

# All Search is not Created Equal: Advanced Text Search within the Text Processing/Mining Continuum



<http://library.thinkquest.org/19300/data/Hist/slide4.htm>

**Tim Allison, Ph.D.**  
**November 2015**

Approved for Public Release;  
Distribution Unlimited. Case  
Number 15-3426

**MITRE**

# Overview

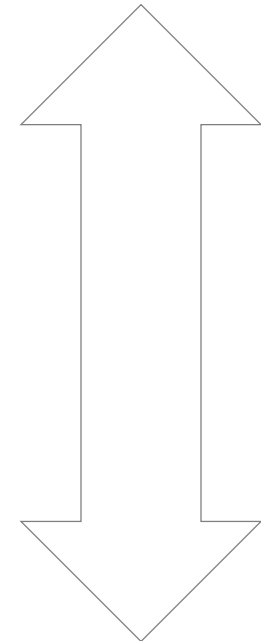
---

- **Text Processing and Text Mining**
- **Considerations for selection/integration and small bit on evaluation**
- **Desktop search prototype examples**

# Text Processing and Text Mining\*

- Event resolution
- Event extraction
- Machine translation
- Entity resolution
- Entity extraction
- Anomaly detection, Clustering, Correlation
- Document similarity
- Document zoning
- Document categorization
- Entity search (name matching)
- Search

Very Difficult



Easy

---

Text Extraction and/or Optical Character Recognition (OCR)

# As Capabilities Becomes More Complex

- The tools are less accurate.
- There are fewer vendors and less competition in the marketplace.
- The tools become more sensitive to unreliable/noisy text.
- The tools become more sensitive to heterogeneity in the documents.
- The tools require more care and feeding/tuning/professional services to yield reasonable results.

**Note: There must be a good fit between the task and the nature of the documents. Even with the best tools and the best tuning, some of these tools may yield nothing of value on some document sets.**

# To integrate or not to integrate?

- The hierarchy presented two slides ago must not be taken as an indicator of the acquiring organization's technical maturity!
- The sole question in determining what makes sense for an organization is: does the value returned by a given capability exceed the amount of resources required to acquire it, integrate it and keep it up to date.
- Added value might include:
  - Organization can carry out current tasks more quickly.
  - Organization can carry out current tasks more accurately.
  - Organization can gain new insight and start performing entirely new tasks to meet mission needs.

# Common Approach

---

- **Vendor: tool can do “x”.**
- **System designer: Let’s buy the tool; “x” would be great!**
- **Implementation team: Wait, the tool requires several preprocessing steps/conditioning of the data/fields we don’t even have/tuning of the data. This requires far more than we thought, and the performance out of the box is awful on my data.\*\*\***
- **Vendor: “But it can do x. Need professional services to help you tune for your data?”**

# Planning

---

- **What exactly are we trying to accomplish?**
- **Do we have an ongoing need on a mostly homogeneous data stream?**
- **What exactly does the tool need to do a good job?**
- **What do we need to do to our data/to the tool to make it work for our data?**
- **Can the tool be tuned for our data?**
- **How will we evaluate whether it is working?**
- **Are the likely costs of acquisition, training and professional services worth the expected performance of the tool for our mission?**

# Evaluation (the often missed step)

- **Intrinsic**

- Common scientific measures of success for different capabilities on a representative test set: Accuracy, Precision, Recall, F-Measure, Mean Average Precision, Character Error Rate, Cluster quality...

- **Extrinsic**

- Does the tool help the user accomplish the tasks (time/volume/quality)
  - Formal: Rigorous study on controlled task
  - Informal: Opinion survey/gut-feeling of analysts



# Example Text

---

**DMV clerk Tracey Lynette Jones, 33, of Long Beach, surrendered to federal authorities Monday and was arraigned Monday afternoon in U.S. District Court. Jones, who entered a not guilty plea, was released on a \$25,000 bond and ordered to stand trial Nov. 24 before U.S. District Judge Cormac J. Carney.**

**Four of the remaining defendants were arrested last month ... They are: Wilfredo Montero, 36, of Los Angeles; Adolfo Maria Cruz, 47, of Corona; Roberto Ruiz, 35, of Tustin; and Jose Cruz, 49, of Corona.**

**Three of the four defendants arrested in September have been freed on bond. Cruz remains in custody.**

**The sixth defendant charged in the case, Jorge “Diablo” Perez, aka Pedro Josue Figueroa-Marquez, 33, originally from Mexico, remains at large and is being sought by authorities.**

# Example Entity Extraction for “Person”

DMV clerk **Tracey Lynette Jones**, 33, of Long Beach, surrendered to federal authorities Monday and was arraigned Monday afternoon in U.S. District Court. **Jones**, who entered a not guilty plea, was released on a \$25,000 bond and ordered to stand trial Nov. 24 before U.S. District Judge **Cormac J. Carney**.

Four of the remaining defendants were arrested last month ... They are: **Wilfredo Montero**, 36, of Los Angeles; **Adolfo Maria Cruz**, 47, of Corona; **Roberto Ruiz**, 35, of **Tustin**; and **Jose Cruz**, 49, of Corona.

Three of the four defendants arrested in September have been freed on bond. Cruz remains in custody.

The sixth defendant charged in the case, **Jorge “Diablo” Perez**, aka Pedro Josue Figueroa-Marquez, 33, originally from Mexico, remains at large and is being sought by authorities.

# Example Entity Extraction for “Person”: Hits, Misses, False Positives, Incorrect Extents

DMV clerk **Tracey Lynette Jones**, 33, of Long Beach, surrendered to federal authorities Monday and was arraigned Monday afternoon in U.S. District Court. **Jones**, who entered a not guilty plea, was released on a \$25,000 bond and ordered to stand trial Nov. 24 before U.S. District Judge **Cormac J. Carney**.

Four of the remaining defendants were arrested last month ... They are: **Wilfredo Montero**, 36, of Los Angeles; **Adolfo Maria Cruz**, 47, of Corona; **Roberto Ruiz**, 35, of **Tustin**; and **Jose Cruz**, 49, of Corona.

Three of the four defendants arrested in September have been freed on bond. **Cruz** remains in custody.

The sixth defendant charged in the case, **Jorge “Diablo” Perez**, aka **Pedro Josue Figueroa-Marquez**, 33, originally from Mexico, remains at large and is being sought by authorities.

# Example Performance Calculation: Precision/Recall/F-Measure

Total person name mentions in document: 10

Total accurately identified: 7

Misses: 3 (or is it 2?)

False positives: 1

Incorrect extent: 1

For convenience, let's count incorrect extent as a miss

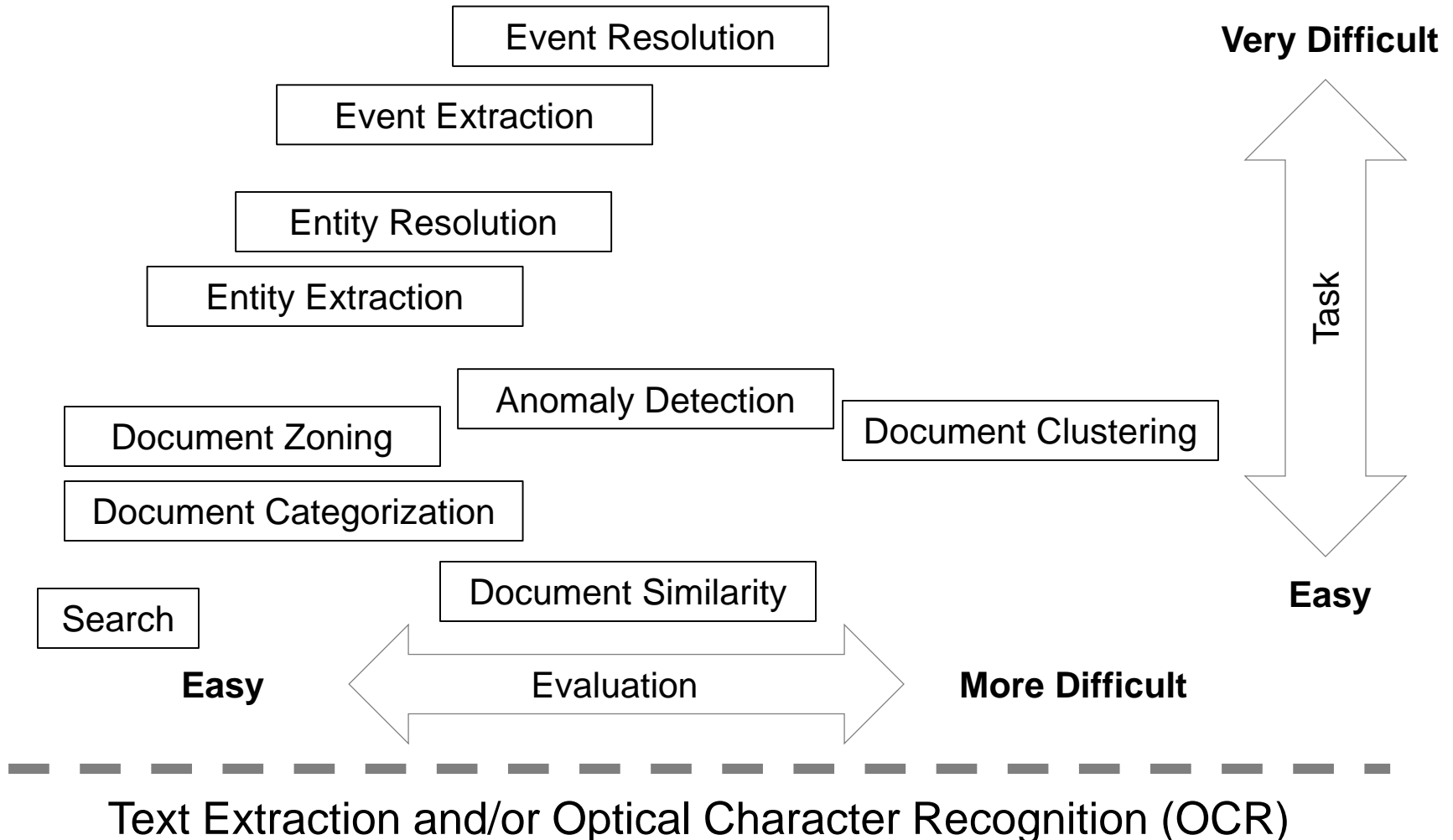
	True Entity	Not an Entity
Extracted by system	7	1
Not extracted by system	3	

Precision:  $7/8$  (0.875)

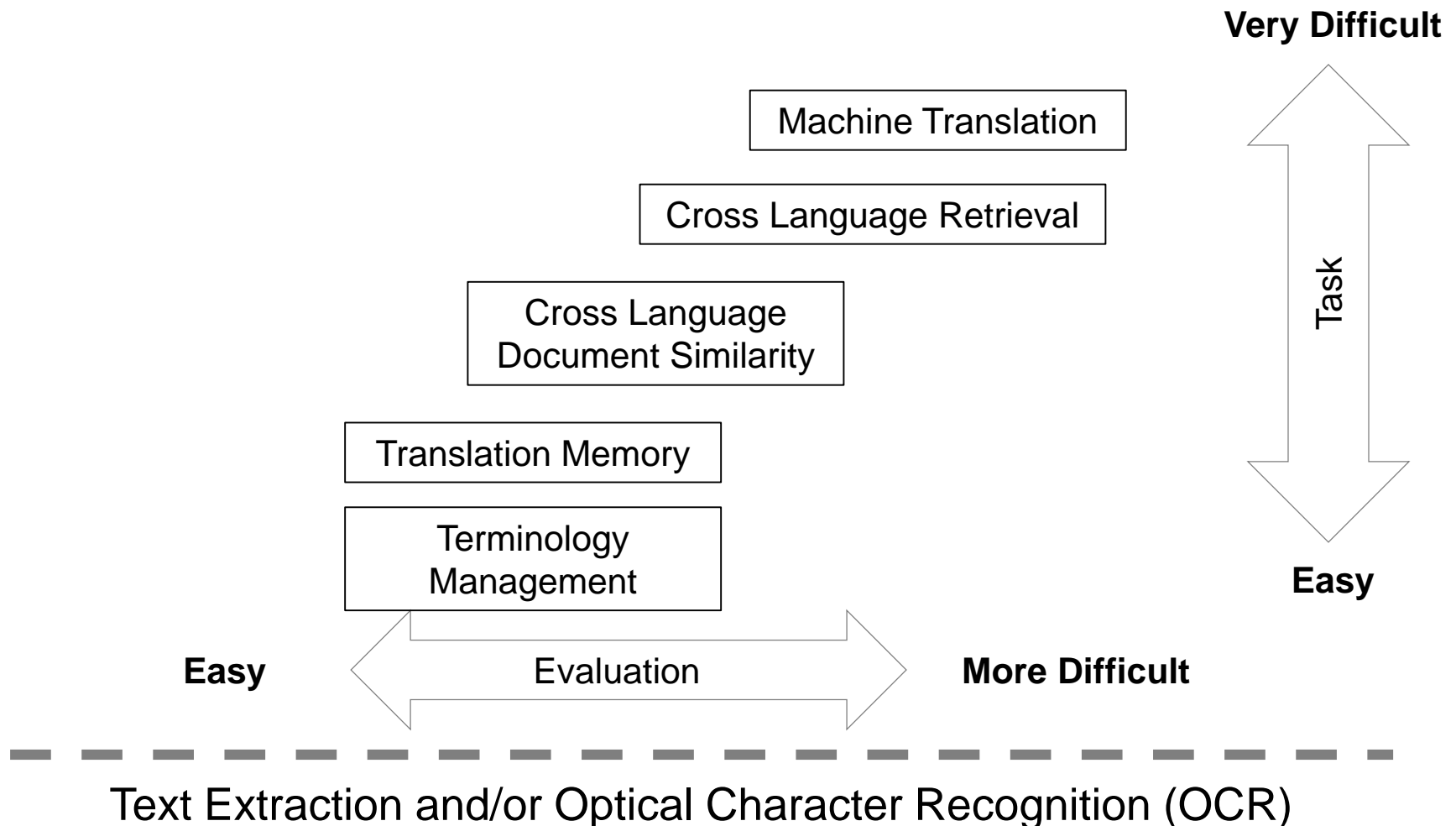
Recall:  $7/10$  (.70)

F-Measure: (.77)

# Monolingual Text Processing and Text Mining\*



# Multilingual Text Processing and Text Mining\*



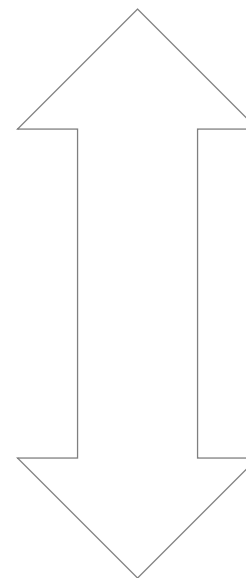
# Text, but what about multimedia?

- **Same principles apply, but performance is typically much lower.**
- **Speech**
  - Language identification
  - Speech Activity Detection
  - Speaker Identification
  - Automatic Speech Recognition (ASR)
  - Search (ASR or phonetic?)
  - Machine translation (speech to speech)
  - Spoken language understanding
- **Images/Video**
  - Optical character recognition from a photo/video
  - Image search
  - Object detection/search
  - Change detection
  - Facial recognition/search
  - Event detection

# And now, to focus on search...

- Event resolution
- Event extraction
- Machine translation
- Entity resolution
- Entity extraction
- Anomaly detection, Clustering, Correlation
- Document similarity
- Document zoning
- Document categorization
- Entity search (name matching)
- **Search**

Very Difficult



Easy

---

Text Extraction and/or Optical Character Recognition (OCR)



# Not all search tasks are created equal

- **Basic search:** answer a specific information need. The user needs one or two of the most relevant documents.
- **Advanced analysis/legal discovery:** find every time x (or some variant of x) appears in our documents. Help me quickly make sense of who/what/where/when/why.
- **Document retrieval vs. sense making vs. discovery**

# Rhapsode: Quick Overview

- **MITRE prototype – in process of tech transfer**
- **Combination of open source and custom code.**
- **Enables desktop search and corpus exploration.**
- **Search features include complex text and basic geo-spatial search.**
- **Four main search features:**
  - Basic search
  - Concordance search
  - Co-occurrence search
  - Variant finder
- **Primary Goal: Demonstrate state of the possible to enable requirements development and drive enterprise level change.**
- **Secondary Goals:**
  - Enable users to experiment in a sandbox with components that are part of major enterprise search systems (Tika and Lucene).
  - Build new components and improve existing components in open source packages that **many** of our sponsors are using.

# Not intended to be this



[http://www.thetorquereport.com/assets\\_c/2011/05/2012\\_lamborghini\\_aventador\\_lp700\\_4\\_fr\\_new3-11397.html](http://www.thetorquereport.com/assets_c/2011/05/2012_lamborghini_aventador_lp700_4_fr_new3-11397.html)

# Closer to this



[http://mitre.org/news/digest/advanced\\_research/10\\_08/sensing.html](http://mitre.org/news/digest/advanced_research/10_08/sensing.html)

# Query Syntax

- **Key:** charlie
- **Fuzzy:** charlie~0.8
- **Wildcard:** ch?rl\*
- **Regex term queries:** /ch[aeiou]rl[aeiou]+/
- **Boolean:** +charlie +foxtrot
- **Phrase:** “charlie foxtrot”
- **Phrase with slop:** “charlie foxtrot”~10
- **Range:** date:[201109 TO 201112]
- **Complex nesting of proximity with fuzzy/wildcard:**
  - [charlie~0.6 foxtr\* (snafu oops doh)]~10
- **SpanNotNear queries:**
  - [ fever (bieber travlota~1 “Saturday night”)]!~15,15
- **Field:**
  - content:charlie file:foxtrot\*doc

# Basic Search

The screenshot displays the Rhapsode Search Prototype v0.1 web application. The browser address bar shows the URL `http://localhost:8090/rhapsode/basic/`. The application has a navigation bar with five tabs: BASIC SEARCH, CONCORDANCE SEARCH, CO-OCCURRENCES, VARIANT COUNTER, and ANALYZER TEST. The 'BASIC SEARCH' tab is active.

**Basic Search**

Query:

Results per page:

Language Direction, Left to Right: ☒ Right to Left: ☐

There were 40 results within 5793 documents.

1.	<a href="#">091127santaana.htm.txt</a>	<input type="checkbox"/>	11/27/2009	Sheriff's deputies rescued a 4-year-old girl from a suspected human <b>smuggler</b> yesterday. Authorities.... Once the two arrived in the United States, the human <b>smuggler</b> refused to return the child to the
2.	<a href="#">080509phoenix.htm.txt</a>	<input type="checkbox"/>	05/09/2008	PHOENIX - A convicted human <b>smuggler</b> from Michoac n, Mexico, was sentenced to life in prison today... from the Yuma County Sheriff's Office and the Federal Bureau of Investigation. "This <b>smuggler</b>
3.	<a href="#">080611mavaguez.htm.txt</a>	<input type="checkbox"/>	06/11/2008	the alleged Puerto Rican <b>smuggler</b> and initially located and recovered approximately 200 bricks of... over custody of the drugs, the vessel and the alleged <b>smuggler</b> to ICE special agents. ICE agents
4.	<a href="#">090306santaana.htm.txt</a>	<input type="checkbox"/>	03/06/2009	, who was living illegally in the United States, paid a <b>smuggler</b> \$14,000 to bring his 24-year-old... one <b>smuggler</b> to another as they made their way through Mexico and eventually crossed the border
5.	<a href="#">090507madison.htm.txt</a>	<input type="checkbox"/>	05/07/2009	the serial numbers so they could not be traced back to him. A Canadian <b>smuggler</b> then traveled to... gun <b>smuggler</b> admitted that he picked up the firearms from Sundal and was on his way back to Canada
6.	<a href="#">100920phoenix.htm.txt</a>	<input type="checkbox"/>	09/20/2010	<b>smuggler</b> for more than four months. On Thursday, ICE HSI agents assigned to the Drop House Response..., and took custody of the suspected <b>smuggler</b> , Mario Fernandez-Fernandez. The 25-year-old Mexican
7.	<a href="#">100810tucson.htm.txt</a>	<input type="checkbox"/>	08/10/2010	pickup truck of the suspected <b>smuggler</b> escort a tractor trailer from a Tucson-area truck stop to a
8.	<a href="#">081024boston.htm.txt</a>	<input type="checkbox"/>	10/24/2008	supporting affidavit alleges that Santandrea-Giron was identified as an alien <b>smuggler</b> from Guatemala. It..., another putative alien <b>smuggler</b> , led to a series of personal meetings in El Salvador with the
9.	<a href="#">100621sarasota.htm.txt</a>	<input type="checkbox"/>	06/21/2010	reported that an alien <b>smuggler</b> was holding his 30-year-old brother, "Cesar," for an unpaid smuggling fee... Guatemala to Florida. They paid smuggling fees first to a <b>smuggler</b> in Guatemala to transport Cesar
10.	<a href="#">101208houston.htm.txt</a>	<input type="checkbox"/>	12/08/2010	location by an unidentified <b>smuggler</b> and that he was extremely ill on arrival. Aburto purchased a... the interviews with the aliens, Guzman-Villa was identified as having been an alien <b>smuggler</b> for at

**MITRE**

# Concordance Search

The screenshot shows a web browser window with the address bar displaying `http://localhost:8090/rhapsode/concordance/`. The page title is "Rhapsode Search Prototype, v0.1". Below the title is a navigation bar with five tabs: "BASIC SEARCH", "CONCORDANCE SEARCH" (which is active), "CO-OCCURRENCES", "VARIANT COUNTER", and "ANALYZER TEST".

The "Concordance Search" section contains the following form fields and controls:

- Concordance Query:
- Filter Query:
- Number of Words Before:  Number of Words After:
- Sort Order:
- Show Duplicate Windows: ☒ Hide Duplicate Windows: ☐
- Language Direction, Left to Right: ☒ Right to Left: ☐
- 
- 

Below the form, a message states: "There were 51 windows in 40 documents out of a total of 5793 documents." Below this message is a table with four columns: document ID, date, text snippet, and the word "smuggler".

<a href="#">091130newyorkcity.htm.txt</a>	<input type="checkbox"/>	11/30/2009	the victim became pregnant with Salazar's child, Salazar paid a	smuggler	to transport himself and the victim across the U.S.-Mexico
<a href="#">090306santaana.htm.txt</a>	<input type="checkbox"/>	03/06/2009	who was living illegally in the United States, paid a	smuggler	\$14,000 to bring his 24-year-old wife, Ana R
<a href="#">100330washingtondc.htm.txt</a>	<input type="checkbox"/>	03/30/2010	Mexico illegally, at which point he referred aliens to a	smuggler	who brought the aliens into the United States. In one
<a href="#">100621sarasota.htm.txt</a>	<input type="checkbox"/>	06/21/2010	Guatemala to Florida. They paid smuggling fees first to a	smuggler	in Guatemala to transport Cesar illegally into the United States



# Concordance Search

U.S. by sea the reality that anytime you trust a	smuggler	you are putting your life in grave danger," said William
affidavit alleges that Santandrea-Giron was identified as an alien	smuggler	from Guatemala. It is alleged that a number of telephone
aliens, Guzman-Villa was identified as having been an alien	smuggler	for at least two years with his wife, Cuadros, and
females were brought to the United States by an alien	smuggler	and forced to work at a bar to pay off
United States contacted MCSO deputies and reported that an alien	smuggler	was holding his 30-year-old brother, "Cesar," for an
a load of 13 aliens being transported by another alien	smuggler	To hijack the pickup truck in which the 13 aliens
Santandrea-Giron and a cooperating witness, another putative alien	smuggler	led to a series of personal meetings in El Salvador
Portillo De Cruz, 36, of Laredo, and a third alien	smuggler	Bertha Alicia "La Guera" Esquivel, 40, of Nuevo Laredo, Tamaulipas
over custody of the drugs, the vessel and the alleged	smuggler	to ICE special agents. ICE agents, personnel of the Key
BROWNSVILLE, Texas A convicted ammunition	smuggler	was sentenced on Wednesday to more than 11 years in
target a load of 13 aliens being transported by another	smuggler	In the effort to hijack the pickup truck in which
target a load of 13 aliens being transported by another	smuggler	In the effort to hijack the pickup truck in which
Monterrey, Mexico, Gerald's co-conspirator paid a Mexico-based	smuggler	to illegally transport the individual across the Mexico-U.S. border
they could not be traced back to him. A Canadian	smuggler	then traveled to Wisconsin to pick up the firearms from
that has destroyed lives and ravaged communities. As a cocaine	smuggler	he was an integral part of the illegal distribution network
PHILADELPHIA - A drug	smuggler	was sentenced Oct. 1 to 20 years in prison after
SEATTLE - A Canadian truck driver and drug	smuggler	with ties to the Hells Angels motorcycle gang was sentenced
of his arrest. "This sentencing not only takes another drug	smuggler	off the street, it demonstrates the serious consequences that narcotics



# Co-Occurrence Counts

## Rhapsode Search Prototype, v0.21

BASIC SEARCH    CONCORDANCE SEARCH    CO-OCCURRENCES    VARIANT COUNTER    ANALYZER TEST

### Concordance Co-Occurrence Counter

Concordance Query:  And: ☐ Or: ☒  
 Geo Query:  Query Radius in KM:  And: ☐ Or: ☒  
 Filter Query:   
 Number of Words Before:  Number of Words After:   
 Number of Results:   
 Minimum Term Frequency:   
 Minimum IDF:   
 Minimum NGram:  Maximum NGram:   
 Maximum Windows   
 Count Duplicate Windows: ☒ Ignore Duplicate Windows: ☐

Search

There were 133 windows in 104 documents out of a total of 5,793 documents.

Rank	Term	Term Frequency	IDF	TF*IDF
1.	semi automatic	31	8.5	262.9
2.	automatic	31	4.3	132.4
3.	semi	31	4.2	130.5
4.	caliber	30	4.2	125.3
5.	ashike	11	8.7	95.3
6.	rifles	22	4.3	94.5
7.	ammunition	22	3.7	81.4

# Co-Occurrences: handg\*

There were 133 windows in 104 documents out of a total of 5,793 documents.

Rank	Term	Term Frequency	IDF	TF*IDF
1.	semi automatic	31	8.5	262.9
2.	automatic	31	4.3	132.4
3.	semi	31	4.2	130.5
4.	caliber	30	4.2	125.3
5.	ashike	11	8.7	95.3
6.	rifles	22	4.3	94.5
7.	ammunition	22	3.7	81.4
8.	loaded	19	3.8	72.6
9.	40 caliber	11	6.5	71.4
10.	rounds of ammunition	8	8	64.3
11.	9mm	10	5.4	54.5
12.	rifle	13	4.1	53.6
13.	shotguns	9	5.8	52
14.	assault rifles	7	6.9	48.1
15.	9 mm	6	7.9	47.4
16.	shotgun	9	5.1	46

# Co-Occurrences: “counterfeit”

There were 2,047 windows in 413 documents out of a total of 5

Rank	Term	Term Frequency	IDF	TF*IDF
1.	goods	537	2.7	1458.8
2.	products	353	3	1075.9
3.	counterfeit	398	2.6	1051.2
4.	merchandise	195	3.7	713.8
5.	items	211	2.5	521.2
6.	counterfeit products	80	5.7	455.1
7.	only on keeping	79	5.5	438.3
8.	pharmaceuticals	95	4.6	434.2
9.	selling	146	2.9	427
10.	integrated circuits	42	9.8	413.5
11.	seized	250	1.6	398.9
12.	cisco	65	6.1	396.5
13.	trafficking	219	1.7	371.4
14.	products off	65	5.6	363.2
15.	china	111	3.2	351.1

# Co-Occurrences: “counterfeit” in documents that also contain “Miami”

Rank	Term	Term Frequency	IDF	TF*IDF
1.	products	38	3	115.8
2.	cigars	16	7.1	112.9
3.	merchandise	28	3.7	102.5
4.	super bowl	9	10.5	94.7
5.	goods	26	2.7	70.6
6.	counterfeit	24	2.6	63.4
7.	bowl	11	5.6	61.3
8.	parts	15	3.9	58.2

# Variant Counter: Simple Single Term

## Variant Counter

Concordance Query: salmonella~2

Filter Query:

Number of Results: 20

Show Code Points: ☐

Normalize: ☒

Simple Single Term Search

Advanced Search

There were 110 unique terms out of a total of 339 documents.

Term	Document Frequency	IDF
salmonella	305	0.1
sulmonella	35	2.3
Osalmoneella	26	2.6
sulmonellu	17	3
salnzonella	16	3.1
salmonellu	15	3.1
saimonella	14	3.2
snlmonella	11	3.4
salmonel	9	3.6
salmonelia	9	3.6
xsalmoneella	7	3.9

# Variant Finder: Phrases

## Variant Counter

Concordance Query: "osama~1 bin laden~1" "bin laden~1"

Filter Query:

Number of Results: 20

Show Code Points: ☐

Normalize: ☒

Simple Single Term Search

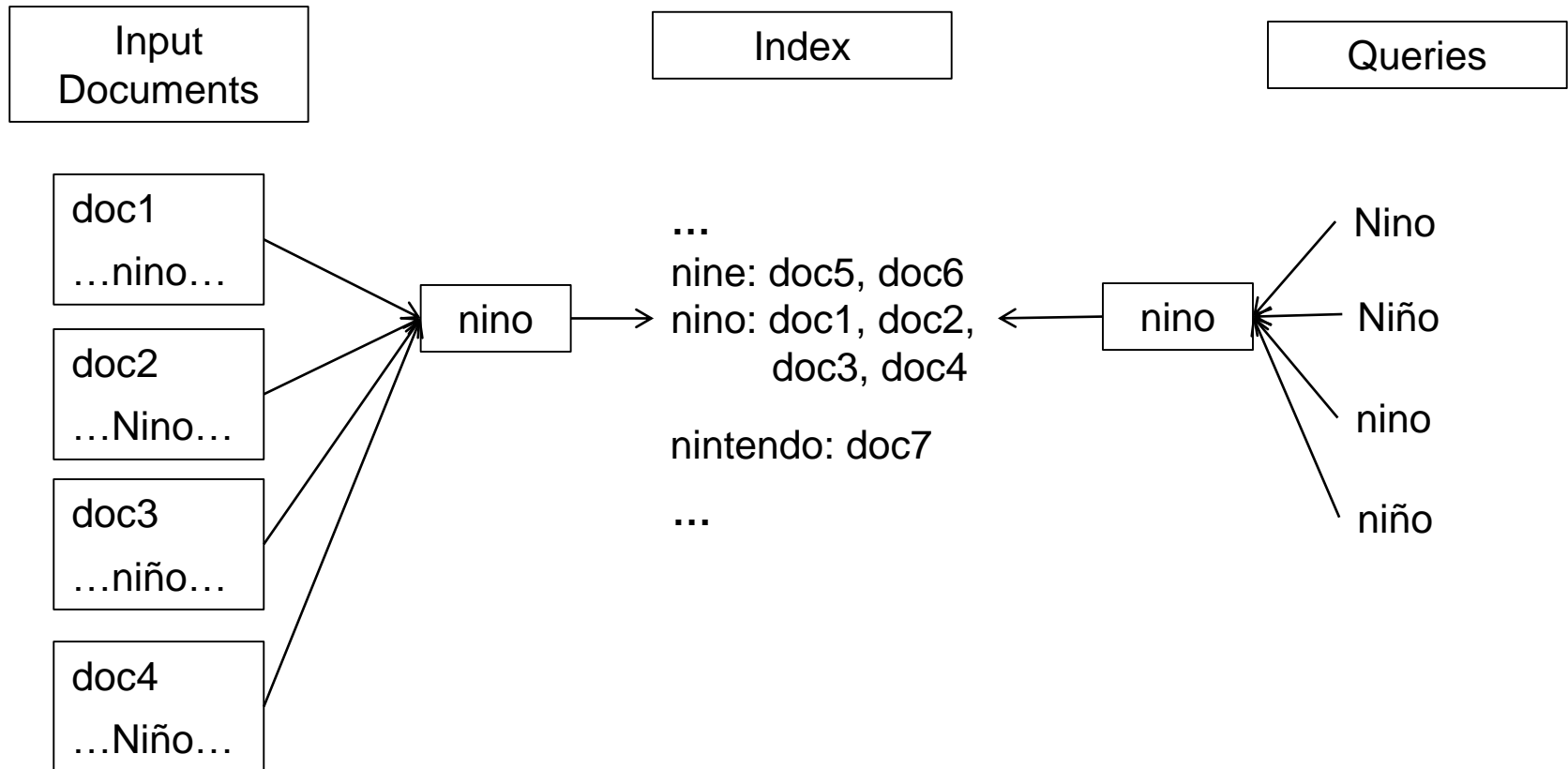
Advanced Search

There were 4 unique terms that appeared 31 times in 13 documents out of a total of 9,779 documents.

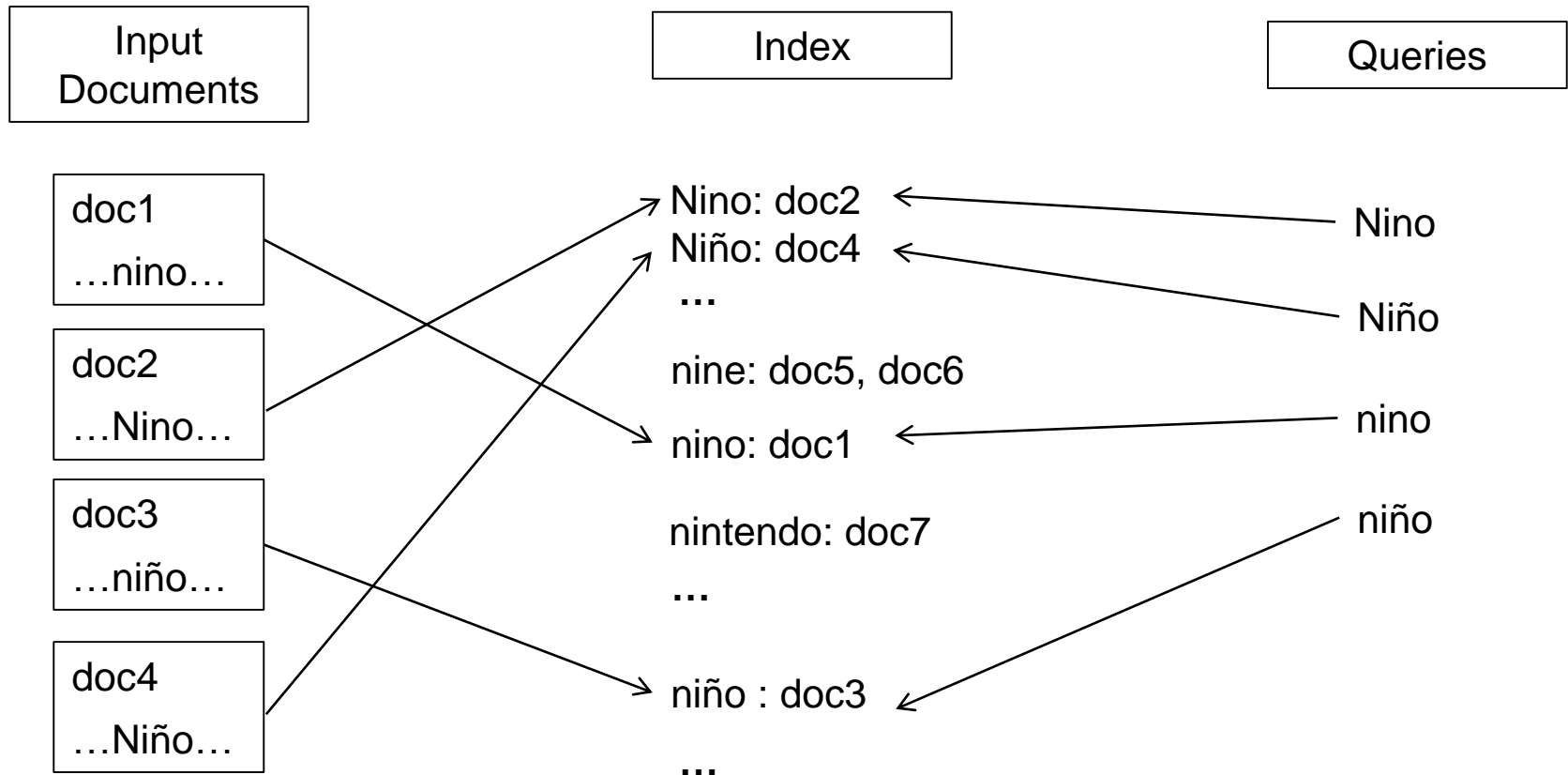
Term	Document Frequency	IDF	Term Frequency
osama bin laden	11	6.8	28
bin laden	1	9.2	1
osama bin ladin	1	9.2	1
usama bin laden	1	9.2	1

("osama bin laden", "bin laden", "osama bin ladin", "usama bin laden")

# Example of Normalization for Search



# Example of Failure to Normalize for Search





# Variant Counter (Advanced)

## Variant Counter

Concordance Query:

Filter Query:

Number of Results:

Show Code Points: ☒

Normalize: ☐

Language Direction, Left to Right: ☐ Right to Left: ☒

Simple Single Term Search

Advanced Search

QueryString	Unicode Code Points (Storage Order)
أمريكا	622, 645, 631, 6cc, 6a9, 627

There were 78 unique terms that appeared 202,229 times in 53,426 documents out of a total of 779,109 documents.

Term	Unicode Code Points (Storage Order)	Document Frequency	IDF	Term Frequency
أمريكا	622, 645, 631, 6cc, 6a9, 627	24000	3.5	75009
أمريكا	622, 645, 631, 64a, 643, 627	16371	3.9	66017
أمريكا	627, 645, 631, 6cc, 6a9, 627	6892	4.7	20069
أمريكا	622, 645, 631, 64a, 6a9, 627	5634	4.9	18376
أمريكا	627, 645, 631, 64a, 643, 627	3138	5.5	9785
أمريكا	622, 645, 631, 6cc, 643, 627	1821	6.1	6252
أمريكا	627, 645, 631, 64a, 6a9, 627	1600	6.2	5059
أمريكا	627, 645, 631, 6cc, 643, 627	417	7.5	1112
أمريكا	623, 645, 631, 64a, 643, 627	115	8.8	257
أمريكا	622, 645, 640, 631, 64a, 643, 640, 627	12	11.1	28

# Extras

---

# Some Commercial Enterprise Options

## ■ Commercial

- Google Search Appliance
- HP's Autonomy
- IBM's Watson Search (formerly Vivisimo)
- Oracle
- Microsoft
- SAP
- Marklogic
- Lexmark
- Attivio

## ■ Open Source Consultants

- Cloudera
- LucidWorks
- Searchblox
- Alfresco (document management+search)

# Some Commercial Desktop Options

---

- **Windows/Mac built in search capability**
- **Copernic**
- **Orion Magic**
- **PowerGrep**
- **dtSearch**
- <http://www.concordancesoftware.co.uk>

# Document Similarity

- Find documents like these
- Pairwise document similarity between all documents in a corpus
- Find similar documents between two corpora
  
- **Similarity metrics**
  - Symmetric vs. asymmetric
  - Overlap, cosine, etc.
- **Granularity**
  - Document
  - Sub-document
  - Sentence

# Corpus Contrasts

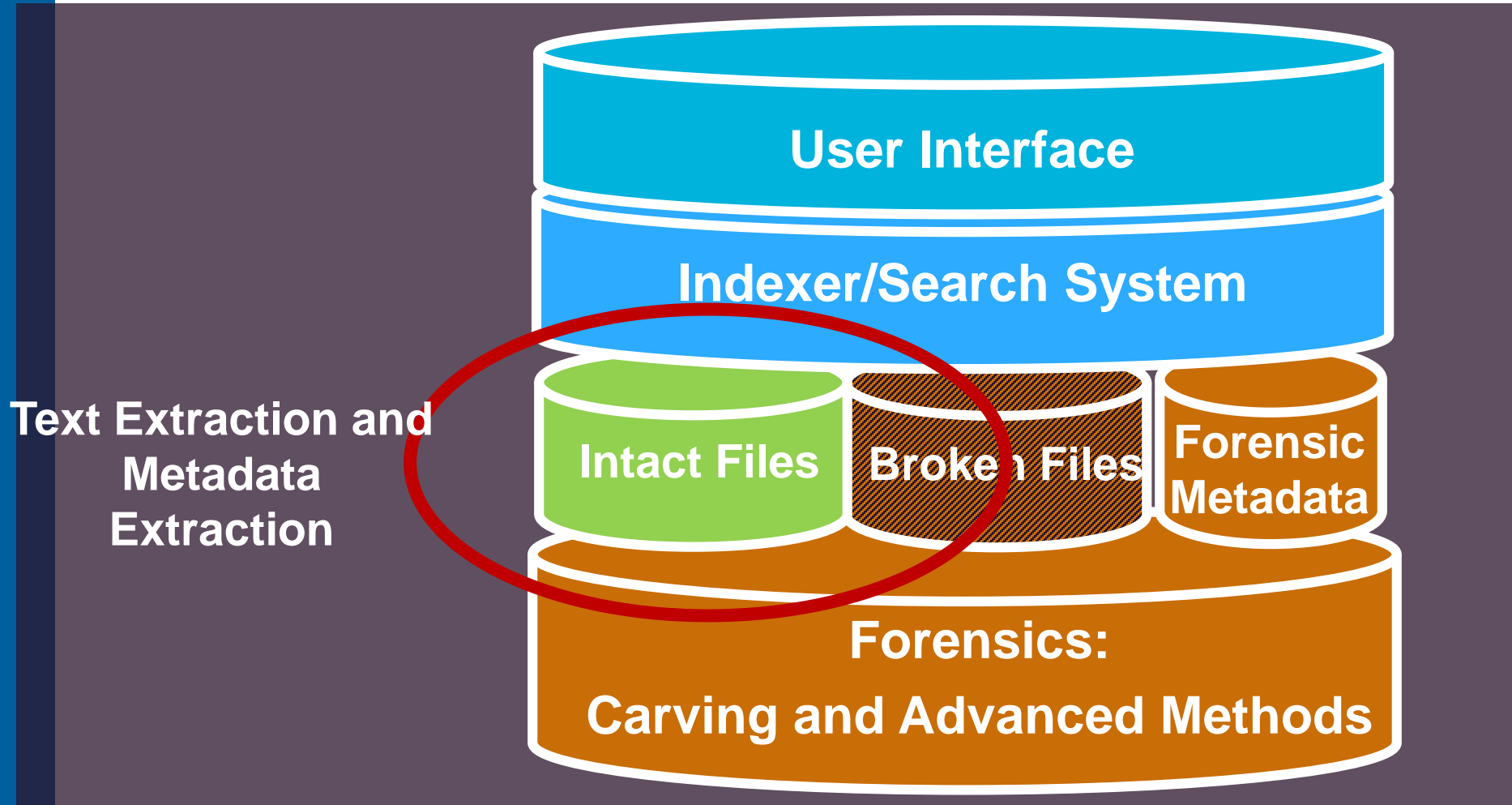
---

- **What terms best distinguish this batch of documents from another?**
- **This month vs. last month.**
- **This region vs. other regions.**
- **This subtopic vs. that subtopic.**

# Chi on “Thai”

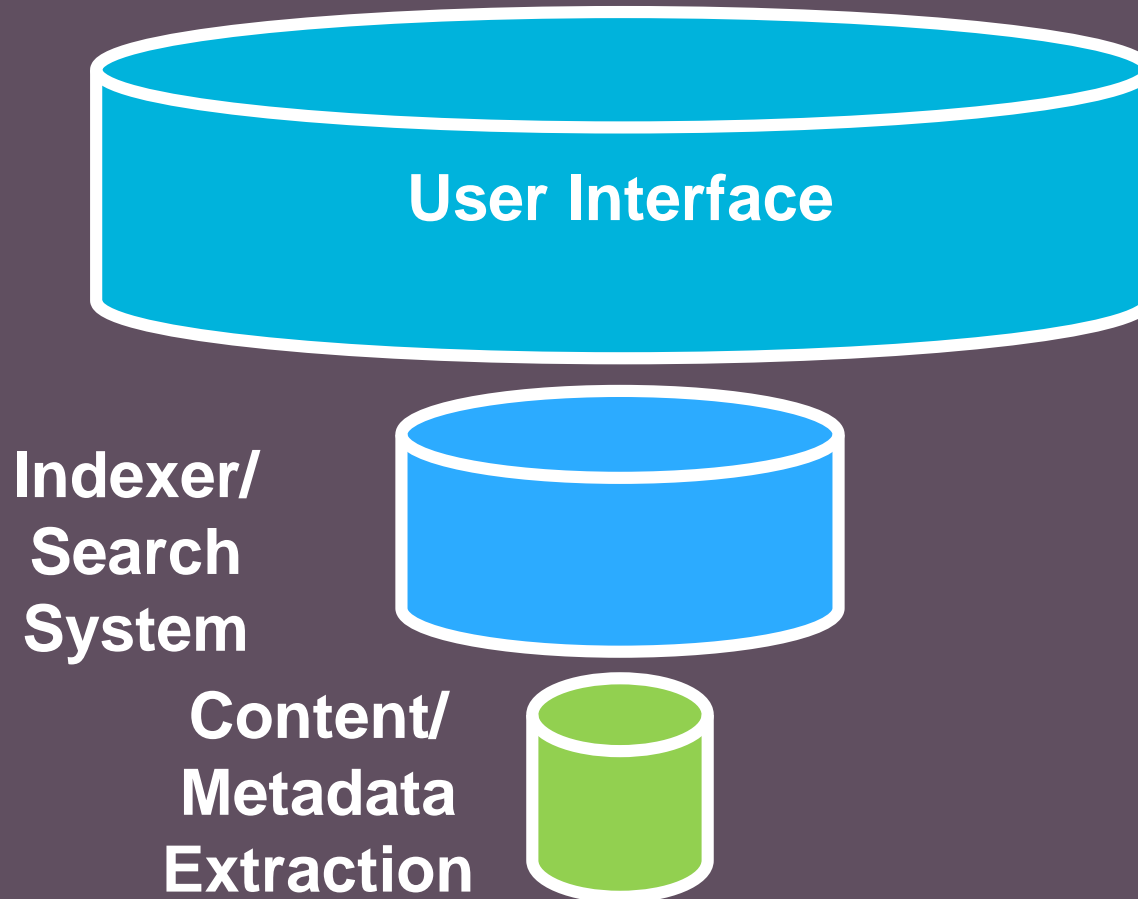
	A	B	C	D	E	F	
1	thai restaurant	10	0	79.2	0.000379	0	in
2	sticky rice	11	3	64.8	0.000417	0.000013	in
3	stir fried	10	2	62.9	0.000379	0.000009	in
4	the thai	8	0	61.5	0.000303	0	in
5	a thai	8	0	61.5	0.000303	0	in
6	green curry	8	1	53	0.000303	0.000004	in
7	pad thai	7	0	52.7	0.000265	0	in
8	thai restaurants	6	0	43.9	0.000227	0	in
9	of thai	6	0	43.9	0.000227	0	in
10	coconut milk	11	8	42.6	0.000417	0.000034	in
11	red curry	6	1	36.1	0.000227	0.000004	in
12	drunken noodles	5	0	35.1	0.000189	0	in
13	thai cooking	5	0	35.1	0.000189	0	in
14	pork or	6	2	30.2	0.000227	0.000009	in
15	rice noodles	6	2	30.2	0.000227	0.000009	in
16	with chilies	5	1	27.8	0.000189	0.000004	in
17	another thai	4	0	26.3	0.000151	0	in
18	thai basil	4	0	26.3	0.000151	0	in
19	ground chicken	4	0	26.3	0.000151	0	in
20	thai cuisine	4	0	26.3	0.000151	0	in
21	noodle dish	4	0	26.3	0.000151	0	in

# High Level Components of a Search Stack





# What the User Sees




# When Things Go Wrong with a Foundation



W. Lloyd MacKenzie, via Flickr  
@ [http://www.flickr.com/photos/saffron\\_blaze/](http://www.flickr.com/photos/saffron_blaze/)

# When Things Go Wrong with Content Extraction

Taking a close look at the forest or open meadows reveals that there are often subtle differences in plant species across a wide landscape. Unique micro-climates, exposure to the sun, soil types, moisture availability, and a variety of other factors influence the types of plant species present in any given location. Changes in any of these factors will cause changes to



BGQOTM G IRUYK RUUQ GZ ZNK LUXKYZ UX UVKT SKGJU]Y  
 XK\KGRY ZNGZ ZNKXK GXK ULZKT Y[HZRK JOLLKXKTIKY OT VRGTZ  
 YVKIOKY GIXUYY G ]OJK RGTJYIGVK% CTOW[K SOIXU-  
 IROSGZKY\$ K^VUY[XK ZU ZNK Y[T\$ YUOR Z\_VKY\$ SUOYZ[XK  
 G\GORGHORZ\_\$ GTJ G \GXOKZ\_ UL UZNKX LGIZUXY OTLR[KTIK ZNK  
 Z\_VKY UL VRGTZ YVKIOKY VXKYKTZ OT GT\_ MO\KT RUIGZOUT%  
 4NGTMKY OT GT\_ UL ZNKYK LGIZUXY ]ORR IG[YK INGTMKY ZU

# When Things Go Wrong with Content Extraction

## Statement

Seasoned professional with a skilled ability to connect co-workers and clients with the information, products and services they are seeking by utilizing professional experiences, organizational and client skills both as a team and an individual.

## Experience

OLS: Office Liquidations Solutions

May 2010 – May 2013

## Statement

**OLS: Office Liquidations Solutions May  
2010 – May 2013**

## Experience

**Bialek Healthcare Environments June 2001  
– May 2010**

Bialek Healthcare Environments

June 2001 – May 2010

Design Associate, Client Services Coordinator

Furniture bid package review, quotation, response and presentation. Small office design, space planning, and assessment, presentation and quotation for commercial projects and for relocation.

# When Things Go Wrong with Text Extraction

---

**You (often) don't  
know what you can't  
find**

# Accessibility and Searchability for Published Documents

## Image:

19 There was documentation of calibration but not of observation of the actual monitoring of the critical limits during production.

## Text Extracted:

19 There was documentation of calibration but not of observation of the actual monitoring of the critical limits during production.

## Search Results:



The screenshot shows a Google search interface. The search bar contains the text "iiionitoring site:www.fsis.usda.gov". Below the search bar, the "Web" tab is selected. The search results show "1 result (0.17 seconds)". The result is a PDF document titled "[PDF] France - 2002 - Food Safety and Inspection Service" with the URL "www.fsis.usda.gov/OPPDE/FAR/France/France2002.pdf". The file format is listed as "PDF/Adobe Acrobat". The snippet of the document text is: "Mar 12, 2003 – 19 There was documentation of calibration but not of observation of the actual iiionitoring of the critical limits during production. 22 Documentation ...".