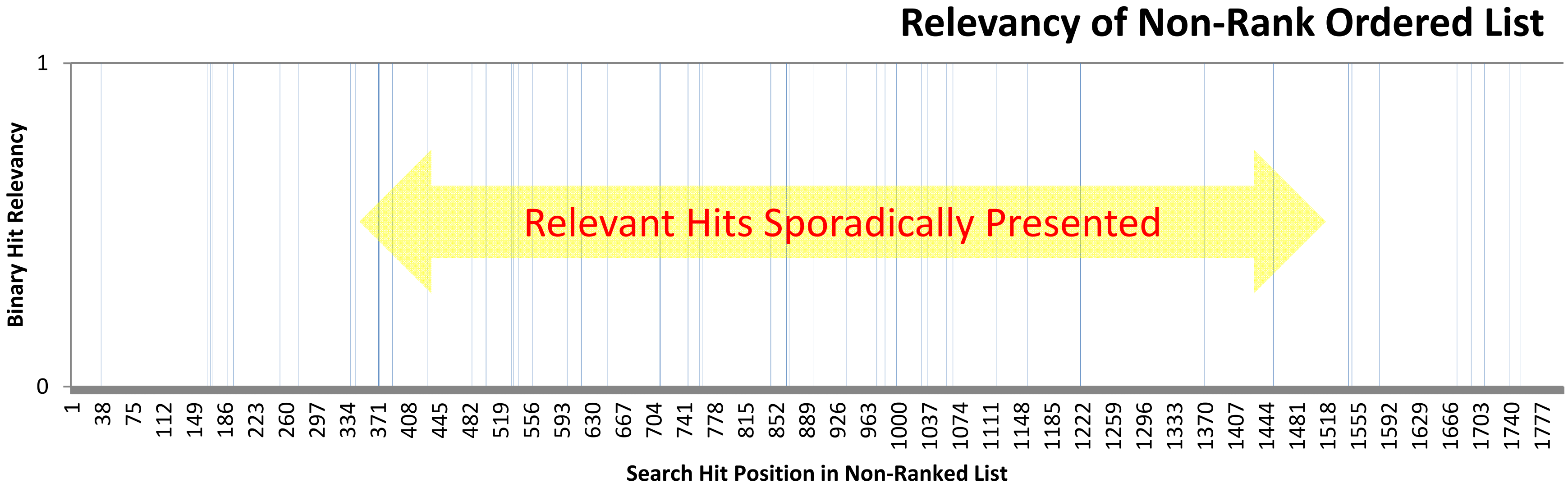
**Sceadan (UTSA GPLv2 licensed data classifier) used to type classify unallocated blocks

Comparative Performance



Relevancy of Rank Ordered List

Relevant Hits Presented Earlier



Relevancy of Non-Rank Ordered List

Relevant Hits Sporadically Presented

Key Points

- Sifter
  - Ingests disk image files (e.g., dd, .E01)
  - Indexes evidence via Lucene and Tika
  - Clusters and ranks string search hits
  - Provides means for improved exploratory search
    - Visual SOM clustering of hits (cells and regions)
  - Improves analytical efficiency
    - Rank ordered lists
    - Ability to include/exclude clusters with query
  - Leverages other research advances
    - Empirically derived digital forensic stoplists
    - Open source naïve statistical data type classifier (Sceadan)

Future Dev. Plans (hopes)

- Linux/OSX installers
- Added features
  - Table view sorting
  - Hit level bookmarking
  - Bookmark removal capability
  - Ranking algorithm parameter tuning via GUI
    - Search term prioritization
    - Temporal reference point (e.g. date of hack vs. date of analysis)
    - Data type prioritization
- Automated mapping configuration
  - Iterative, sample-based SOM build/re-build
- SOM usability improvements
  - Layout stability
- Ranking function improvements
  - Additional R&D (train ranking algorithm(s) on more cases)
  - Case type specific ranking algorithms
- TSK 3.0 integration

Sifter v1.0 (beta)

Released 10/16/13

Apache 2.0 Licensed

Email Nicole Beebe for a copy

(source for Mac/Linux, Windows Installer)

Nicole.Beebe@utsa.edu

210-458-8040 (w) 210-269-5647 (c)

COMMENTS?? QUESTIONS??