Text Analytics Workshop

Tom Reamy
Chief Knowledge Architect
KAPS Group
http://www.kapsgroup.com
Author: Deep Text

KAPS Group
Knowledge Architecture Professional Services

Agenda

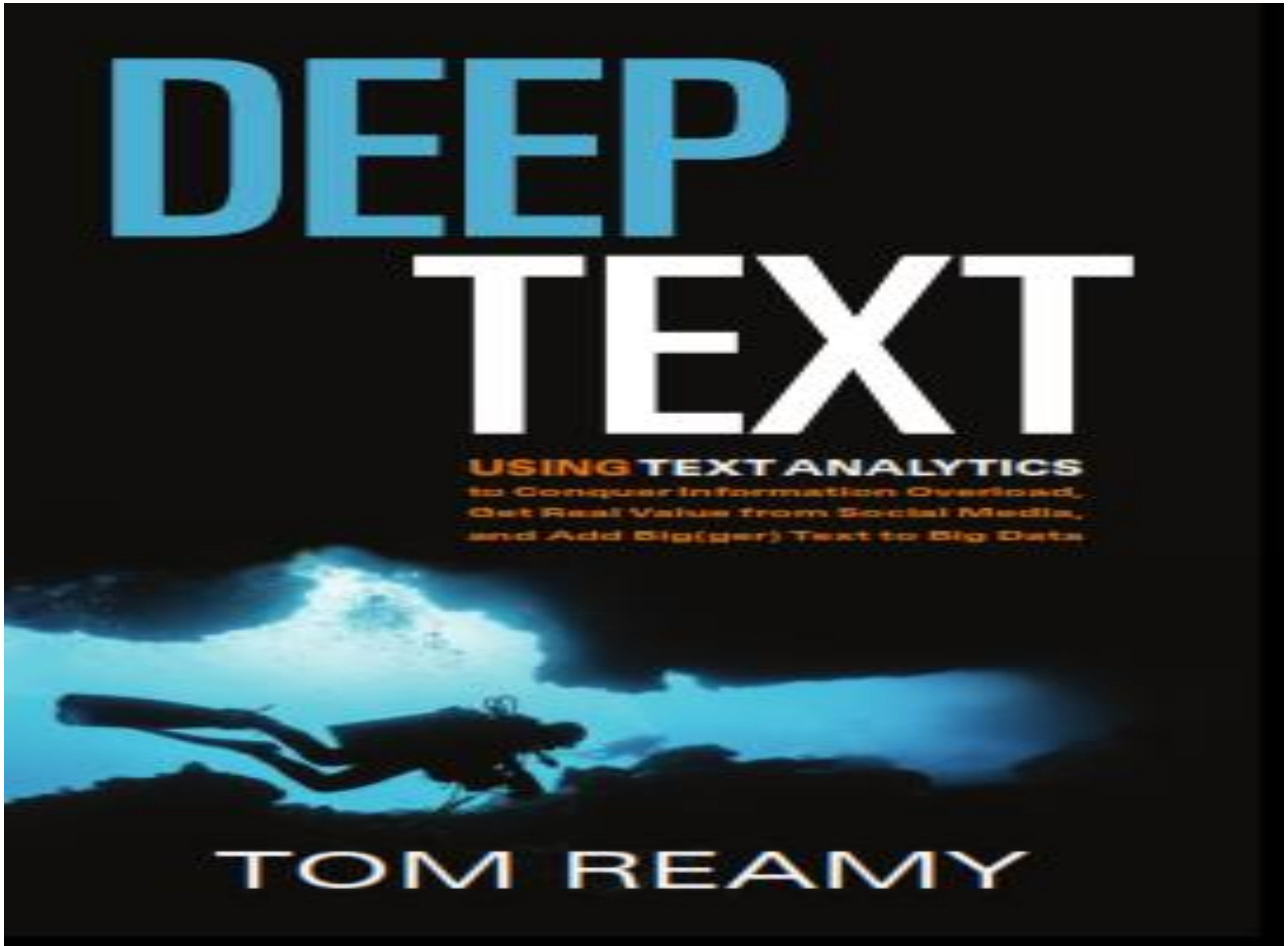
- Introduction – Elements of Text Analytics
- Development –Categorization, Data Extraction
- Text Analytics Applications
- AI and Text Analytics
  - Building a Text Analytics and Gen AI Foundation
- Questions / Discussions

Introduction: KAPS Group

- Network of Consultants and Partners - 2002
- Text analytics consulting: Strategy, Development-taxonomy, text analytics foundation & applications, Prompt Engineering
- Mini-Projects – get started or take to next level
  - Strategy-TA & Gen AI, Mini-POC - Categorization
- Partners – Expert AI, Synaptica, SAS, Smartlogic, Lexalytics, BA Insight, BiText
- Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, IMF, IFC, Dept. of Transportation, RWJF, IDG/Foundry, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Program Chair – Text Analytics Forum – Nov. 20-21

A treasure trove of technical detail, likely to become a definitive source on text analytics – *Kirkus Reviews*

Book signing



Text Analytics Forum (TAF) Introduction
Elements of Text Analytics

- Auto-categorization - Statistical – Rules
  - Brains of the outfit: (Rules) Makes everything else smarter
  - Sentiment Analysis – positive & negative, attitudes
    - Advanced - racial equality, social media analysis
- Data Extraction – entities, concepts, events, facts
  - Analytical applications, enhance search - facets
- Supplemental – NLP, summarization, variety of analytical
  - NLG – Language generation
  - Text Mining – NLP, feed analytical apps, terms for categorization

File Edit View Model Help

Phrase	Covera... ▼	Docume...	Co...
health	100%	20	1121
care	100%	20	491
also	95%	19	160
use	90%	18	134
state	90%	18	660
year	90%	18	311
new	90%	18	192
health care	90%	18	226
information	90%	18	186
including	90%	18	232
one	90%	18	134
services	90%	18	224
plan	90%	18	276
data	85%	17	163
coverage	85%	17	646
time	85%	17	69
federal	85%	17	154
changes	85%	17	103
public	85%	17	205
costs	85%	17	165
higher	85%	17	167
made	85%	17	85
found	85%	17	96
individuals	85%	17	93
insurance	85%	17	336
provided	85%	17	114
reports	85%	17	168
number	80%	16	141
source	80%	16	187
rates	80%	16	177

Source document: **C:\Text Files\RWJF Content Repository 2021\Initial Rule Development Proj Coverage Access\First 20\195706.pdf**

(...) JD

Research Analyst

Lisa Cacari Stone. MS

Research Analyst

Robert W. Seifert, MPA

Senior Policy Analyst The Access Project

Heller Graduate School, Brandeis University

Prepared for United Power for Action and Justice, March 2000

The Access Project is a national initiative of The Robert Wood Johnson Foundation, in partnership with Brandeis University's H Collaborative for Community Health Development. It began its efforts in early 1998. The mission of The Access Project is to im by assisting local communities in developing and sustaining efforts that improve healthcare access and promote universal cov who are without health insurance.

Ifyou have any additional questions, or would Uke to learn more about our work, please contact us.

The Access Project

30 Winter Street, Suite 930 Boston, MA 02108 Phone; 617-654-9911 Fax: 617-654-9922

E-mail: info@accessproject.org Web site: www.accessproject.org

United Povi^er for Action and Justice is an organization of 330 dues-paying member congregations, community organizatio community health centers in Chicago and its suburbs. It is committed to citizen-initiated democracy and action for justice on metropolitan Chicago. United Power's Gilead Campaign for the Uninsured is an initiative to build political will and find practic of the uninsured. United Power advocates the funding of a comprehensive enrollment campaign to register those eligible but benefits programs; is asking the Cook County Board to include $20 miUion in next year's budget to create a pilot program to e County; and has begun to advocate for the use of tobacco settiement moneys to be used to create expanded health care cove

United Power for Action and Justice Phone:773-334-7281

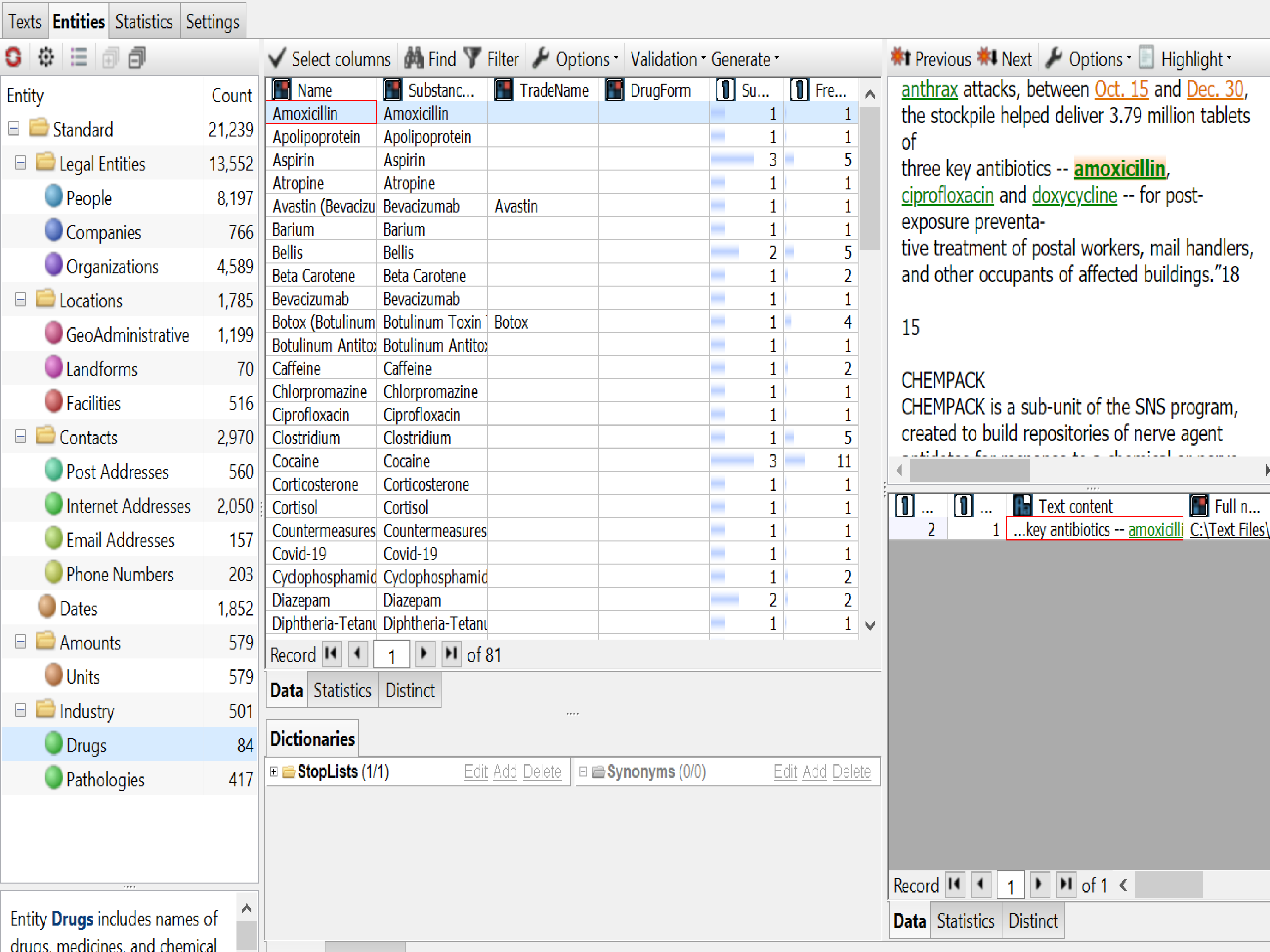
The authors would Uke to thank Kara Sokol of United Power for Action and Justice for her contributions to this report.

This Report may be reproduced or quoted with appropriate credit.

INTRODUCTION

This report describes the growing number of people without health insurance in Illinois. It discusses trends within the state ar growing sub-groups among the uninsured. The rising number of the uninsured in IUinois harms overaU health and weU-being begun or expanded innovative programs to cover the uninsured, which might serve as models for Illinois. These programs can uninsurance rate, thereby improving the overaU health of the state's residents. The national tobacco settiement is an importar coverage expansion, and several states have proposed devoting sizable portions of their settiement doUars to aUeviating the s growing ranks of uninsured.

Previous document Next document Show all



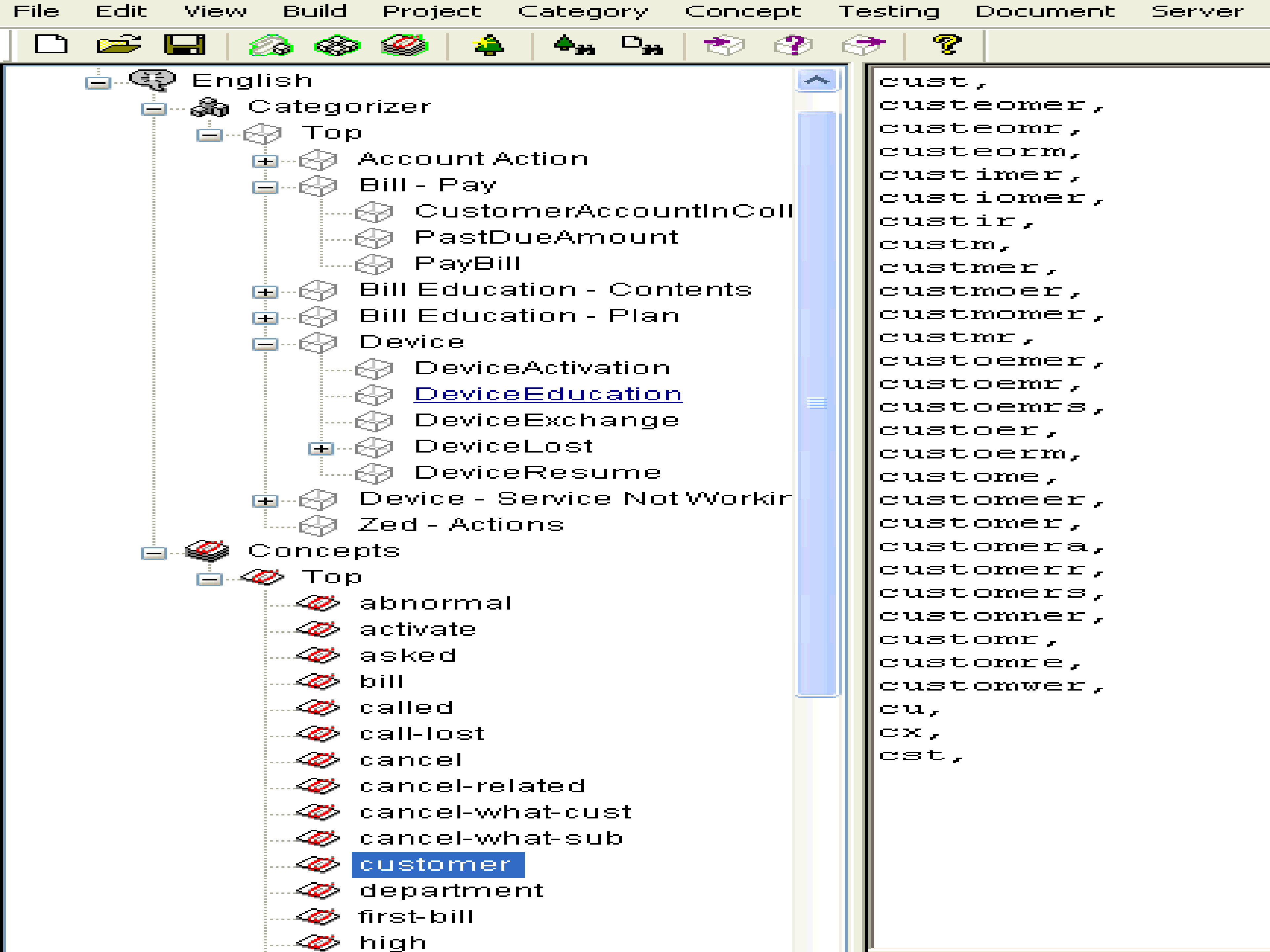
Texts
Entities
Statistics
Settings
Select columns
Find
Filter
Options
Validation
Generate
Previous
Next
Options
Highlight
Entity
Count
Standard 21,239
Legal Entities 13,552
People 8,197
Companies 766
Organizations 4,589
Locations 1,785
GeoAdministrative 1,199
Landforms 70
Facilities 516
Contacts 2,970
Post Addresses 560
Internet Addresses 2,050
Email Addresses 157
Phone Numbers 203
Dates 1,852
Amounts 579
Units 579
Industry 501
Drugs 84
Pathologies 417
Entity Drugs includes names of drugs, medicines, and chemical
Name
Substanc...
TradeName
DrugForm
Su...
Fre...
Amoxicillin Amoxicillin 1 1
Apolipoprotein Apolipoprotein 1 1
Aspirin Aspirin 3 5
Atropine Atropine 1 1
Avastin (Bevacizu Bevacizumab Avastin 1 1
Barium Barium 1 1
Bellis Bellis 2 5
Beta Carotene Beta Carotene 1 2
Bevacizumab Bevacizumab 1 1
Botox (Botulinum Botulinum Toxin Botox 1 4
Botulinum Antito: Botulinum Antito: 1 1
Caffeine Caffeine 1 2
Chlorpromazine Chlorpromazine 1 1
Ciprofloxacin Ciprofloxacin 1 1
Clostridium Clostridium 1 5
Cocaine Cocaine 3 11
Corticosterone Corticosterone 1 1
Cortisol Cortisol 1 1
Countermeasures Countermeasures 1 1
Covid-19 Covid-19 1 1
Cyclophosphamid Cyclophosphamid 1 2
Diazepam Diazepam 2 2
Diphtheria-Tetanu Diphtheria-Tetanu 1 1
Record 1 of 81
Data
Statistics
Distinct
Dictionaries
StopLists (1/1) Edit Add Delete
Synonyms (0/0) Edit Add Delete
anthrax attacks, between Oct. 15 and Dec. 30,
the stockpile helped deliver 3.79 million tablets of
three key antibiotics -- amoxicillin,
ciprofloxacin and doxycycline -- for post-exposure preventa-
tive treatment of postal workers, mail handlers,
and other occupants of affected buildings."18
15
CHEMPACK
CHEMPACK is a sub-unit of the SNS program,
created to build repositories of nerve agent
...
...
Text content
Full n...
2 1 ...key antibiotics -- amoxicilli C:\Text Files\
Record 1 of 1
Data
Statistics
Distinct

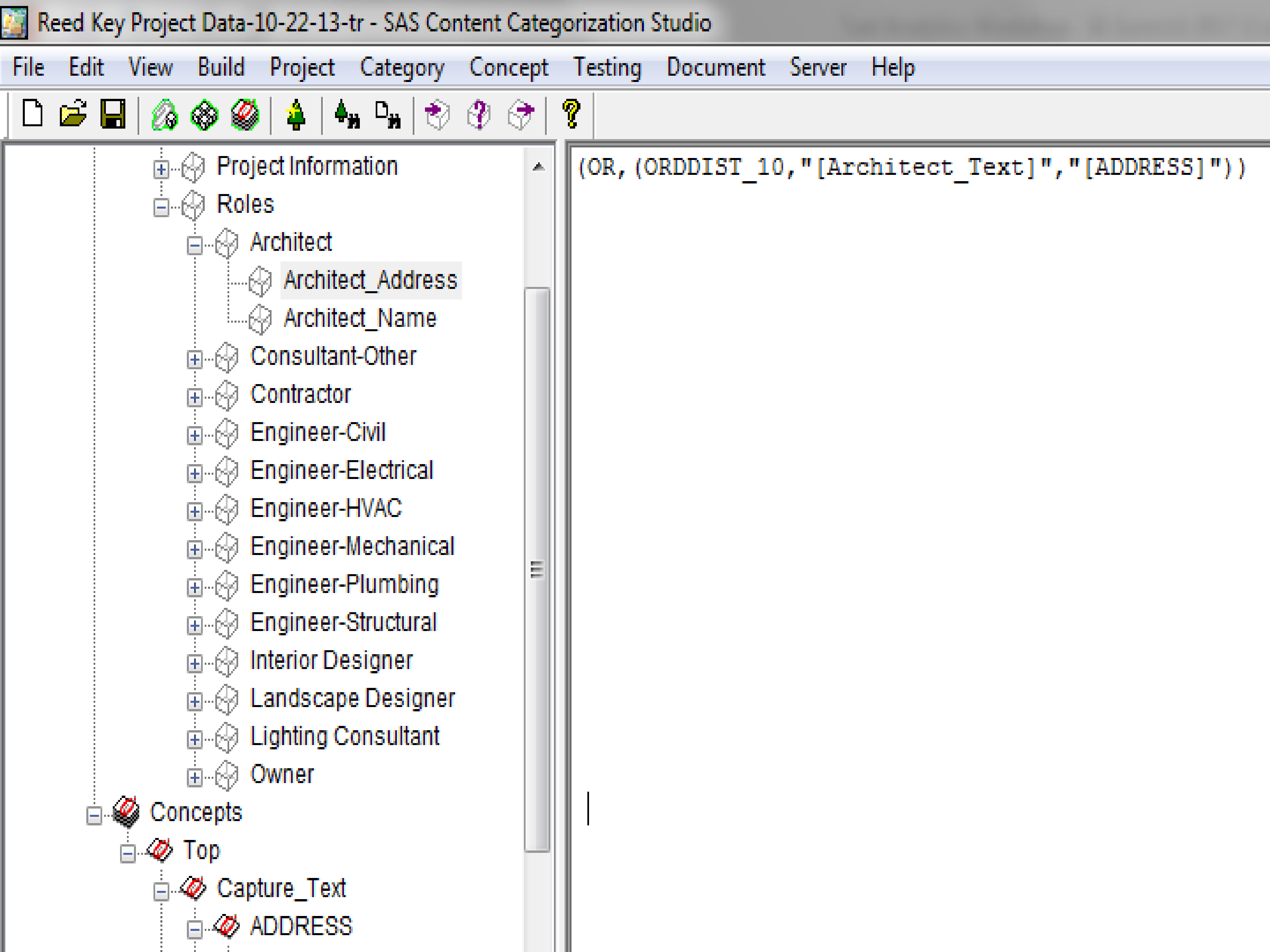
	A	B	C	D	E	F	G
1	Term (1-gram)	Term Phrase (2-gram to n-gram)	Term N	Doc N	Doc %	tf-idf	Sentence
2	agile		258	20	100%	-0.0209	
3		agile teams	29	8	40%	-0.001	Finally, agile development leads to better software because people on agile tea
4		agile development	27	13	65%	-0.0016	Agile development teams and organizations that follow devops should review a
5		agile principles	14	7	35%	-0.0004	Others define agile principles and governance models so that agile product owr
6		agile methodology	13	5	25%	-0.0001	agile methodology establishes both a mindset and process for that continuous
7		scaled agile	10	3	15%	0.0001	Large organizations adopting the Scaled Agile Framework (SAFe) use Program In and understand team dependencies.
8		agile software	9	3	15%	0.0001	But one of the really cool and powerful aspects of Git is that you can use it to w parallel, which is crucial for agile software development.
9		agile development teams	9	9	45%	-0.0004	Agile development teams and organizations that follow devops should review a
10		agile product	8	5	25%	-0.0001	Others define agile principles and governance models so that agile product owr
11		safe agile	7	1	5%	0.0005	The 45-question exam covers candidates' ability to: Explain SAFe agile principle iterations and drive value Improve ART processes Work with other teams on A Candidates must be familiar with Scrum, Kanban, and Extreme Programming (X working knowledge of software or hardware development processes.
12		agile methodologies	6	5	25%	-0.0001	But what is agile, and how do developers and organizations incorporate agile m
13		disciplined agile	6	1	5%	0.0004	Once you've earned your DASM certification, it also opens you up to take the D Certification or Disciplined Agile Value Stream Consultant (DAVSC) Certification
14		agile planning	6	2	10%	0.0002	For distributed software development teams, I advise formalizing agile plannin customer expectations.

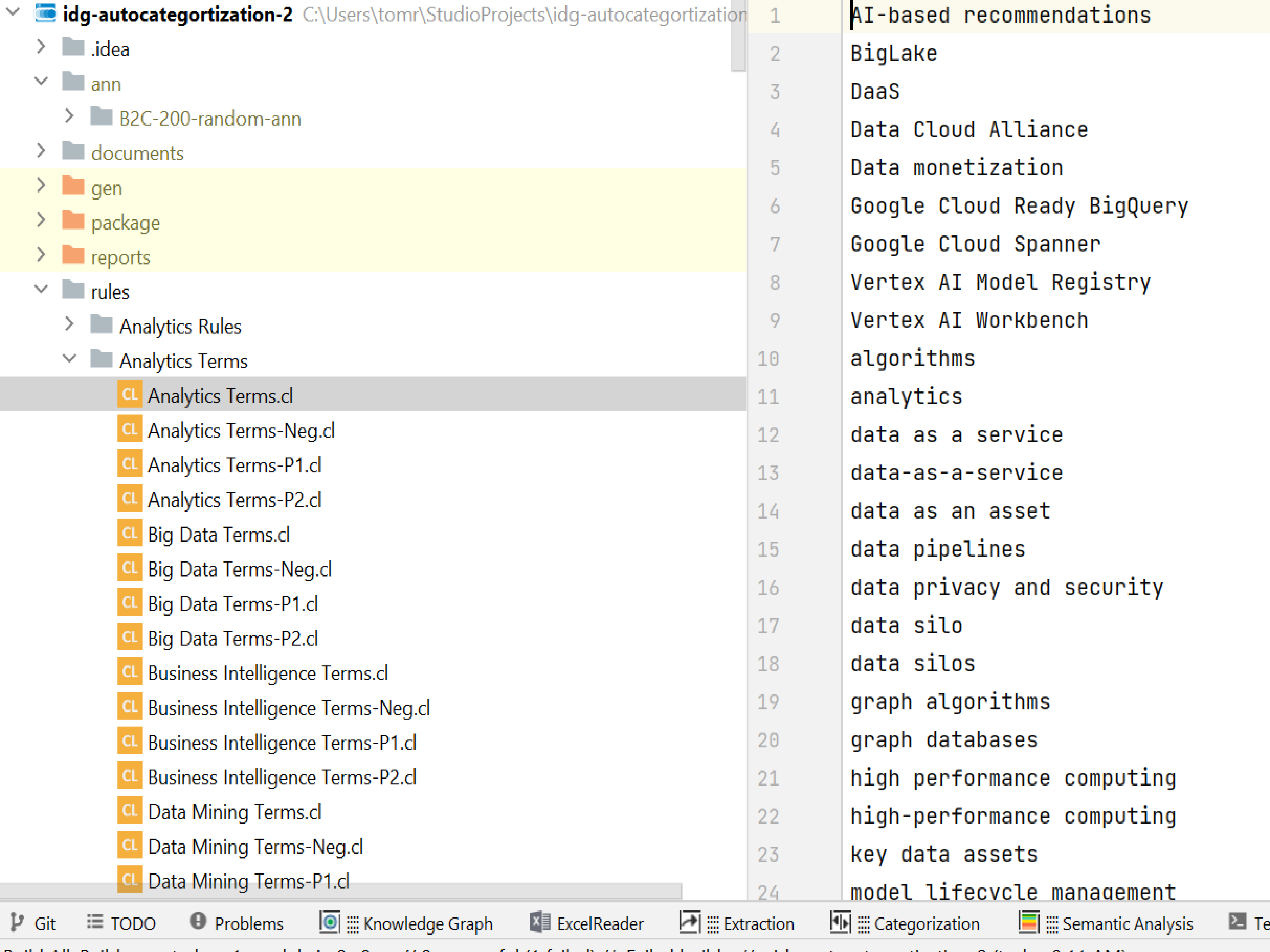
terms | phrases | term combos | keywords | keyword combos | h2 & h3 tags | headlines | stop v ...

- Reed Key Project Data
  - English
    - Categorizer
      - Top
        - Dates-Times
        - Project Attributes
        - Project Information
        - Roles
    - Concepts
      - Top
        - Capture_Text
          - ADDRESS
          - Bond_Amount
          - COMPANY
          - DATE
          - IFB_NUMBER
          - LOCATION
          - MONEY
          - ORGANIZATION
          - PERSON
          - Project_Text
          - TIME
        - Dates-Times-Text
        - Project_Attributes_Text
        - Project_Information_Text
        - Roles_Text
        - Support_Concepts

```
# ROOT = *PERSON

*PERSON = !INITCAP !INITCAP says
*PERSON = !INITCAP !INITCAP said
```


File Edit View Build Project Category Concept Testing Document Server
English
Categorizer
Top
Account Action
Bill - Pay
CustomerAccountInColl
PastDueAmount
PayBill
Bill Education - Contents
Bill Education - Plan
Device
DeviceActivation
DeviceEducation
DeviceExchange
DeviceLost
DeviceResume
Device - Service Not Workir
Zed - Actions
Concepts
Top
abnormal
activate
asked
bill
called
call-lost
cancel
cancel-related
cancel-what-cust
cancel-what-sub
customer
department
first-bill
high
cust,
custeomer,
custeomr,
custeorm,
custimer,
custiomer,
custir,
custm,
custmer,
custmoer,
custmomer,
custmr,
custoemer,
custoemr,
custoemrs,
custoer,
custoerm,
custome,
customeer,
customer,
customera,
customerr,
customers,
customner,
customr,
customre,
customwer,
cu,
cx,
cst,


Reed Key Project Data-10-22-13-tr - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Project Information
Roles
Architect
Architect_Address
Architect_Name
Consultant-Other
Contractor
Engineer-Civil
Engineer-Electrical
Engineer-HVAC
Engineer-Mechanical
Engineer-Plumbing
Engineer-Structural
Interior Designer
Landscape Designer
Lighting Consultant
Owner
Concepts
Top
Capture_Text
ADDRESS
(OR,(ORDDIST_10,"[Architect_Text]","[ADDRESS]"))


idg-autocategortization-2 C:\Users\tomr\StudioProjects\idg-autocategortization
.idea
ann
B2C-200-random-ann
documents
gen
package
reports
rules
Analytics Rules
Analytics Terms
Analytics Terms.cl
Analytics Terms-Neg.cl
Analytics Terms-P1.cl
Analytics Terms-P2.cl
Big Data Terms.cl
Big Data Terms-Neg.cl
Big Data Terms-P1.cl
Big Data Terms-P2.cl
Business Intelligence Terms.cl
Business Intelligence Terms-Neg.cl
Business Intelligence Terms-P1.cl
Business Intelligence Terms-P2.cl
Data Mining Terms.cl
Data Mining Terms-Neg.cl
Data Mining Terms-P1.cl
AI-based recommendations
BigLake
DaaS
Data Cloud Alliance
Data monetization
Google Cloud Ready BigQuery
Google Cloud Spanner
Vertex AI Model Registry
Vertex AI Workbench
algorithms
analytics
data as a service
data-as-a-service
data as an asset
data pipelines
data privacy and security
data silo
data silos
graph algorithms
graph databases
high performance computing
high-performance computing
key data assets
model lifecycle management
Git
TODO
Problems
Knowledge Graph
ExcelReader
Extraction
Categorization
Semantic Analysis

Expression

Functions Columns M Check syntax ?

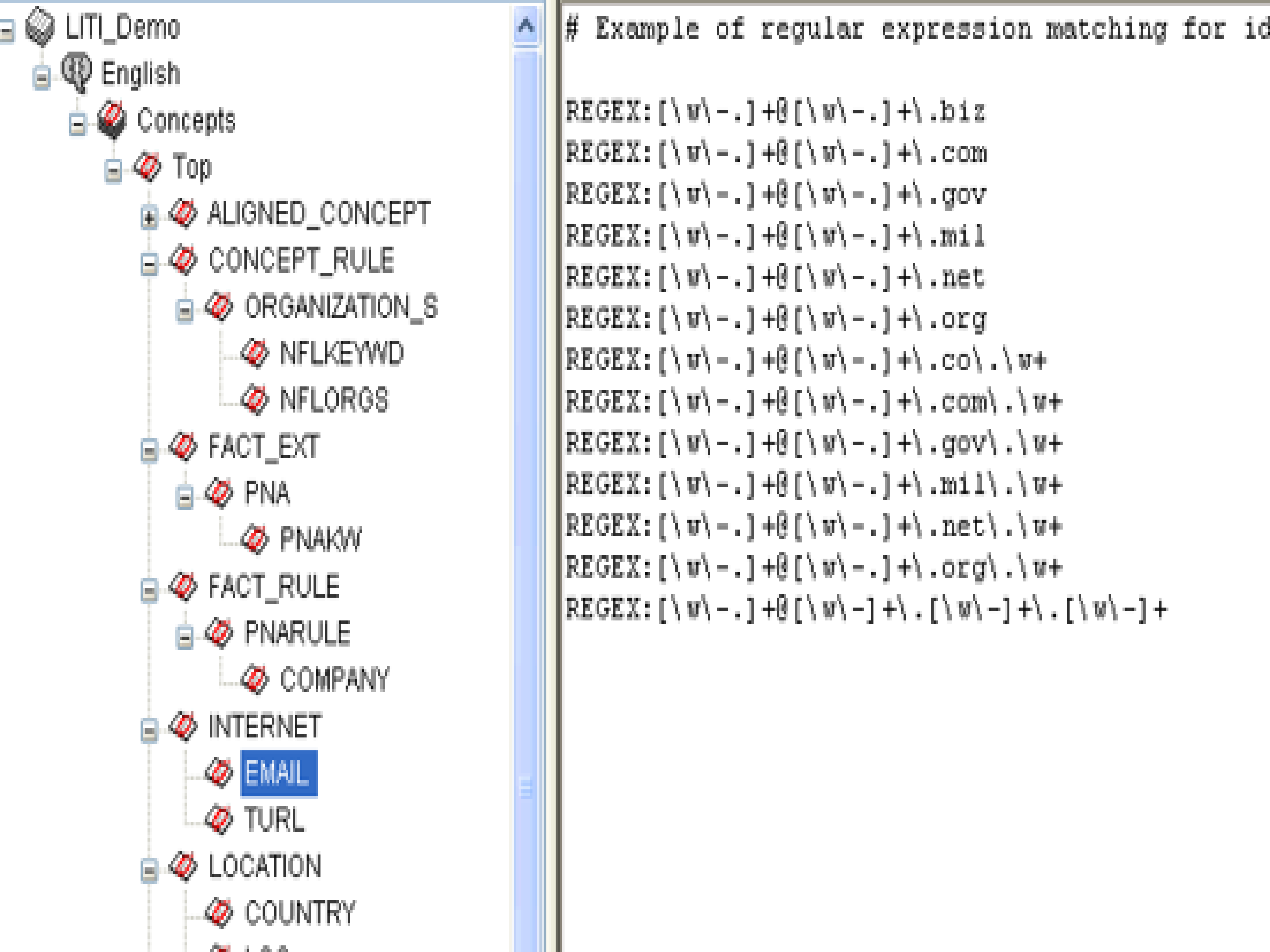
```
(position(1000,
    (position (100,
        (term(Childhood Obesity Terms)
        / sentence(term(Childhood Obesity Terms),term(Childhood Obesity Negative Terms)))
        or
        near(7,term(Childhood Terms), term(Obesity Terms))
        / sentence(near(7,term(Childhood Terms), term(Obesity Terms)),term(Childhood Obesity Ne
    ))
    or
        (fnear(200, term(Summary Text),
            term(Childhood Obesity Terms)
            / sentence(term(Childhood Obesity Terms), term(Childhood Obesity Negative Terms)))
    or
        (fnear(200,term(Summary Text),
            near(7,term(Childhood Terms), term(Obesity Terms))
                / sentence(near(7,term(Childhood Terms), term(Obesity Terms)),term(Childhood Obe
        ))
)
```

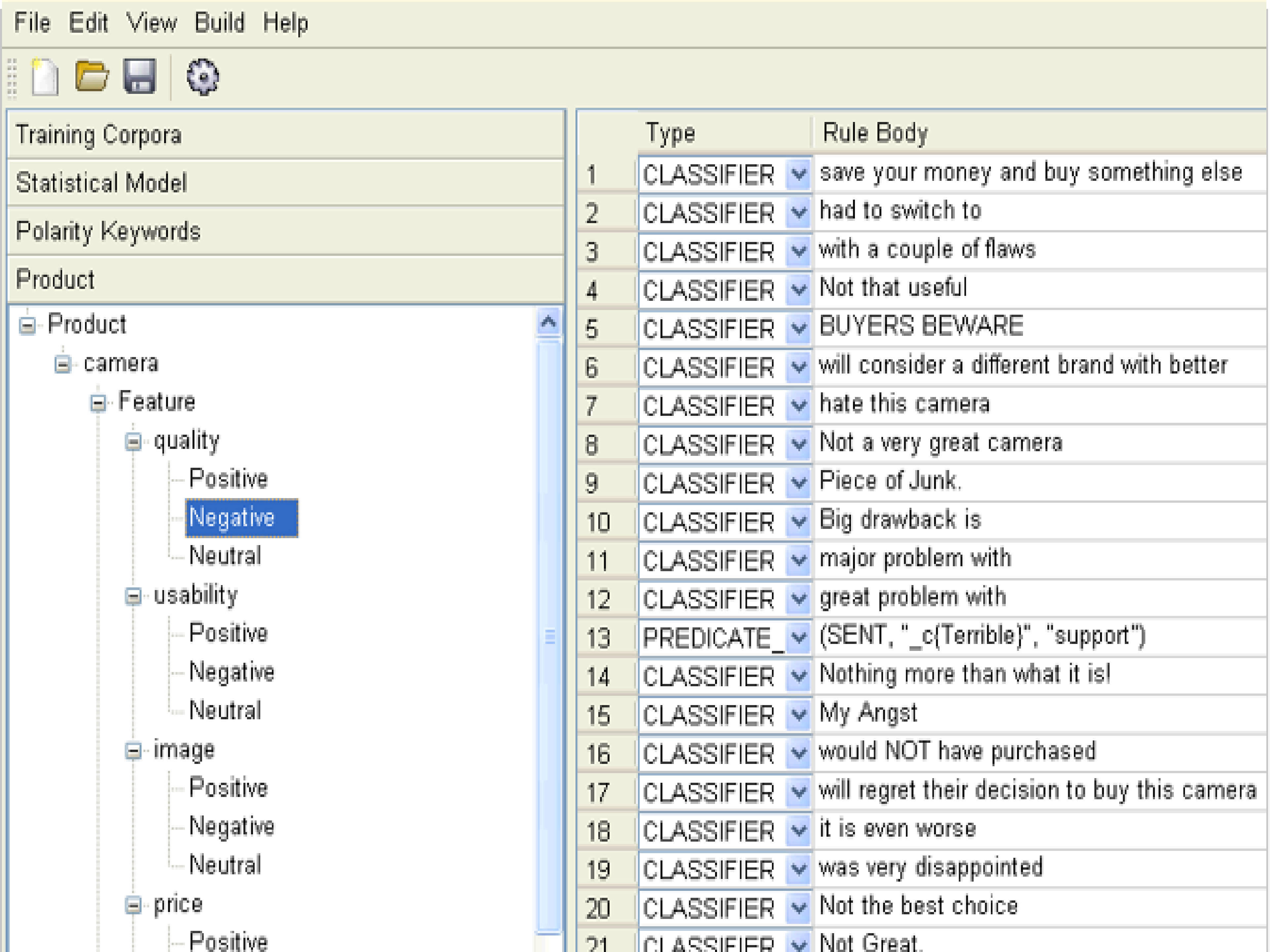
JR main.jr × CL Analytics Terms.cl × CR Analytics Rules.cr × CR main.cr ×

```
// Title Rules
SCOPE SENTENCE IN SEGMENT (DOCTITLE)
{
    DOMAIN("Analytics":HIGHEST)
    {
        KEYWORD( EXPAND "Analytics Terms\Analytics Terms.cl")
       AND NOT
        KEYWORD( EXPAND "Analytics Terms\Analytics Terms-Neg.cl" )
    }

    DOMAIN ("Analytics":HIGHEST)
    {
          KEYWORD( EXPAND "Analytics Terms\Analytics Terms-P1.cl" )
          <-7:7>
          KEYWORD( EXPAND "Analytics Terms\Analytics Terms-P2.cl" )
          AND NOT
    KEYWORD( EXPAND "Analytics Terms\Analytics Terms-Neg.cl" )
    }
}

// Summary Rules - <desc></desc>
SCOPE SENTENCE IN SEGMENT (DOCSUMMARY)
{
```

Position

xtraction | Categorization | Semantic Analysis | Terminal | Segments | Console | Tagger


LITI_Demo
English
Concepts
Top
ALIGNED_CONCEPT
CONCEPT_RULE
ORGANIZATION_S
NFLKEYWD
NFLORGS
FACT_EXT
PNA
PNAKW
FACT_RULE
PNARULE
COMPANY
INTERNET
EMAIL
TURL
LOCATION
COUNTRY
Example of regular expression matching for id
REGEX:[\w\-.]+@[\w\-.]+\.biz
REGEX:[\w\-.]+@[\w\-.]+\.com
REGEX:[\w\-.]+@[\w\-.]+\.gov
REGEX:[\w\-.]+@[\w\-.]+\.mil
REGEX:[\w\-.]+@[\w\-.]+\.net
REGEX:[\w\-.]+@[\w\-.]+\.org
REGEX:[\w\-.]+@[\w\-.]+\.co\.\w+
REGEX:[\w\-.]+@[\w\-.]+\.com\.\w+
REGEX:[\w\-.]+@[\w\-.]+\.gov\.\w+
REGEX:[\w\-.]+@[\w\-.]+\.mil\.\w+
REGEX:[\w\-.]+@[\w\-.]+\.net\.\w+
REGEX:[\w\-.]+@[\w\-.]+\.org\.\w+
REGEX:[\w\-.]+@[\w\-]+\.[\w\-]+\.[\w\-]+



File Edit View Build Help
Training Corpora
Statistical Model
Polarity Keywords
Product
Product
camera
Feature
quality
Positive
Negative
Neutral
usability
Positive
Negative
Neutral
image
Positive
Negative
Neutral
price
Positive
Type
Rule Body
1 CLASSIFIER save your money and buy something else
2 CLASSIFIER had to switch to
3 CLASSIFIER with a couple of flaws
4 CLASSIFIER Not that useful
5 CLASSIFIER BUYERS BEWARE
6 CLASSIFIER will consider a different brand with better
7 CLASSIFIER hate this camera
8 CLASSIFIER Not a very great camera
9 CLASSIFIER Piece of Junk.
10 CLASSIFIER Big drawback is
11 CLASSIFIER major problem with
12 CLASSIFIER great problem with
13 PREDICATE_ (SENT, "_c{Terrible}", "support")
14 CLASSIFIER Nothing more than what it is!
15 CLASSIFIER My Angst
16 CLASSIFIER would NOT have purchased
17 CLASSIFIER will regret their decision to buy this camera
18 CLASSIFIER it is even worse
19 CLASSIFIER was very disappointed
20 CLASSIFIER Not the best choice
21 CLASSIFIER Not Great.

Analysis Details

Description **B2C random 200 1** **ID** **147**

Documents | Taxonomy | Properties | Profiling

Filter by result: Filter by file name:

Validati...	File	Size	Duration	Success	Categories	Extractions	Categorization					
							TP	FP	FN	Precision	Recall	F-Measure
	B2C-200-r...	1,562	00:00.145	✔	5	0	1	0	1	100.00%	50.00%	67.00%
	B2C-200-r...	1,198	00:00.147	✔	9	0	0	0	1	0.00%	0.00%	0.00%
	B2C-200-r...	3,363	00:00.346	✔	21	0	1	0	7	100.00%	12.00%	22.00%
	B2C-200-r...	2,142	00:00.199	✔	14	0	0	2	4	0.00%	0.00%	0.00%
	B2C-200-r...	4,313	00:00.428	✔	14	0	1	2	0	33.00%	100.00%	50.00%
	B2C-200-r...	2,331	00:00.236	✔	7	0	2	2	0	50.00%	100.00%	67.00%
	B2C-200-r...	1,389	00:00.194	✔	7	0	1	2	1	33.00%	50.00%	40.00%
	B2C-200-r...	1,998	00:00.229	✔	13	0	4	0	5	100.00%	44.00%	62.00%
	B2C-200-r...	7,892	00:01.041	✔	29	0	3	3	6	50.00%	33.00%	40.00%
	B2C-200-r...	3,086	00:00.254	✔	9	0	1	1	1	50.00%	50.00%	50.00%
	B2C-200-r...	3,341	00:00.353	✔	8	0	2	0	6	100.00%	25.00%	40.00%
	B2C-200-r...	3,038	00:00.278	✔	5	0	0	0	3	0.00%	0.00%	0.00%

Files: **195**

Close

KAPS Group

Knowledge Architecture Professional Services

main.jr × | Analytics Terms.cl × | Analytics Rules.cr × | Consumer-Electronics_579117_This-glori.txt × | main.cr ×

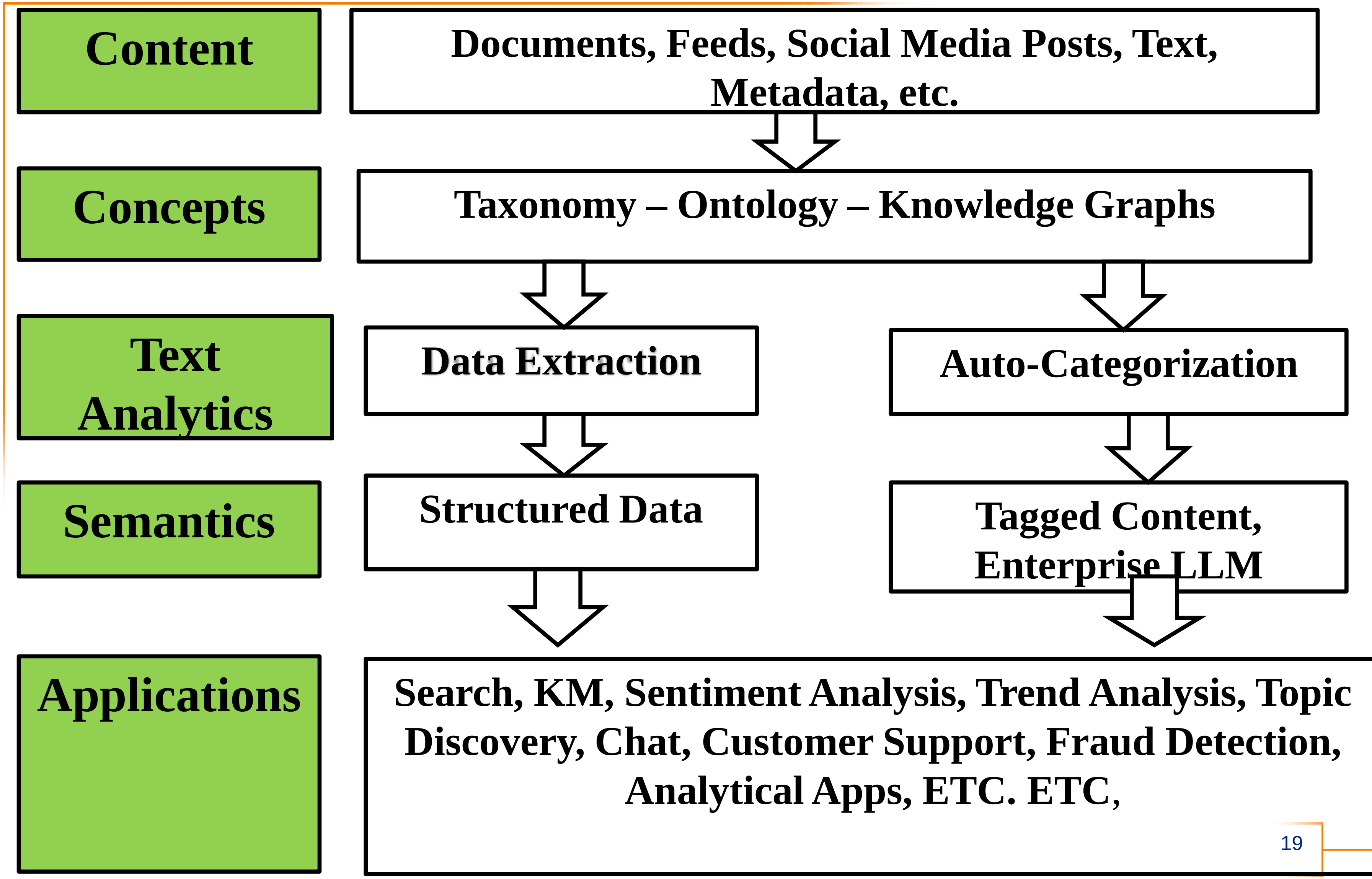
```
<title>This glorious 65-inch 4K Roku TV is a jaw-dropping $400 today</title>
<subheadline>Watch the Super Bowl in style and save $200 for snacks.</subheadline>
<desc>Best Buy is selling a Westinghouse Roku smart TV for just $400.</desc>
<story_type>Deal</story_type>
<article_type>default</article_type>
The Super Bowl is coming, and if you plan to watch the big game you're going to need a bigger screen. Best Buy has you covered today with a Westinghouse 65-inch 4K Roku smart TV for a measly $400Remove non-product link'$200 off the MSRP and one of the best prices you'll ever see.
The TV itself features 4K resolution, as well as HDR 10. We wouldn't call that true HDR as this TV doesn't get quite bright enough to qualify, but you'll certainly notice the improved picture over similarly sized sets without it.'ith Roku on board, you have access to all your favorite streaming services including Apple TV, HBO, Hulu, Netflix, Starz, and more. This TV also works with Amazon Alexa and Google Assistant for voice control, as well as integrating with other smart home devices.
For ports, it has three HDMI one of which has HDCP 2.2, as well as one USB 2.0 for thumb drives and other storage devices. It's also packing one digital optical audio out and analog audio out. For connectivity, you get Wi-Fi and Ethernet, but no Bluetooth.
Add it all up and you've got one heck of a bargain for just $400. So go grab one and get read to watch the big game in style.
[<strong>Today's deal:</strong> Westinghouse 65-inch 4K Roku smart TV for $400 at Best BuyRemove non-product link]
```

Analysis Details

Extraction | Categorization | Semantic Analysis | Terminal | Statistics | Segments

Description B2C ran

Text Analytics and GenAI in the Enterprise



Text Analytics Software – Full Platform Essential Functionality

- Taxonomy – structure for auto-cat, minimum management
  - Orthogonal categories, good hierarchy
- Content – multiple document types, conversion, translation
  - Importance of good example documents for each category
- Rules – categorization, disambiguation – programming language
- Terms – manage sets of terms – few to 100's
- Testing – apply rules and generate scores
  - Analytics – quality of rules
  - Testers – SMEs, end users – need multiple to overcome bias
- Supplemental
  - Text mining

Text Analytics Workshop

- Text Analytics Development

Text Analytics Development: Categorization Basics

- Representation of domain knowledge – taxonomy, ontology
- Categorization – Most basic to human cognition
  - Most difficult to do with software
  - Subject, tacit knowledge, sentiment, expertise
- Beyond Categorization – making everything else smarter
  - Disambiguation – within categorization and entity extraction
- No single correct categorization
  - Women, Fire, and Dangerous Things

Intel Mini-POC
Categorization Techniques – Two Basic Approaches

- Machine Learning – Bayesian, Vector space, CNN, RNN
  - Create a statistical/neural net signature and compare new content
  - Results are poor, difficult to improve, needs large numbers of representative documents
- Categorization language - AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
  - Good results, flexible and power – DIST, etc.
  - Need to learn a categorization language

- Boehringer Pilot One Drug Names Disease
  - English
    - Categorizer
      - Top
        - Diseases
          - arthritis
          - Benign Prostatic Hyperpla
          - Cancer
          - Hypertension
          - Deep Vein Thrombosis
          - HIV
          - Pulmonary Disease
        - Drug Names
          - afatinib
          - clonidine
          - dabigatran
          - meloxicam
          - tamsulosin
          - telmisartan
          - tiotropium
    - Concepts
      - Top
        - BI Drugs
        - Diseases
          - arthritis

```
(OR,
_/article/title:"[arthritis]",

(AND, _/article/mesh:"[arthritis]",_/article/abstract:"[arthritis]"),

(MINOC_2, _/article/abstract:"[arthritis]"),

(START_500, (MINOC_2,"[arthritis]"))
)
```

Text Analytics Workshop
Machine/Deep Learning and Rules

- Claim – ML is faster to develop – only if unsupervised – typically bad results
- Selecting documents takes time and effort – and difficult to do well
- Rules (and Taxonomy) can provide structure and better training sets
- ML can provide terms for rules
- Current trend – how to combine
- One solution is content model – statistical based on sections

Text Analytics Development: Categorization Process
Start with Taxonomy and Content

- Starter Taxonomy - If no taxonomy, develop initial high level
  - Textbooks, glossaries, Intranet structure
  - Organization Structure – facets, not taxonomy
- Analysis of taxonomy – suitable for categorization
  - Structure – not too flat, not too large, Orthogonal categories
  - Best = rich synonyms – starter cat rules
- Selection of “training sets” – 20-50-100 per category
  - SME input, search logs, information interviews
  - Trick – category name in file name
- Automated selection of training sets
  - Taxonomy nodes as first categorization rules
- Social Media – external searches
  - Sentiment – Forums – ranked posts – 1-5

Text Analytics Workshop
Text Analytics Development: Categorization Process

- Start: Term building – from content
  - Text Mining – basic set of terms that are unique to topic
  - Multiple passes – sub-types of content
  - Clustering – word or tag clouds
- Develop initial rules – per category
  - 1.) ½ of training set – add terms to rules – 90-100% recall
  - 2.) Test against ½ of all training sets – remove terms – precision
  - 3.) Multiple refinement rounds
- Test against more, new content – more terms, refine logic of rules – distance operators, utilize metadata - carefully
- Develop templates – separate logic and vocabulary
- Repeat until “done”

idg-autocategortization-2 C:\Users\tomr\StudioProjects\idg-autocategortization

- .idea
- ann
  - B2C-200-random-ann
- documents
- gen
- package
- reports
- rules
  - Analytics Rules
  - Analytics Terms
    - Analytics Terms.cl
    - Analytics Terms-Neg.cl
    - Analytics Terms-P1.cl
    - Analytics Terms-P2.cl
    - Big Data Terms.cl
    - Big Data Terms-Neg.cl
    - Big Data Terms-P1.cl
    - Big Data Terms-P2.cl
    - Business Intelligence Terms.cl
    - Business Intelligence Terms-Neg.cl
    - Business Intelligence Terms-P1.cl
    - Business Intelligence Terms-P2.cl
    - Data Mining Terms.cl
    - Data Mining Terms-Neg.cl
    - Data Mining Terms-P1.cl

```
AI-based recommendations
BigLake
DaaS
Data Cloud Alliance
Data monetization
Google Cloud Ready BigQuery
Google Cloud Spanner
Vertex AI Model Registry
Vertex AI Workbench
algorithms
analytics
data as a service
data-as-a-service
data as an asset
data pipelines
data privacy and security
data silo
data silos
graph algorithms
graph databases
high performance computing
high-performance computing
key data assets
model lifecycle management
```

Git | TODO | Problems | Knowledge Graph | ExcelReader | Extraction | Categorization | Semantic Analysis | Te

Text Analytics Workshop
What Makes a Good Term?

- Keywords – NO!!!
  - Mostly related terms, not terms that indicate what a document is about
  - Evidence terms – appear in document about X, not in general
  - New project either 0 or over 800 “keywords”
- 3 types of evidence terms
  - Single phrases that appear in target document and not others
  - 2 words/phrases that are near each other (7-10 words)
  - Negative terms – if found, discount - deal with overlapping taxonomy

JR main.jr × CL Analytics Terms.cl × CR Analytics Rules.cr × CR main.cr ×

```
// Title Rules
SCOPE SENTENCE IN SEGMENT (DOCTITLE)
{
    DOMAIN("Analytics":HIGHEST)
    {
        KEYWORD( EXPAND "Analytics Terms\Analytics Terms.cl")
       AND NOT
        KEYWORD( EXPAND "Analytics Terms\Analytics Terms-Neg.cl" )
    }

    DOMAIN ("Analytics":HIGHEST)
    {
          KEYWORD( EXPAND "Analytics Terms\Analytics Terms-P1.cl" )
          <-7:7>
          KEYWORD( EXPAND "Analytics Terms\Analytics Terms-P2.cl" )
          AND NOT
    KEYWORD( EXPAND "Analytics Terms\Analytics Terms-Neg.cl" )
    }
}

// Summary Rules - <desc></desc>
SCOPE SENTENCE IN SEGMENT (DOCSUMMARY)
{
```

Position

xtraction Categorization Semantic Analysis Terminal Segments Console Tagger

Content Structure Models
No Such Thing as Unstructured Text

- Documents are not unstructured – poly-structure
  - Words, Sentences, and Paragraphs
  - Sections and Clusters
- Sections – Variety - “Abstract” to Function “Evidence”
  - Categorization – Title, Sub-title, Abstract, Executive Summary
  - Special - Results / Methods / Objectives
  - Systemic Text – Acknowledgements, References
  - Data Sections – Major and throughout – Tables, etc.
- “Summary” – human judgement on what the document is about
- Bag of Words = Bag of S**t

Providing actuarial analyses and modeling of health reform ideas to stabilize individual insurance markets and continuing RWJF's actuarial challenge

Fund Description: To continue dissemination and analytical activities associated with the results of the 2017 RWJF Actuarial Challenge, in which teams of actuaries proposed solutions to stabilize the individual insurance market.

SUMMARY

This project will continue the work of the RWJF Actuarial Challenge. The Actuarial Challenge took place in early 2017. The final results included policy suggestions for stabilizing the individual market, including elements such as reinsurance, auto enrollment, and other market reforms. Milliman organized the challenge and simulated the winning proposals, providing estimates of how they would impact enrollment and public and private spending. As the prospect for bipartisan health reform increases, there is an increased demand for disseminating these results and for potentially engaging in some additional actuarial modelling. The challenge process and results are reviewed as technically credible and politically nonpartisan. As the effort to repeal and replace has abated, there may be an opportunity to bring some bipartisan suggestions for reform forward. Several organizations and RWJF are planning meetings and presentations for the next several months, with the goal of sharing the challenge results with policymakers and other stakeholders. At some point, this will most likely result in engaging Milliman in simulating some refined version of some elements of the winning proposals. It may ultimately be recommended to stage a second round of the challenge. The deliverables will include meetings, presentations, discussions with stakeholders, and, potentially, additional simulations. The policy environment and demand for these products will help determine the size and scope of this project.

COMMONWEALTH OF VIRGINIA
DEPARTMENT OF TRANSPORTATION

WORK ORDER

Contract ID. No.: P00091296B00 FHWA No.: BH-BR03(259); BH-BR03(261) Work Order No.: 2

State Project No.: BRDG-041-718, B660; BRDG-041-719, B661 Category: MISC

Original Contract Value $ 646,308.25 Total of Other Work Orders $ 0

NOTE: If additional space is needed, use an additional sheet(s) and label as Supplemental Attachment #.

I. LOCATION AND DESCRIPTION OF PROPOSED WORK:

Time Extension

Dec. 22, 2010 to March 13, 2011 Suspension of work.

March 14, 2009 to April 15, 2011 Extension of 33days

50 days total time extension

One month additional Maintenance of Traffic

II. RESPONSIBLE CHARGE ENGINEER'S EXPLANATION OF NECESSITY FOR PROPOSED WORK:

This Work Order is needed to extend the contract time to allow the contractor to place the Asphalt Concrete TY. SM9.5A, during warmer weather. Asphalt producers have shut down and will not be open until warmer weather returns. All remaining work to be completed at current contract prices.

"Burleigh Construction Company Inc. and VDOT agree that this Work Order fully resolves and settles all claims, demands or damages of any kind relating to or arising out of the work set forth in this Work Order, including but not limited to delay, impact and acceleration."

The additional Maintenance of Traffic cost are to cover the cost of rented traffic control equipment during the time when additional work was taking place.

III. FUNDING SOURCE/CHARGE Federal 80% / State 20%

IV. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT PRIOR TO APPRVOAL OF THIS WORK ORDER IS Dec. 21, 2010

V. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT UPON APPROVAL OF THIS WORK ORDER IS Apr. 15, 2011

Content Structure Models
Structure Rules Basic Logic

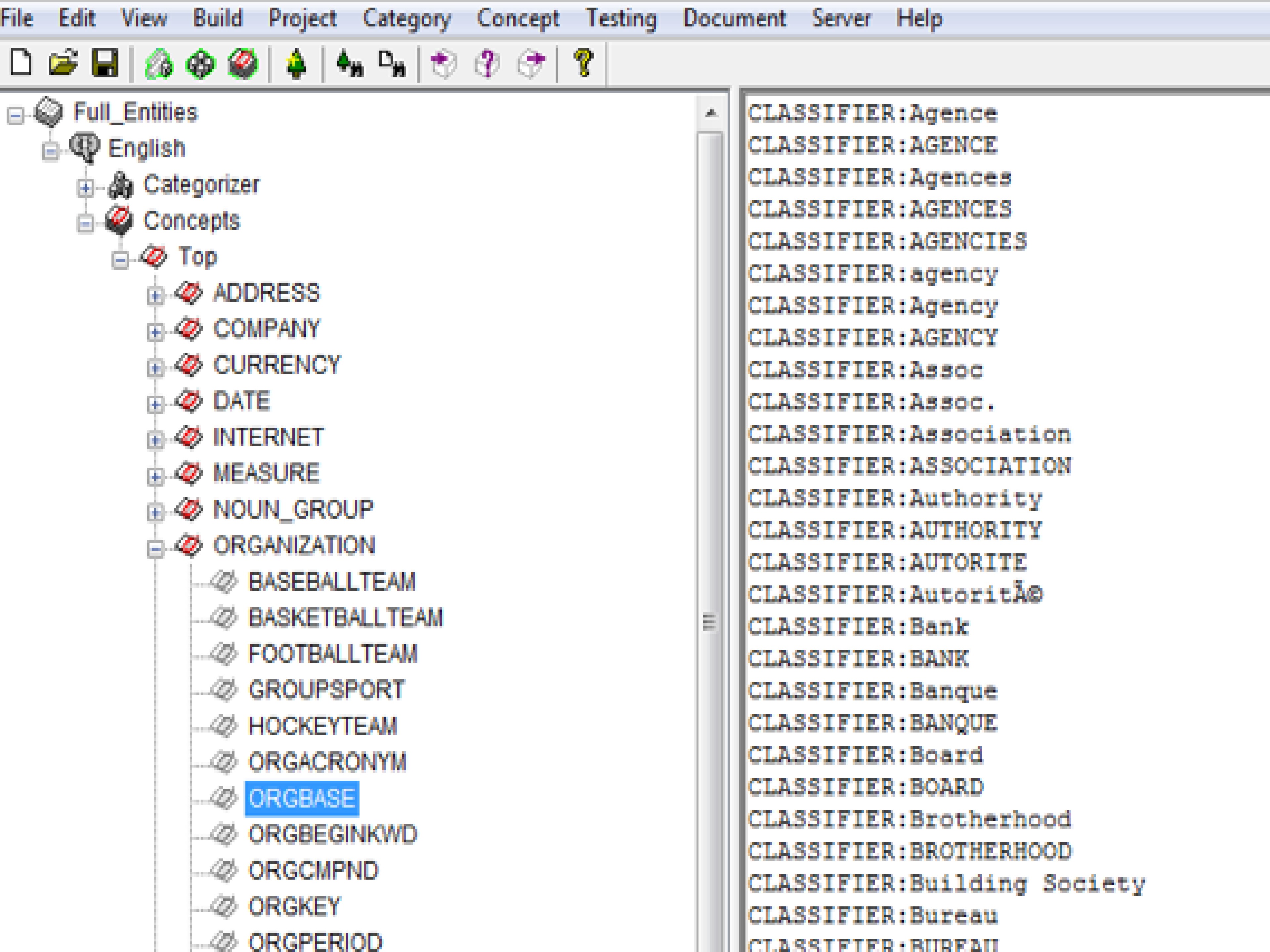
- Count terms that are in the list and in the first 100 words unless there are negative terms within 7 words
- Count terms that are in the list and that are within 500 words after a Document Summary Indicator unless there are negative terms within 7 words
  - Document Summary Indicators – 29 terms “Executive Summary”, “Issue Brief”, “Abstract”
- Terms in the list can be phrases or sets of terms within 7 words of each other
- Negative terms are ones that often show up but should belong to another category – they vary by category
  - Child & Family Well-being – “Coverage”, “Obesity”, “Nurses”

RWJF Mini-POC
Overview Average Scores

	Recall	Precision	Precision Top 10
With Sections	95%	92%	99%
Full Text	71%	41%	81%
Difference	24%	51%	18%

Text Analytics Workshop
Development: Entity Extraction Process

- Facet Design – from Knowledge Audit, K Map
- Catalogs – linked data or convert to internal:
  - Organization – internal resources
  - People – corporate yellow pages, HR
  - Include variants
  - Scripts to convert catalogs – programming resource
- Build initial rules – follow categorization process
  - Differences – scale, threshold – application dependent
  - Recall – Precision – balance set by application
  - Issue – disambiguation – Ford company, person, car
- Unknown entities – NLP rules – “cap cap said”


File Edit View Build Project Category Concept Testing Document Server Help
Full_Entities
English
Categorizer
Concepts
Top
ADDRESS
COMPANY
CURRENCY
DATE
INTERNET
MEASURE
NOUN_GROUP
ORGANIZATION
BASEBALLTEAM
BASKETBALLTEAM
FOOTBALLTEAM
GROUPSPORT
HOCKEYTEAM
ORGACRONYM
ORGBASE
ORGBEGINKWD
ORGCMPND
ORGKEY
ORGPERIOD
CLASSIFIER:Agence
CLASSIFIER:AGENCE
CLASSIFIER:Agences
CLASSIFIER:AGENCES
CLASSIFIER:AGENCIES
CLASSIFIER:agency
CLASSIFIER:Agency
CLASSIFIER:AGENCY
CLASSIFIER:Assoc
CLASSIFIER:Assoc.
CLASSIFIER:Association
CLASSIFIER:ASSOCIATION
CLASSIFIER:Authority
CLASSIFIER:AUTHORITY
CLASSIFIER:AUTORITE
CLASSIFIER:AutoritÃ©
CLASSIFIER:Bank
CLASSIFIER:BANK
CLASSIFIER:Banque
CLASSIFIER:BANQUE
CLASSIFIER:Board
CLASSIFIER:BOARD
CLASSIFIER:Brotherhood
CLASSIFIER:BROTHERHOOD
CLASSIFIER:Building Society
CLASSIFIER:Bureau

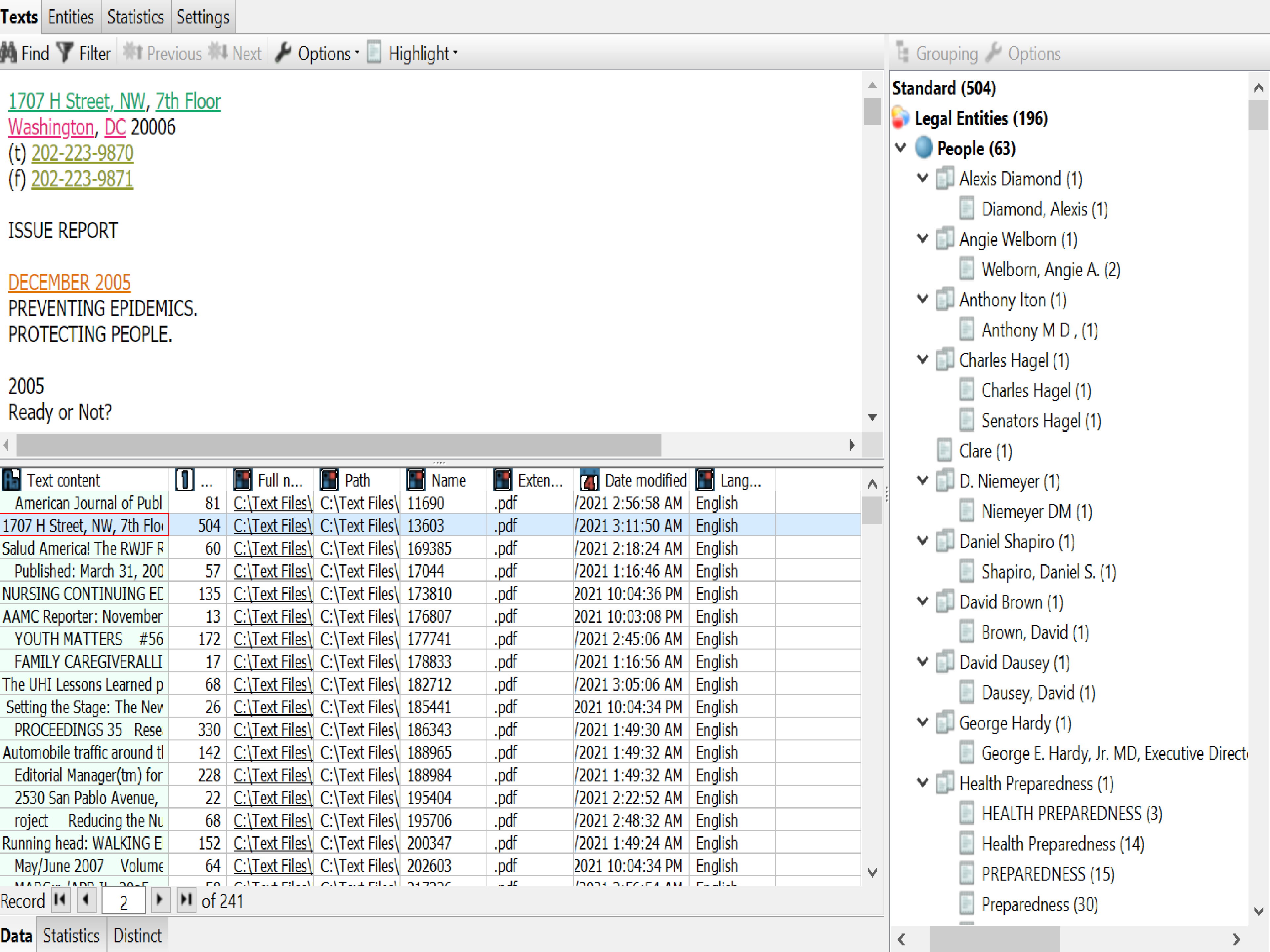
Solution Development

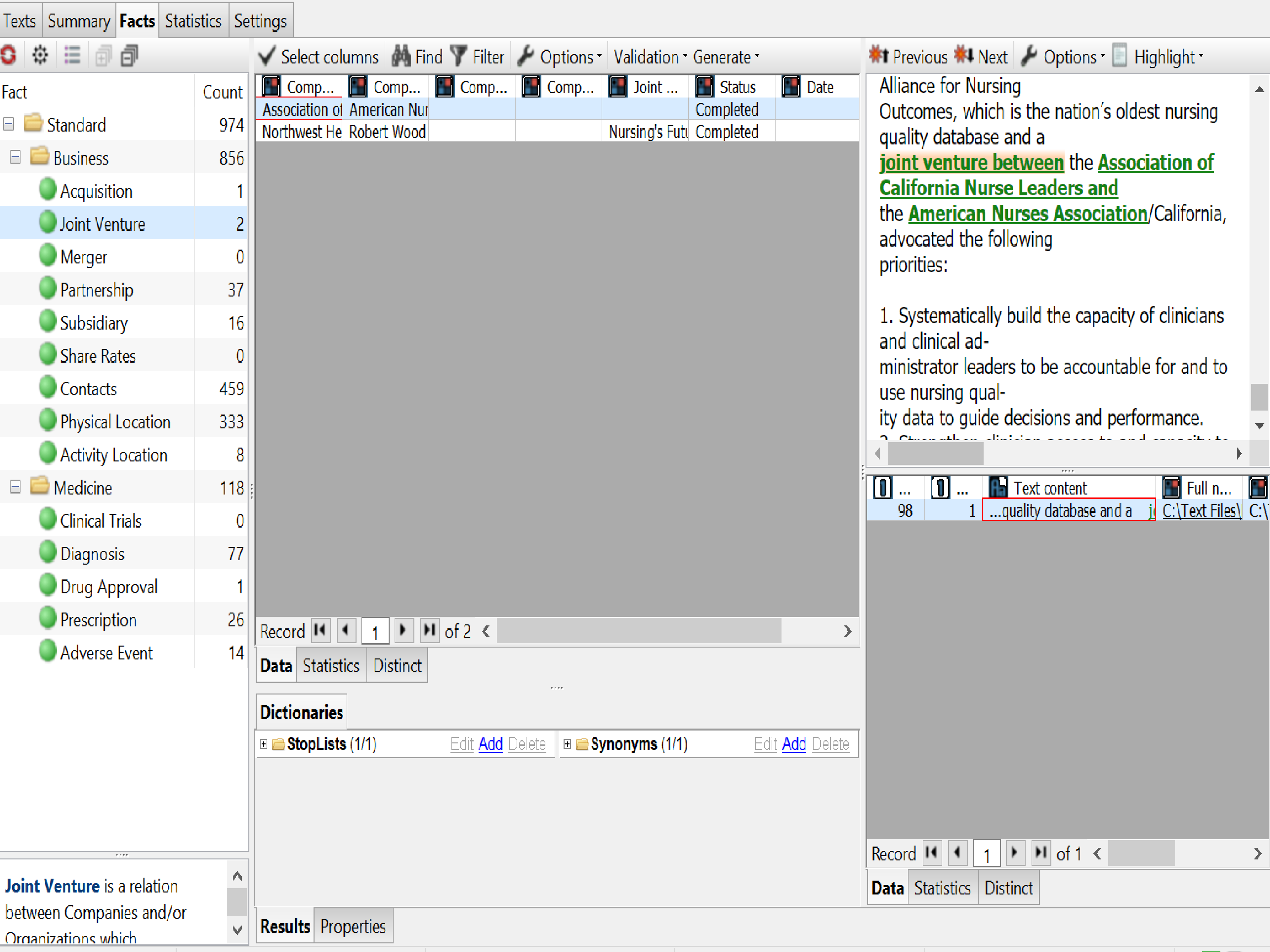
Semantic Model – Elements ("facets")

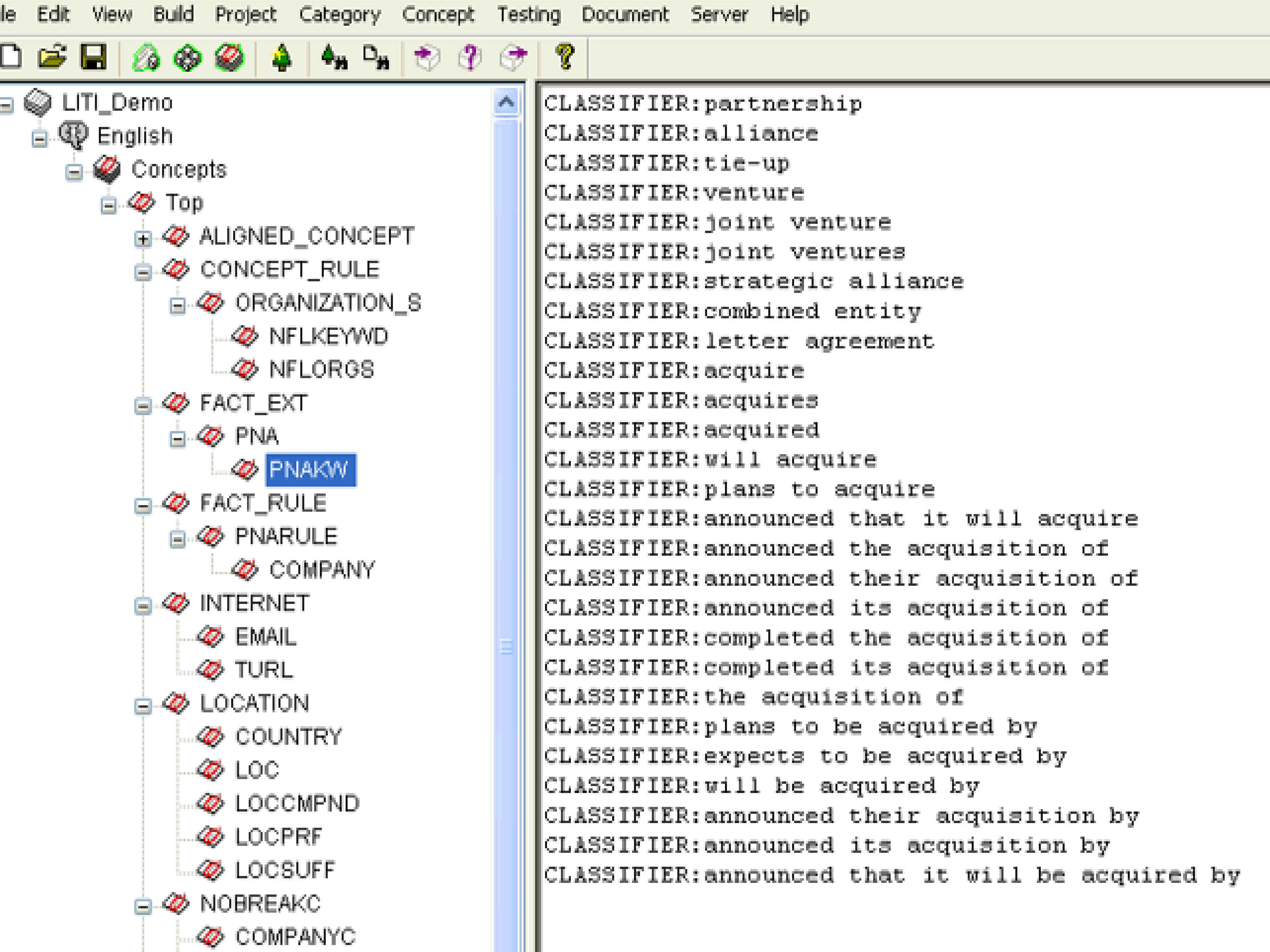
- Content Type
  - Source of Materials
  - DWR,
  - Work Order,
  - Work Order-Related
  - Project Profile
- Project No/Contract No/UPC
- Location: District, Jurisdiction, Route
- Type of Work
- Award Amount
- Manufacturers and Suppliers
- Contractors
- Materials
- Equipment
- Pay Items
- Work Order Category
- Work Issue
  - Drainage
  - Utility
  - Weather
  - Plan-Related
  - Work Zone-Related

Text Analytics Forum (TAF) Introduction Context: Fact Extraction

- Two types
  - Find specific entities
    - Not all addresses, Company addresses
  - Find relationship of two entities
    - Company A merges with Company B
- Need rules that can process context around key data
  - Dictionaries
  - Patterns – CAP CAP said
- Software selection is a key - rules
  - If only ML, poor results


Texts
Entities
Statistics
Settings
Find
Filter
Previous
Next
Options
Highlight
Grouping
Options
1707 H Street, NW, 7th Floor
Washington, DC 20006
(t) 202-223-9870
(f) 202-223-9871
ISSUE REPORT
DECEMBER 2005
PREVENTING EPIDEMICS.
PROTECTING PEOPLE.
2005
Ready or Not?
Text content
Full n...
Path
Name
Exten...
Date modified
Lang...
American Journal of Publ 81 C:\Text Files\ C:\Text Files\ 11690 .pdf /2021 2:56:58 AM English
1707 H Street, NW, 7th Flo 504 C:\Text Files\ C:\Text Files\ 13603 .pdf /2021 3:11:50 AM English
Salud America! The RWJF R 60 C:\Text Files\ C:\Text Files\ 169385 .pdf /2021 2:18:24 AM English
Published: March 31, 200 57 C:\Text Files\ C:\Text Files\ 17044 .pdf /2021 1:16:46 AM English
NURSING CONTINUING ED 135 C:\Text Files\ C:\Text Files\ 173810 .pdf 2021 10:04:36 PM English
AAMC Reporter: November 13 C:\Text Files\ C:\Text Files\ 176807 .pdf 2021 10:03:08 PM English
YOUTH MATTERS #56 172 C:\Text Files\ C:\Text Files\ 177741 .pdf /2021 2:45:06 AM English
FAMILY CAREGIVERALLI 17 C:\Text Files\ C:\Text Files\ 178833 .pdf /2021 1:16:56 AM English
The UHI Lessons Learned p 68 C:\Text Files\ C:\Text Files\ 182712 .pdf /2021 3:05:06 AM English
Setting the Stage: The New 26 C:\Text Files\ C:\Text Files\ 185441 .pdf 2021 10:04:34 PM English
PROCEEDINGS 35 Rese 330 C:\Text Files\ C:\Text Files\ 186343 .pdf /2021 1:49:30 AM English
Automobile traffic around tl 142 C:\Text Files\ C:\Text Files\ 188965 .pdf /2021 1:49:32 AM English
Editorial Manager(tm) for 228 C:\Text Files\ C:\Text Files\ 188984 .pdf /2021 1:49:32 AM English
2530 San Pablo Avenue, 22 C:\Text Files\ C:\Text Files\ 195404 .pdf /2021 2:22:52 AM English
roject Reducing the Nu 68 C:\Text Files\ C:\Text Files\ 195706 .pdf /2021 2:48:32 AM English
Running head: WALKING E 152 C:\Text Files\ C:\Text Files\ 200347 .pdf /2021 1:49:24 AM English
May/June 2007 Volume 64 C:\Text Files\ C:\Text Files\ 202603 .pdf 2021 10:04:34 PM English
Record 2 of 241
Data
Statistics
Distinct
Standard (504)
Legal Entities (196)
People (63)
Alexis Diamond (1)
Diamond, Alexis (1)
Angie Welborn (1)
Welborn, Angie A. (2)
Anthony Iton (1)
Anthony M D , (1)
Charles Hagel (1)
Charles Hagel (1)
Senators Hagel (1)
Clare (1)
D. Niemeyer (1)
Niemeyer DM (1)
Daniel Shapiro (1)
Shapiro, Daniel S. (1)
David Brown (1)
Brown, David (1)
David Dausey (1)
Dausey, David (1)
George Hardy (1)
George E. Hardy, Jr. MD, Executive Directo
Health Preparedness (1)
HEALTH PREPAREDNESS (3)
Health Preparedness (14)
PREPAREDNESS (15)
Preparedness (30)


Texts
Summary
Facts
Statistics
Settings
Select columns
Find
Filter
Options
Validation
Generate
Previous
Next
Options
Highlight
Fact
Count
Standard 974
Business 856
Acquisition 1
Joint Venture 2
Merger 0
Partnership 37
Subsidiary 16
Share Rates 0
Contacts 459
Physical Location 333
Activity Location 8
Medicine 118
Clinical Trials 0
Diagnosis 77
Drug Approval 1
Prescription 26
Adverse Event 14
Comp... Comp... Comp... Comp... Joint ... Status Date
Association of American Nur Completed
Northwest He Robert Wood Nursing's Futu Completed
Record 1 of 2
Data
Statistics
Distinct
Dictionaries
StopLists (1/1) Edit Add Delete
Synonyms (1/1) Edit Add Delete
Results
Properties
Joint Venture is a relation between Companies and/or Organizations which
Alliance for Nursing
Outcomes, which is the nation's oldest nursing
quality database and a
joint venture between the Association of
California Nurse Leaders and
the American Nurses Association/California,
advocated the following
priorities:
1. Systematically build the capacity of clinicians
and clinical ad-
ministrator leaders to be accountable for and to
use nursing qual-
ity data to guide decisions and performance.
Text content
Full n...
98 1 ...quality database and a
C:\Text Files\
Record 1 of 1
Data
Statistics
Distinct



le Edit View Build Project Category Concept Testing Document Server Help
LITI_Demo
English
Concepts
Top
ALIGNED_CONCEPT
CONCEPT_RULE
ORGANIZATION_S
NFLKEYWD
NFLORGS
FACT_EXT
PNA
PNAKW
FACT_RULE
PNARULE
COMPANY
INTERNET
EMAIL
TURL
LOCATION
COUNTRY
LOC
LOCCMPND
LOCPRF
LOCSUFF
NOBREAKC
COMPANYC
CLASSIFIER:partnership
CLASSIFIER:alliance
CLASSIFIER:tie-up
CLASSIFIER:venture
CLASSIFIER:joint venture
CLASSIFIER:joint ventures
CLASSIFIER:strategic alliance
CLASSIFIER:combined entity
CLASSIFIER:letter agreement
CLASSIFIER:acquire
CLASSIFIER:acquires
CLASSIFIER:acquired
CLASSIFIER:will acquire
CLASSIFIER:plans to acquire
CLASSIFIER:announced that it will acquire
CLASSIFIER:announced the acquisition of
CLASSIFIER:announced their acquisition of
CLASSIFIER:announced its acquisition of
CLASSIFIER:completed the acquisition of
CLASSIFIER:completed its acquisition of
CLASSIFIER:the acquisition of
CLASSIFIER:plans to be acquired by
CLASSIFIER:expects to be acquired by
CLASSIFIER:will be acquired by
CLASSIFIER:announced their acquisition by
CLASSIFIER:announced its acquisition by
CLASSIFIER:announced that it will be acquired by

Text Analytics Workshop
Multiple Fact Extraction – Key Lessons

- Need rules that can process context around key data
  - Tool and expertise needed
- Separate logic and text – understandable, maintenance
  - Previous rules were too complex – went for pages
- Add dynamic section identification rules
  - Flexible rules needed to handle huge variation in documents
- Software selection is a key
  - Initial estimates of additional 4 months was too high (expensive) and too low (no way to get from here to there)

Text Analytics Workshop

- Applications

Text Analytics Workshop Process – Hybrid Tagging

Content Hub

Content
Metadata

Tagging
Editing

TA / Gen AI
Suggest Tags

Documents
Authoring

Taxonomy
Categorization Rules
Data Extraction Rules

Text Analytics Workshop: Information Environment Metadata – Tagging – Mind the Gap

- Tagging documents with taxonomy nodes is tough
  - And expensive – central or distributed
- Authors – Experts in the subject matter, terrible at categorization
  - Intra and Inter inconsistency, “intertwingleness”
  - Choosing tags from taxonomy – complex task
  - Folksonomy – almost as complex, wildly inconsistent
  - Resistance – not their job, cognitively difficult = non-compliance

Text Analytics Workshop
Hybrid Model: Content Management

- Publish Document -> Text Analytics analysis -> suggestions for categorization, entities, metadata - > present to author
  - Cognitive task is simple -> react to a suggestion instead of select from head or a complex taxonomy
  - Feedback – if author overrides -> suggestion for new category
- External Information - human effort is prior to tagging
  - More automated, human input as specialized process – periodic evaluations
  - Precision usually more important
  - Linked Data – How important? What resources?

Text Analytics Workshop
Applications: Application Areas

- Search and Search-based – Info Apps
- Risk management, insurance price optimization
- Healthcare – image processing, treatment optimization
- Fraud detection, fake news, Anti-Money Laundering
- Contextual advertising, rich personalization
- Automated chat customer support, etc.
- Spam filtering, customer churn prediction, Cybercrime prevention
- Social media analysis, Customer and Business Intelligence
- Robotics process automation, augmented analytics

- All of the above and more

Text Analytics and Search
Multi-dimensional and Smart

- Search continues to underperform
- Faceted Navigation has become the basic/ norm
  - Facets require huge amounts of metadata
  - Entity / noun phrase extraction is fundamental
  - Automated with disambiguation (through categorization)
- Taxonomy – two roles – subject/topics and facet structure
  - Complex facets and faceted taxonomies
- Clusters and Tag Clouds – discovery & exploration
- Auto-categorization – aboutness, subject facets
  - This is still fundamental to search experience
- InfoApps only as good as fundamentals of search

Text Analytics Workshop: Applications Expertise Analysis

- Expertise Analysis
  - Experts think & write differently – process, chunks
- Expertise Characterization for individuals, communities, documents, and sets of documents
  - Automatic profiles – based on documents authored, etc.
- Applications:
  - Business & Customer intelligence, Voice of the Customer
  - Deeper understanding of communities, customers
  - Security, threat detection – behavior prediction
  - Expertise location- Generate automatic expertise characterization
- Political – conservative and liberal minds/texts
  - Disgust, shame, cooperation, openness

Social Media Applications
Voice of the Customer / Voter / Employee

- Detection of a recurring problem categorized by subject, customer, client, product, parts, or by representative.
- Analytics to evaluate and track the effectiveness:
  - Representatives, policies, programs, actions
- Detect recurring or immediate problems – high rate of failure, etc.
- Competitive intelligence – calls to switch from brand X to Y in a particular region
- Subscriber mood before and after a call – and why
- Pattern matching of initial motivation to subsequent actions – optimize responses and develop proactive steps

Social Media Applications
Behavior Prediction – Telecom Customer Service

- Problem – distinguish customers likely to cancel from mere threats
- Basic Rule
  - (START_20, (AND, (DIST_7,"[cancel]", "[cancel-what-cust]"),
  - (NOT,(DIST_10, "[cancel]", (OR, "[one-line]", "[restore]", “[if]”)))))
- Examples:
  - customer called to say he will **cancell** his **account** **if** the does not stop receiving a call from the ad agency.
  - and context in text
- Combine text analytics with Predictive Analytics and traditional behavior monitoring for new applications

Social Media Applications
Pronoun Analysis: Fraud Detection; Enron Emails

- Patterns of “Function” words reveal wide range of insights
- Function words = pronouns, articles, prepositions, conjunctions.
- Areas: sex, age, power-status, personality – individuals and groups
- Lying / Fraud detection: Documents with lies have
  - Fewer and shorter words, fewer conjunctions, more positive emotion words
  - More use of “if, any, those, he, she, they, you”, less “I”
  - More social and causal words, more discrepancy words
- Current research – 76% accuracy in some contexts
- Part of analytical effort – future research

Text Analytics Workshop

- AI and Text Analytics

Text Analytics Forum (TAF) Workshop
AI and Text Analytics

- AI is everywhere and growing
- Two early approaches
  - Symbolic AI – rules – wrong kind of rules
  - Neural networks – ML – too stupid
- More memory and faster computers
  - Enables sets of neural nets (Deep Learning)
- Current Focus – Gen AI and LLMs
- We’ve seen this AI hype before
  - Starting the trough of disillusionment?
- BUT – This is more substantive

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Forum (TAF) Workshop
Gen AI Basics

- ChatGPT and LLMs – AI's huge leap forward
  –Basically type ahead on steroids – predict next word
  - Keys – scale and transformers
- Requires huge amounts of content, huge neural nets
  – GPT-3.5 small = 1.3 B parameters, large = 175 B parameters
  – GPT-4 – rumored to have 1.76T
- Applications seem endless – customer support to writing poetry
  – Write a poem in which every word begins with Q
  – Write code, business plans, plan office parties, etc.
- Test – sentence completion
  – GPT-3 – 86.4% - MT-NLG – 87.2 % correct
- Small gain for 3x the scale, costs millions
- The model is trained on the Selene supercomputer, thousands of GPUs – who can afford?

Analytics Forum (TAF) Workshop
Gen AI and Text Analytics: Conceptual Limits

- General:
  - Tendency to hallucinate
  - Lack of transparency – no one knows how they work
  - Amount and quality of data needed to train
  - Security issues - Jailbreaks – get Gen AI to break its own rules
- Enterprise:
  - Public vocabulary, not reflective of enterprise language
  - Fine tune with enterprise content – costly, low accuracy
- Mimic human language, not understand. No common sense
  - No notion of causality – only correlation
  - Humans understand causality by age 1 year
- Most popular answer, not necessarily the best
  - Superficial, stereotypical writings
- Difficulty with complexity
  - Example: ask should I checkmate?

Text Analytics Forum (TAF) Workshop
How Overcome Limitations - General

- AI Projects fail 2x higher rate than general IT projects
- Two Approaches
  - Improve Gen AI – Prompt Engineering, RAG
  - Add Text Analytics
- Build Text Analytics-LLM Foundation
  - Combine text analytics and Gen AI
    - Build Foundation – accurate training sets
    - Text normalization, noise reduction
    - Apply Foundation – within apps, advanced prompts
      - Combine – Gen AI as rough/first drafts
      - Richer contexts – add taxonomies, knowledge graphs
  - Consistent data
  - Hidden patterns – feed Gen AI

Text Analytics Forum (TAF) Workshop
How to Overcome Limitations - Specifics

- Enterprise LLM – more control over content
- Hallucinations
  - Use text analytics to build better training sets
  - Auto-classification to check Gen AI output
- Transparency
  - Auto-summarization – input into asking what features it used
- Training Data
  - General – data cleaning, remove dups
  - Add context – metadata, summaries,
- Security
  - Detect bias, balanced training set
  - Jailbreaks – content filtering, intent recognition, sentiment analysis

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Forum (TAF) Workshop
Gen AI and Text Analytics

- What do you get by adding TA? Accuracy! Which impacts all attempts to utilize and analyze documents
- Enterprise Weak Link – training sets
  - Issue – quality of content – cost of getting good content
  - Statistical – there is more bad content than good
  - Human curation is expensive and inconsistent (75% agree)
  - Search engines not accurate enough
- Answer – semi-automatic content curation – auto-categorization and human curation
- Gen AI provides general answer (Recall, draft), TA adds precision

Text Analytics Forum (TAF) Workshop
Gen AI and Text Analytics Together

- Text Analytics can add structure (conceptual and linguistic) to AI
  - Multiple types of Knowledge Organization
  - Taxonomy, ontology, knowledge graphs
  - Content structure models – eliminate noise of Bag of Words
  - Brain is more than a network – universal language detector
- Knowledge Models
  - Use in tagging – training sets and more
  - Use in prompts – adding context
  - Use in applications – variety, platform
- Gen AI suggests taxonomy nodes, terms for rules – based on documents, general
- Gen AI capabilities – sentiment analysis, summarization, tagging

Text Analytics Workshop

- Building a Text Analytics Foundation

Text Analytics Workshop
Smart Start: Think Big, Start Small, Scale Fast

- Think Big: Infrastructure Foundation
  - Based on deep understanding of entire information environment – Iterative process
  - Avoid expensive mistakes – dead end technology
- Start Small: Pilot or POC
  - Immediate value and learn by doing
  - Easier to get management buy-in
- Scale Fast: Build on the foundation
  - Semantic Infrastructure – catonomies, ontologies
  - First project +10%, Subsequent projects – 50%

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
The start and foundation: Knowledge Audit

- Knowledge Map - Understand what you have, what you are, what you want
- Contextual interviews, content analysis, surveys, focus groups, ethnographic studies, text mining
- Category modeling – Monkey, Panda, Banana
- 4 Dimensions – Content, People, Technology, Activities
- Strategic Vision and Change Management
  - Format – reports, enterprise ontology
  - Political/ People and technology requirements

Text Analytics Software
Different Kind of software evaluation

- No single leader - Vendors have different strengths in different environments
- Map output of K Audit to current software offerings
- Select 1-2 for a pilot/POC
- POC use cases – basic features needed for initial projects
- 2-4 week POC – 2+ rounds of develop, test, refine / Not OOB
- Majority of time is on auto-categorization
- False Model – all you need is our software and your SME's
  - Categorization is not a skill that SME's have
  - Rule Building is more esoteric – part library science, part business analysis, part cognitive science

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
POC and Early Development: Risks and Issues

- IT Problem - This is not a regular software process
  - IT favors fully automatic – poorer results
- Semantics is messy not just complex
  - 30% accuracy isn't 30% done – could be 90%
- Variability of human categorization
- Categorization is iterative, not "the program works"
  - Need realistic budget and flexible project plan
- Not enough or bad content – need good example documents – ML or semantic rules
- Anyone can do categorization

Text Analytics Workshop
Benefits: Selling the Vision – Mini-POC

- One week – build categorization rules to 90% accuracy
  - 10 concepts, 20 documents each, simple content model
- Something that people can see, touch, play with
- Real application with real content
- See the value of Taxonomy + Text Analytics
- Appeal to all audiences – Taxonomists to KM to technology geeks to executives
- Start of building a foundation for full enterprise
  - Full POC can build (most of) that foundation

Text Analytics Workshop
Where in the organization?

- Text Analytics impacts all aspects, departments in an organization
  - Needs input from all departments
- Text Analytics requires both IT and Language skills
  - Computational Linguistics
- IT – often the default – budget and software expertise
- KM or Marketing – business focus, business language
- Ideal – library - if it exists (rare).
- Text Analytics requires inter-department cooperation
  - Often requires extra-organization resources

Text Analytics Forum (TAF) Workshop
Select & Build Enterprise LLM

- LLM Foundation:
  - LLM part of text analytics foundation
  - Define Use Cases and Requirements – K Audit
    - Open Source vs. Private
  - Select LLM – Cloud or specialized SLM
  - Fine Tune LLM, merge public and private LLMs
  - Security and Compliance – ingest documents, test output
  - Cost and Resource Requirements
  - Scalability and Performance
  - POC/Pilot

Text Analytics Forum (TAF) Workshop
Fine Tune LLM / Merge LLM and Enterprise Data

- Fine-Tuning – basic steps
  - Data Collection – mostly automated
  - Preprocessing – Hybrid – human and TA
  - Training - see below
  - Evaluation – multiple, SME & auto
  - Deployment
- Multiple Approaches
  - Hybrid Models – RAG + data
  - LLM Orchestration-frameworks
  - API Integration
  - Data Augmentation
  - Many more

Text Analytics Workshop

- Text Analytics and Prompt Engineering

Text Analytics Forum (TAF) Workshop
Prompt Engineering

- Iteratively deriving prompt – 3 parts
  - Context – Task – Output format
  - Personas – detailed description – suitable for X
  - Varieties of prompts – Lance Eliot – 50 types
  - NEW – meta-prompts – Open AI – software improves prompt before submitting
  - Meta – model checks accuracy of other models
- Text Analytics Prompt
  - Taxonomy/Knowledge Graph – incorporate into context
  - Apply auto-categorization, tags into context
  - Iterate through taxonomy – top down

Text Analytics Forum (TAF) Workshop
Prompt Engineering – Example Types

- **Persistent Context –** prepend to all/some prompts
- **Multi-Persona Prompting** – you are Lincoln meeting Gandi
- **Chain-of-Thought (CoT) Prompting -** telling generative AI to proceed in a stepwise fashion
- **Retrieval-Augmented Generation (RAG) Prompting –** combine data/search results with prompt
- **Chain-of-Thought Factored Decomposition Prompting -** prod the generative AI to generate a series of sub-questions and sub-answers
- **Skeleton-of-Thought (SoT) Prompting –** produce an outline, itself or with CoT
- **Self-Reflection Prompting –** think about your answer
- 43 other types

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
Conclusions

- Text analytics and Gen AI – mutual enrichment
  - AI needs concepts, new kinds of structure
- AI for categorization not ready for prime time
  - Great for data and patterns, emerging trends
- Text analytics turns “unstructured” content into data
- Categorization is the brains of the outfit
  - Smart applications, makes everything else smarter
- Text analytics and Gen AI best approached as infrastructure – platform for multiple applications
- Future = multiple integrations of methods, applications

Questions?

Tom Reamy
tomr@kapsgroup.com
KAPS Group
Knowledge Architecture Professional Services
http://www.kapsgroup.com

Text Analytics Workshop
Additional Reading

- What is Smarter and Safer than Chat GPT? - KAPS Group
- There is No Such Thing as Unstructured Text - KAPS Group
- Enterprise AI's Weak Link - KAPS Group
- Lessons from Chess for Gen AI - KAPS Group
- Benefits of Text Analytics for Data-Driven Insights and AI Initiatives (progress.com)