Processing Capstone Email Using Predictive Coding

Brent M. West
University of Illinois

go.uillinois.edu/capstone

PROJECT OVERVIEW

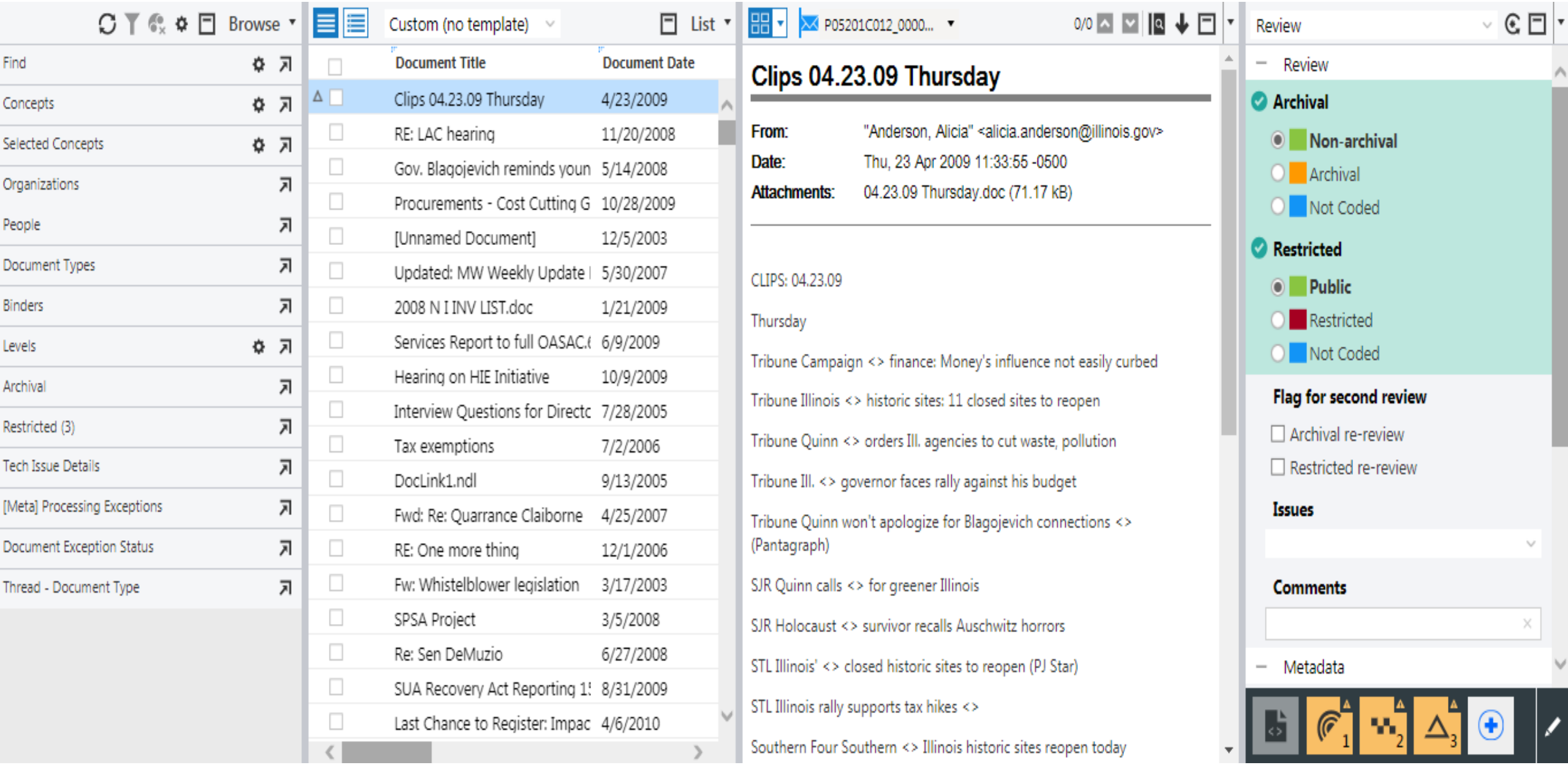
NHPRC-funded 3 year project with Illinois State Archives

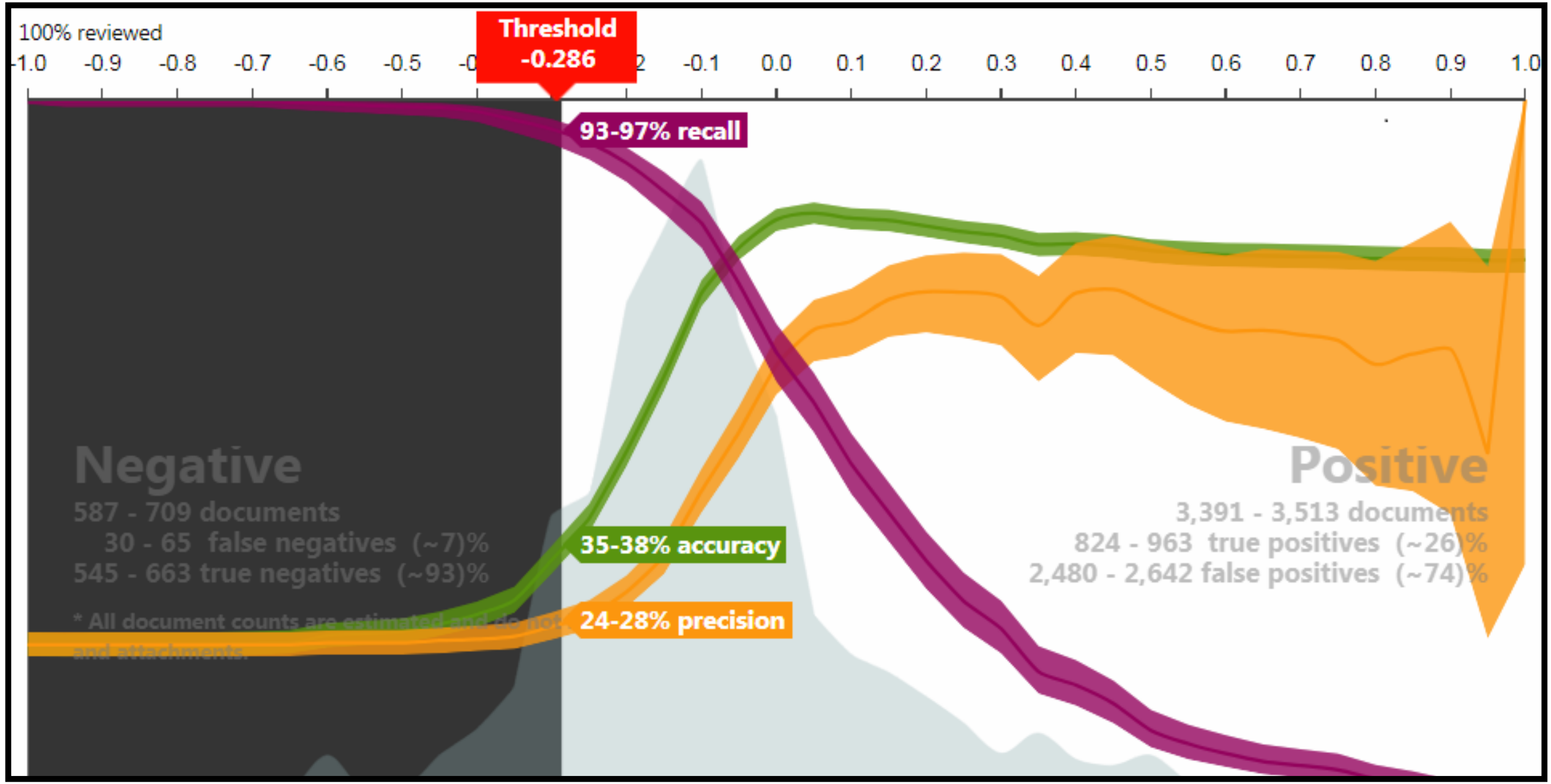
NATIONAL ARCHIVES
NATIONAL HISTORICAL PUBLICATIONS & RECORDS COMMISSION

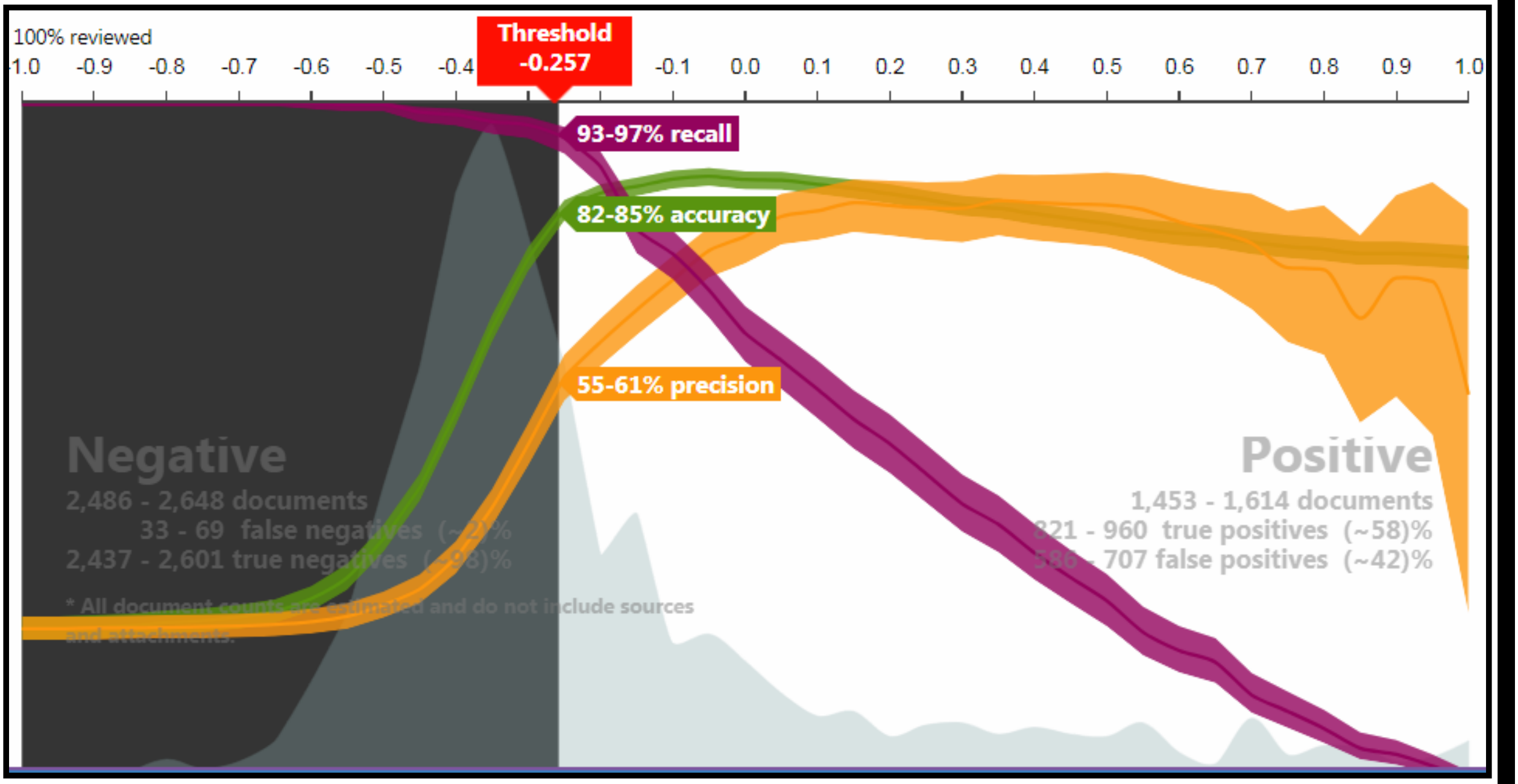
Dataset: 5.3M messages of most senior state officials from 2000-2014

Workflow:

1. File format migration (Emailchemy)
2. Appraise & sanitize (Ringtail)
3. Preservation repository (Preservica)
4. Access (EPADD)


Browse
Find
Concepts
Selected Concepts
Organizations
People
Document Types
Binders
Levels
Archival
Restricted (3)
Tech Issue Details
[Meta] Processing Exceptions
Document Exception Status
Thread - Document Type
Custom (no template)
List
Document Title
Document Date
Clips 04.23.09 Thursday 4/23/2009
RE: LAC hearing 11/20/2008
Gov. Blagojevich reminds youn 5/14/2008
Procurements - Cost Cutting G 10/28/2009
[Unnamed Document] 12/5/2003
Updated: MW Weekly Update I 5/30/2007
2008 N I INV LIST.doc 1/21/2009
Services Report to full OASAC.(6/9/2009
Hearing on HIE Initiative 10/9/2009
Interview Questions for Directc 7/28/2005
Tax exemptions 7/2/2006
DocLink1.ndl 9/13/2005
Fwd: Re: Quarrance Claiborne 4/25/2007
RE: One more thing 12/1/2006
Fw: Whistelblower legislation 3/17/2003
SPSA Project 3/5/2008
Re: Sen DeMuzio 6/27/2008
SUA Recovery Act Reporting 1! 8/31/2009
Last Chance to Register: Impac 4/6/2010
P05201C012_0000...
0/0
Clips 04.23.09 Thursday
From: "Anderson, Alicia" <alicia.anderson@illinois.gov>
Date: Thu, 23 Apr 2009 11:33:55 -0500
Attachments: 04.23.09 Thursday.doc (71.17 kB)
CLIPS: 04.23.09
Thursday
Tribune Campaign <> finance: Money's influence not easily curbed
Tribune Illinois <> historic sites: 11 closed sites to reopen
Tribune Quinn <> orders Ill. agencies to cut waste, pollution
Tribune Ill. <> governor faces rally against his budget
Tribune Quinn won't apologize for Blagojevich connections <> (Pantagraph)
SJR Quinn calls <> for greener Illinois
SJR Holocaust <> survivor recalls Auschwitz horrors
STL Illinois' <> closed historic sites to reopen (PJ Star)
STL Illinois rally supports tax hikes <>
Southern Four Southern <> Illinois historic sites reopen today
Review
Review
Archival
Non-archival
Archival
Not Coded
Restricted
Public
Restricted
Not Coded
Flag for second review
Archival re-review
Restricted re-review
Issues
Comments
Metadata


100% reviewed
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0 2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Threshold
-0.286
93-97% recall
35-38% accuracy
24-28% precision
Negative
587 - 709 documents
30 - 65 false negatives (~7)%
545 - 663 true negatives (~93)%
* All document counts are estimated and do not
and attachments.
Positive
3,391 - 3,513 documents
824 - 963 true positives (~26)%
2,480 - 2,642 false positives (~74)%


100% reviewed
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Threshold
-0.257
93-97% recall
82-85% accuracy
55-61% precision
Negative
2,486 - 2,648 documents
33 - 69 false negatives (~2)%
2,437 - 2,601 true negatives (~98)%
* All document counts are estimated and do not include sources and attachments.
Positive
1,453 - 1,614 documents
821 - 960 true positives (~58)%
586 - 707 false positives (~42)%

LESSONS LEARNED

1. Augmented not automated appraisal
2. Huge labor savings (99.4%)
3. Capstone ≠ all archival (41%)
4. PII small but important (<2%)
5. Quantify and reduce risk
6. “Block box” commercial solutions
7. Forgiving to mistakes but expertise required
8. Trade-offs between privacy and transparency
9. Proceed deliberately: document decisions and context
10. Integrate into a lifecycle workflow

- Pre Processing
  - Creating RDX Guid Tables
  - Batching
  - Pre Processing
- Processing
  - Processing
  - Post Processing
  - Standardizing Data
- Import Files and Metadata
  - Load Data
  - Cleanup
- Hashes
  - Hashes
- Group Coding
  - Group Coding
- Data Filtering
  - Data Filtering
- Transfer Unsuppressed Files
  - File Copy Batching
  - File Transfer
  - File Transfer Confirmation
- OCR
  - OCR
- Indexing
- Update Field Counts
  - Update Field Counts
- Transfer Suppressed Files
  - File Copy Batching
  - File Transfer
  - File Transfer Confirmation
- Gathering Report Data
  - Gathering Report Data
- Submit Index and Enrichment Job
  - Submit Index and Enrichment Job
- Submit Ingestions Cleanup Job
  - Submit Ingestions Cleanup Job
- Finalize Job
  - Cleanup

FY24 Ringtail Project_converted.7z

Description	Started via upload
Job ID	2
Started	11/9/2023 9:54 AM
Duration	**1h 39m 52s**
Status	Completed with exceptions

Processed files: 601,125 (94.91GB)

Total expanded documents (77.14GB)	714,252	
Suppressed documents		
Duplicates	19,513 Master Document	2.7%
Outside date range	Not applicable	
Excluded NIST files	23	0.0%
Excluded by category or extension	0	0
Does not match search term family	Not applicable	
Unsuppressed documents (76.07GB)	**694,716**	**97.3%**

Suppressed document ingestion exceptions	116
Renamed Extension	33
NIST Item	27
Non-Searchable PDF	23
Text Stripped	16
Multimedia	7
Unknown Binary	6
Encrypted	2
Corrupted	1
Export Failed	1

Unsuppressed document ingestion exceptions	19,462
Renamed Extension	4,618
Non-Searchable PDF	4,614
Text Stripped	4,304
Corrupted	2,004
Unknown Binary	1,343
System File	860
Multimedia	642
Encrypted	248
NIST Item	241
Data Type Conversion Failed	228
Field Data Truncated	124
Missing Hash Value	79
Empty	49
Extracted Text Only	30
Export Failed	27
Field Data Extraction Error	25
Inaccessible Content	23
Databases	3

Included fields

BCC
CC
Collection ID
Custodian
Custodian ID
Document Date
Document Type
Evidence Job ID
From
Media ID
To
Appmt Location
Appmt Optional Attendees
Appmt Organizer
Appmt Required Attendees
Chat Action Count
Chat Attachment Count
Chat Disclaimer Count
Chat End Date
Chat End Time
Chat Enter Count
Chat Event Count
Chat Exit Count
Chat Invite Count
Chat Message Count
Chat Participant Count
Chat Sender Count
Chat Start Date
Chat Start Time
Chat Viewed History Count
Conversation Topic
Date Accessed
Date Accessed Time
Date Appmt End
Date Appmt End Time
Date Appmt Start
Date Appmt Start Time
Date Created
Date Created Time
Date Modified
Date Modified Time
Date Received
Date Received Time
Date Sent
Date Sent Time
Date Top Family
Date Top Family Time
Document Category
Document Kind
Email Delivery Receipt Request
Email Folder
Email Importance
Email Message ID
Email Message ID Replied To
Email Read Receipt Request
Email Sensitivity
Entity
EntryID
Evidence ID
Excluded File
Exif Date Time
Exif Date Time Digitized
Exif Date Time Original
Extended File Path
File Application
File Extension - Loaded
File Extension - Original
File Name
File Path
File Size
GUID
GUID - Parent
ISO Media Type
Languages
Latitude
Longitude
Mapi-DeliveredTo
Mapi-Message-Flags
Message Class
Multimedia Audio Codec
Multimedia Duration
Multimedia Video Codec
NSF UNID
Office Exceptions
Office Exceptions - Excel Hidden Sheet Count
Office Exceptions - Notes Count
Office Property - Author
Office Property - Comments
Office Property - Company
Office Property - Date Last Printed
Office Property - Date Last Printed Time
Office Property - Date Last Saved
Office Property - Keyword
Office Property - Last Author Saved By
Office Property - Title
Original Full Path
PDF - Encryption Level
PDF - Portfolio
PDF Properties
Processing Exceptions
Processing Time Zone
Revision Number
Subject
Transport Message Headers
Word-Pdf-Image Page Count
DPM File ID
Ingestion Complete
Ingestion Exception Detail
MD5 Hash

Nuix Discover | P05201C021_UOFI Capstone Email | Case Home | Case Manager | BMWest | Help

Case Home

Case Home

- Case Site Map
- My Assignments
- Review Dashboard
- Concept Cloud
- Case Dashboard

Document set

All documents in the case

Hide the top 0 concepts

40481 documents containing News

Version 10.17.004.2

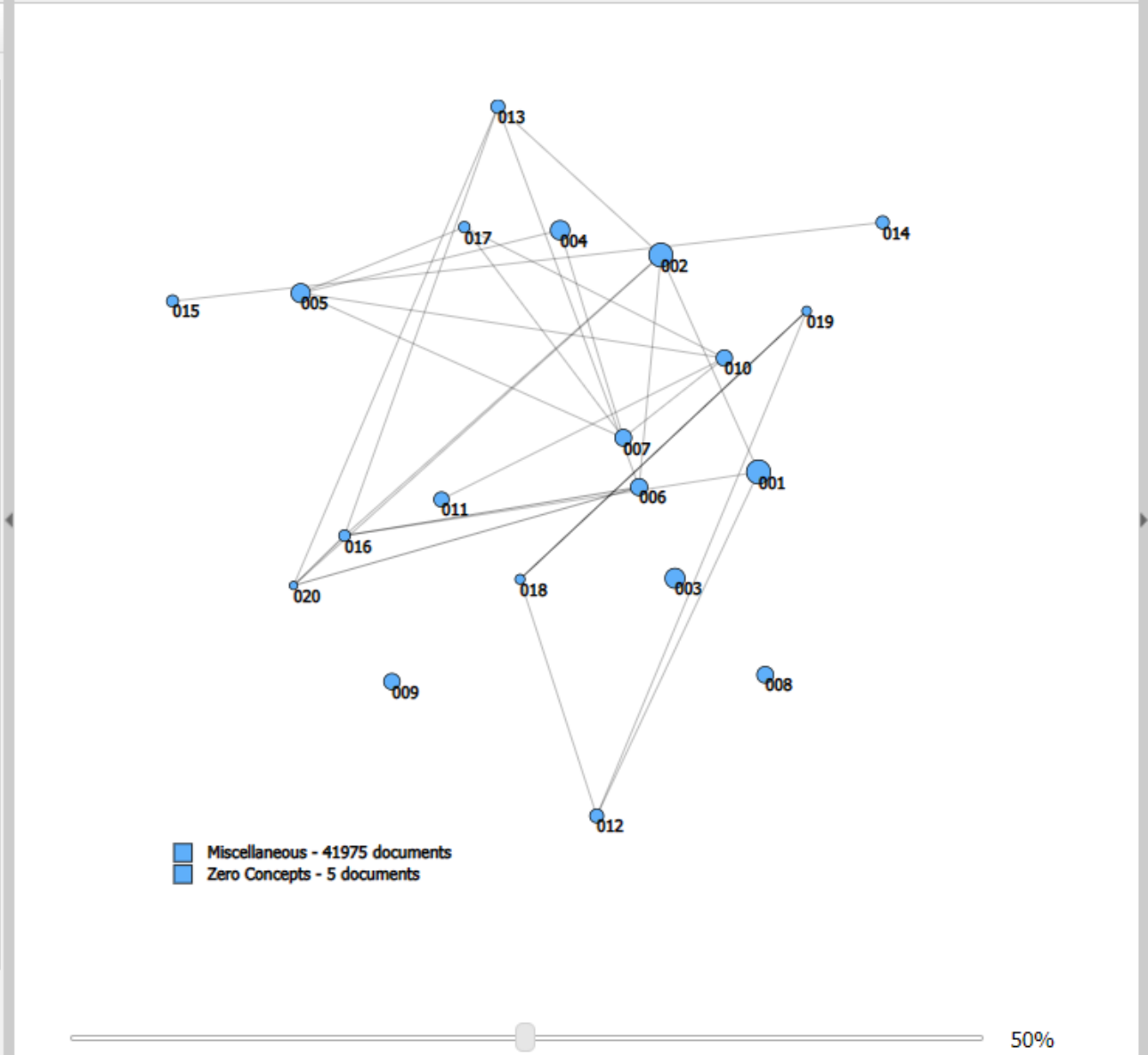
Rows: All × Date ×

	Document Kind	Document Type	Org	People	Relevance: Archival
1996	45	45	22	22	0
1997	22	22	18	18	0
1998	36	36	22	23	0
1999	423	423	357	357	0
2000	7523	7523	6116	6120	0
2001	12317	12317	9169	9176	5
2002	18709	18709	12718	12798	0
2003	22988	22988	17014	17087	0
2004	734	734	570	571	0
2005	358	358	325	325	0
2006	603	603	503	503	0
2007	759	759	664	664	0
2008	1241	1241	1114	1114	0
2009	1432	1432	1292	1292	0
2010	7761	7761	6791	6795	0
2011	9497	9497	7835	7843	0
2012	10055	10055	8392	8394	0
2013	13786	13786	11094	11099	0
2014	19082	19082	14322	14338	0
2015	19391	19391	15260	15261	0
2016	16442	16442	13446	13453	0
2017	16564	16564	13474	13483	0
2018	17782	17782	14358	14363	0
2019	20687	20687	16725	16727	0
2020	204904	204904	178341	178389	0
2021	292629	292629	244332	244397	0
2022	38397	38397	32322	32334	0
2023	2584	2584	2188	2196	0

Cluster List

ID	Name	Total
001	Fenner, Melvin, MEL	45663
002	Shield, Julie, Shield Illinois	45406
003	student, community, Education	37940
004	archive, Library, Ellen	36877
005	Library, Urbana, Archivist	36326
006	Watkins, Ronald, Ron	32510
007	Library, archive, Urbana	32039
008	New, Twitter, Facebook	31825
009	Brewer, Michael, Mike	31022
010	Library, Copyright, archive	30919
011	Senate, Committee, faculty	28857
012	file, Medicat, NAME	26385
013	Yoni, Merid, lab	26235
014	Unsubscribe, privacy, Data	24872
015	SUBSCRIPTION, Service, release	22646
016	icon, vector graphics, white cross Description	21773
017	Archivist, Library, ICA	21284
018	Redwood City, Jefferson Avenue, Redwood	18748
019	Box Health, Redwood City, Redwood	18116
020	Greta, Andrew, Shield	14458
ZERO	No concepts	5

Mine Map



Concepts

Name	Density	Total
Illinois		326037
University		317442
information		235350
business		178976
Service		172326
Director		159205
office		158612
SYSTEM		152808
Urbana-Champaign		152793
Urbana		149917
work		149743
question		146692
Library		142040
Shield		141326

Confidence

Confidence	Count	Density
90 - 100%	4	
80 - 90%	1934	
70 - 80%	11854	
60 - 70%	57942	
50 - 60%	125747	
40 - 50%	212786	

RE: Hoke out at Z bldg today or tomorrow.
From:
"Brewer, Michael K (O&M)" <mkbrewer@university-of-illinois>
To:
"Madsen, Kenneth M (O&M)" <kmmadsen@oandm.uiuc.edu>
Date:
Mon, .. Feb .... ..:..:06 +0000
I've got a message in to ..... to find out what the status is.
We asked them to deliver .-. of the .... a long time ago. If they can not do that now, we will buy some more on Keith's money to get some in stock.
Thanks for ..... me.
-Mike
-----Original Message-----
From: Madsen, Kenneth M (O&M)
Sent: Monday, February .., ...:.. AM
To: Brewer, Michael K (O&M)
Subject: RE: Hoke out at Z bldg today or tomorrow.
Mike
Are we going to get ." pipe from Hoke?
-----Original Message-----
From: Brewer, Michael K (O&M)
Sent: Monday, February .., ...:.. AM
To: Roberts, Joseph P (O&M); Madsen, Kenneth M (O&M)
Cc: Larson, Michael J (O&M); Erickson, Keith R (O&M)
Subject: Hoke out at Z bldg today or tomorrow.
Hoke called and will send someone up to stop the leak at the strainer today or tomorrow. They may have to shut the station down to replace the gasket.
When they do, I'll call so we can look to agree that the leak is stopped and let Mr. ..... know he will need to do .....
Thank you.
Mike Brewer, PE, Utilities, O&M
Phone (...) ...-7067