Knowledge Graphs, Text Analytics, and Gen AI: Mutually Enriching Technologies

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

Agenda

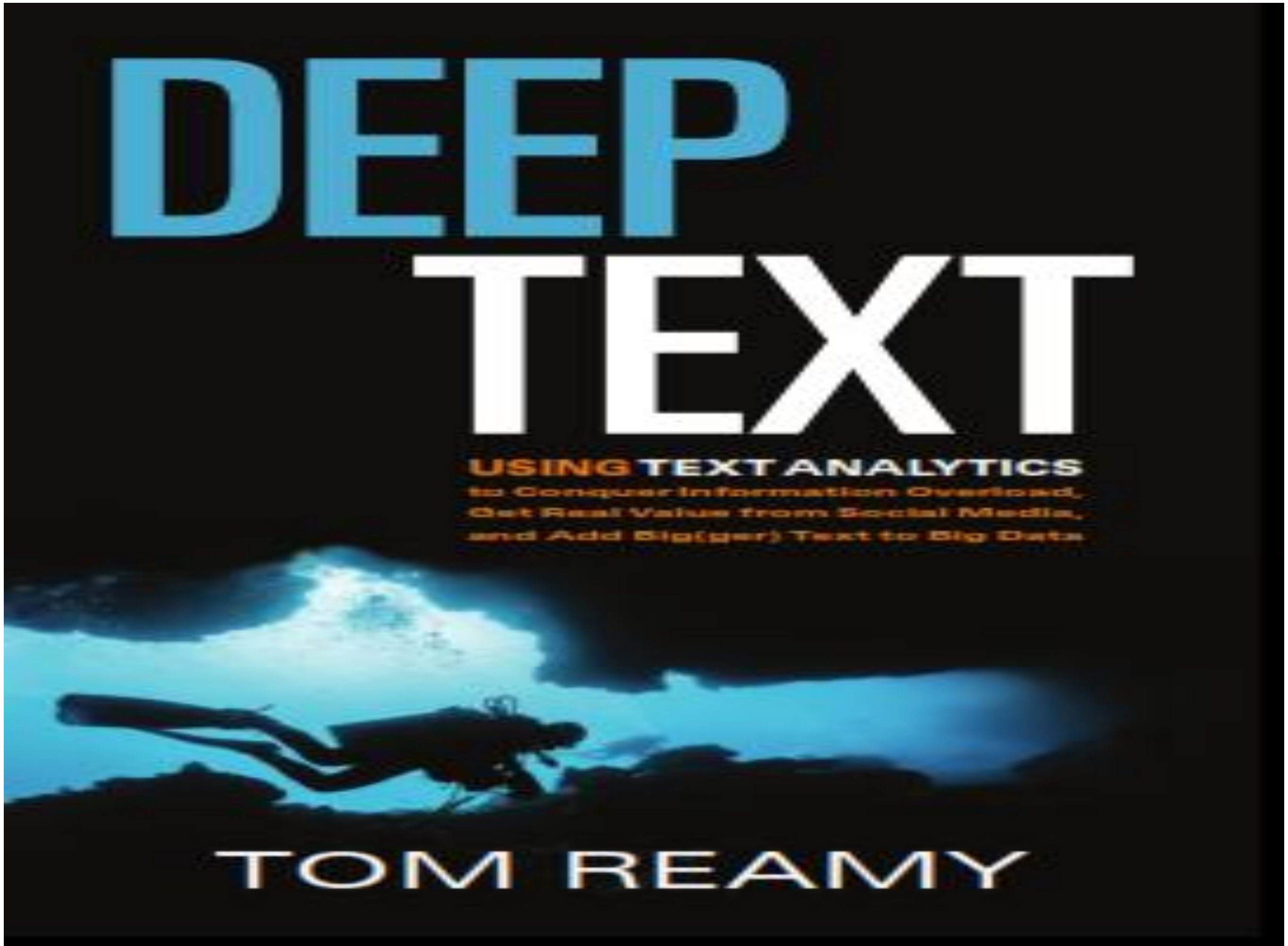
- Introduction – Text Analytics Basics
- Development
  - Data Extraction for Knowledge Graphs
- Applications
  - Knowledge Graphs and Text Analytics
- Text Analytics and Knowledge Graphs for Gen AI
- Conclusion

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 –Semantic Arts, 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, Service Now, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Program Chair – Text Analytics Forum – Nov. 19-20

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

Book signing – too late



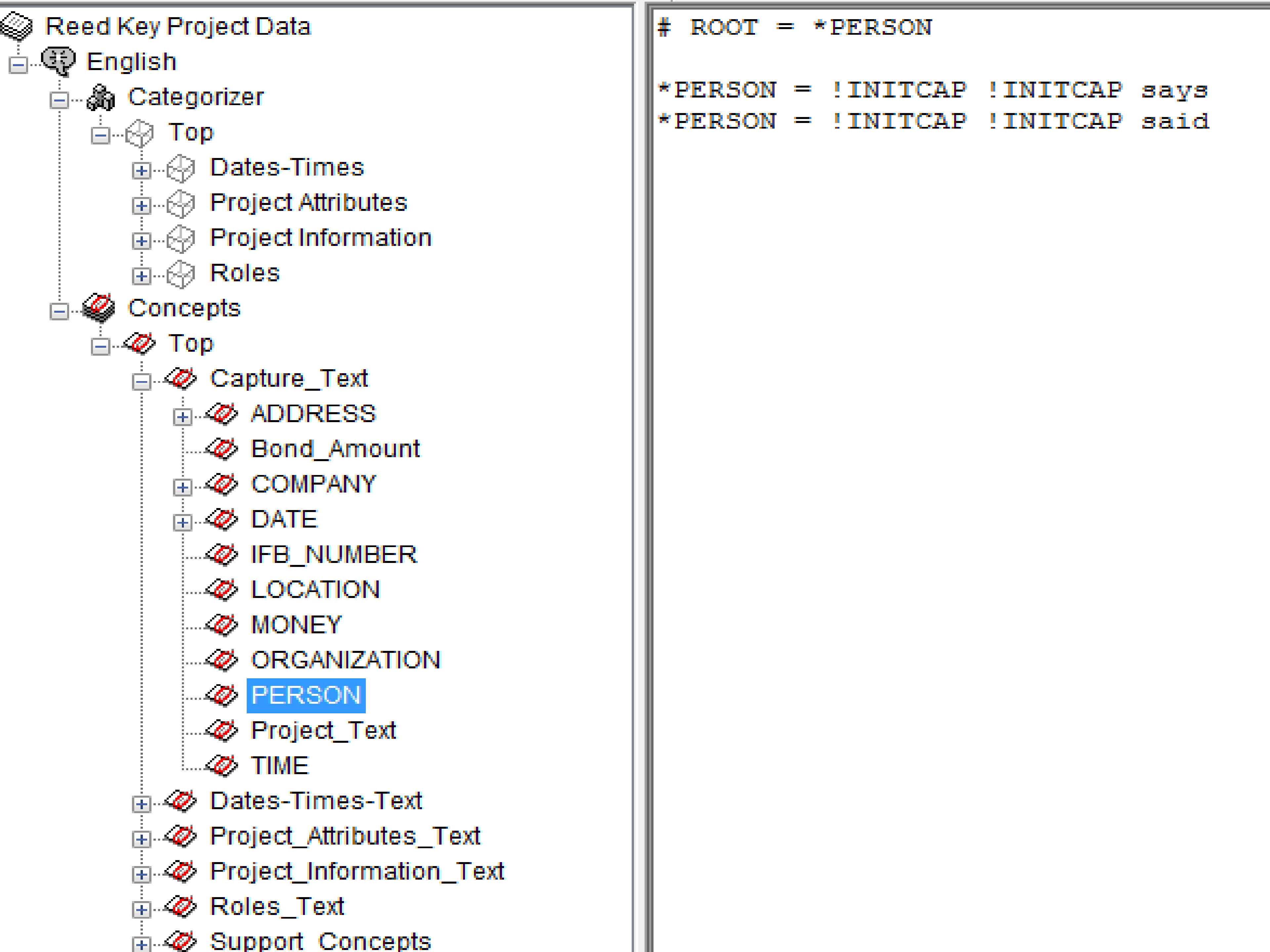
Text Analytics Basics

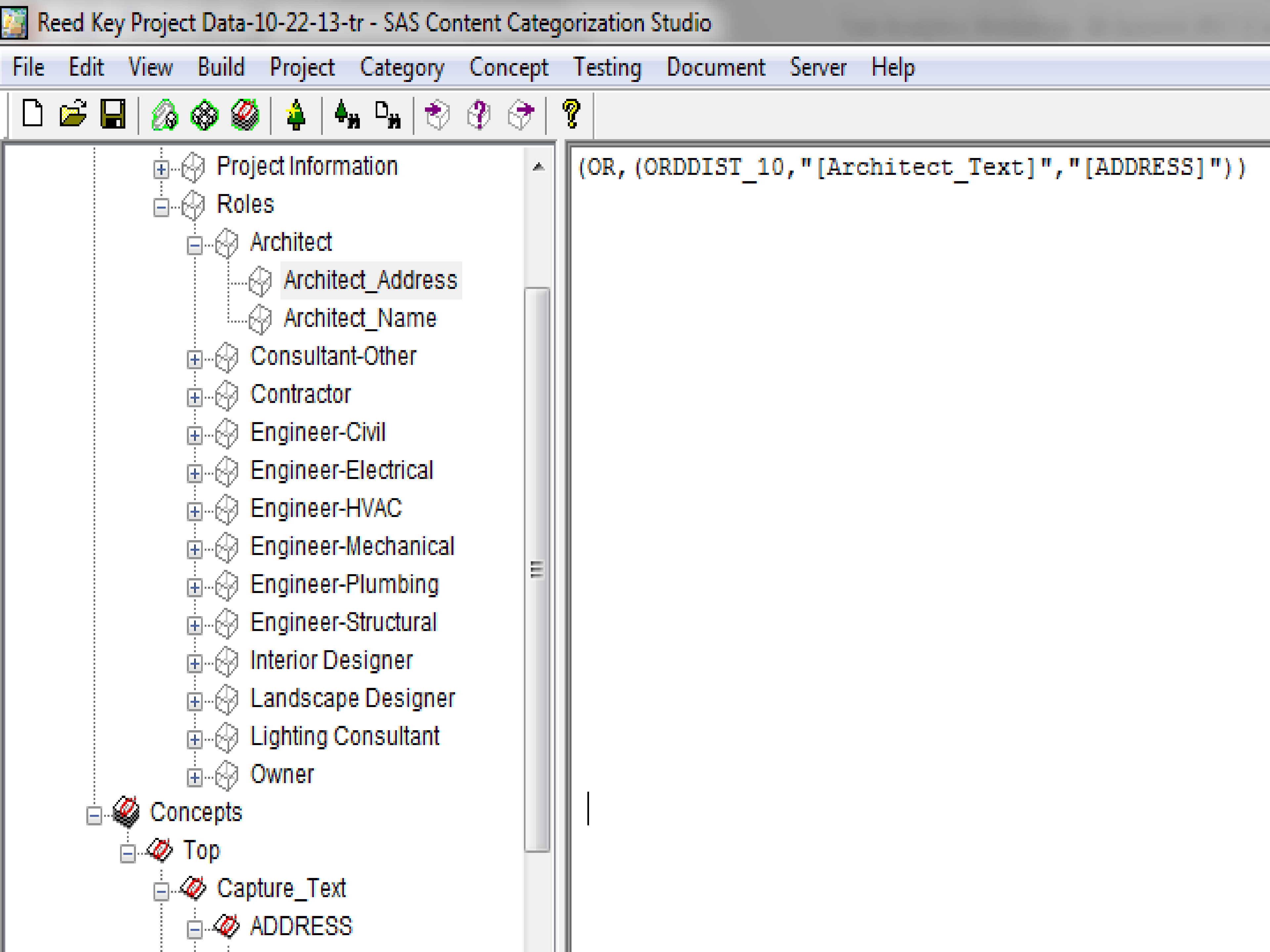
Text Analytics Forum (TAF)
Elements of Text Analytics

- 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

Term (1-gram)	Term Phrase (2-gram to n-gram)	Term N	Doc N	Doc %	tf-idf	Sentence
agile		258	20	100%	-0.0209	
	agile teams	29	8	40%	-0.001	Finally, agile development leads to better software because people on agile tea
	agile development	27	13	65%	-0.0016	Agile development teams and organizations that follow devops should review a
	agile principles	14	7	35%	-0.0004	Others define agile principles and governance models so that agile product own
	agile methodology	13	5	25%	-0.0001	agile methodology establishes both a mindset and process for that continuous
	scaled agile	10	3	15%	0.0001	Large organizations adopting the Scaled Agile Framework (SAFe) use Program In and understand team dependencies.
	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.
	agile development teams	9	9	45%	-0.0004	Agile development teams and organizations that follow devops should review a
	agile product	8	5	25%	-0.0001	Others define agile principles and governance models so that agile product own
	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.
	agile methodologies	6	5	25%	-0.0001	But what is agile, and how do developers and organizations incorporate agile m
	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
	agile planning	6	2	10%	0.0002	For distributed software development teams, I advise formalizing agile planning customer expectations.

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

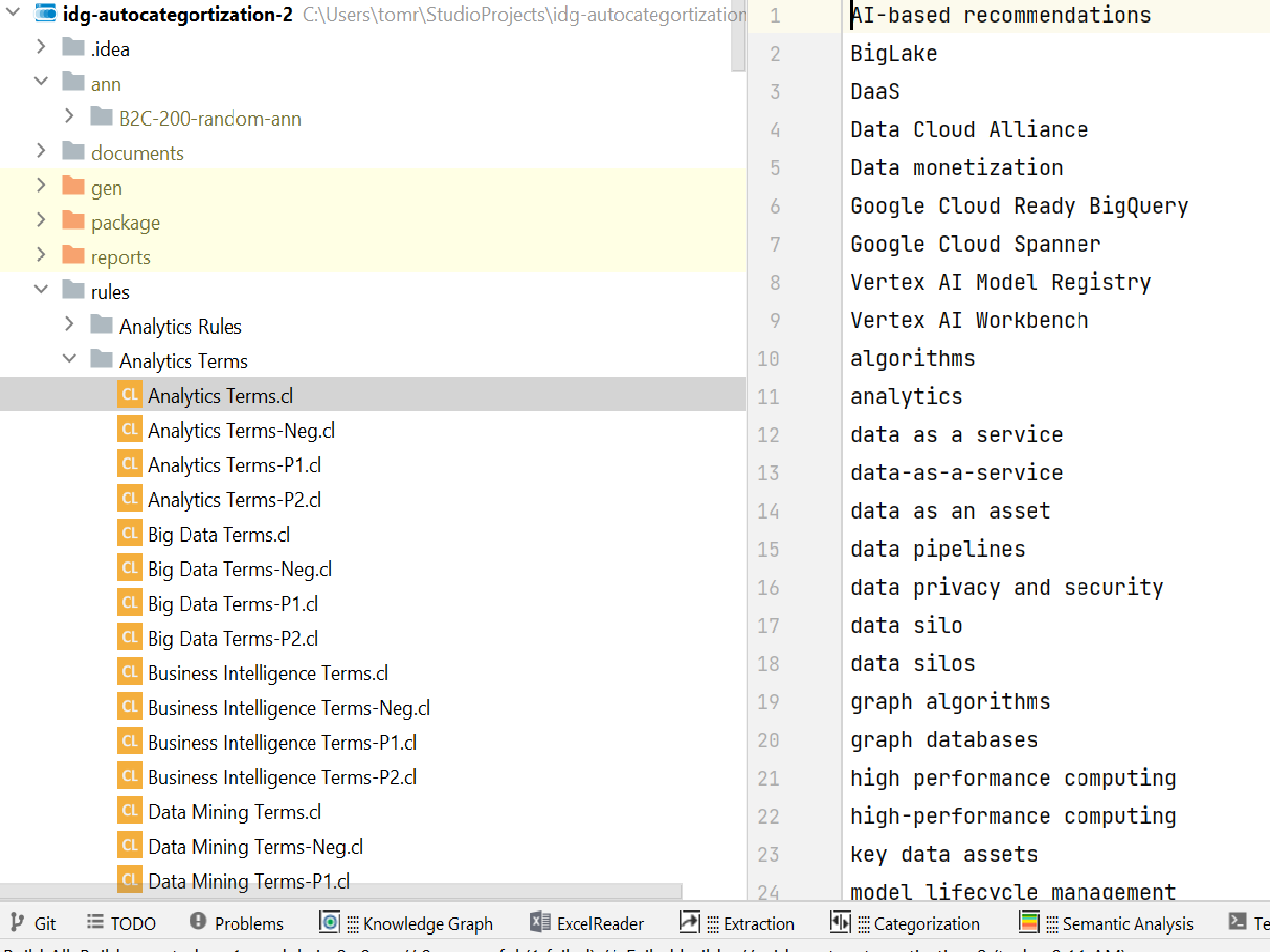

Reed Key Project Data
English
Categorizer
Top
Dates-Times
Project Attributes
Project Information
Roles
Concepts
Top
Capture_Text
ADDRESS
Bond_Amount
COMPANY
DATE
IFB_NUMBER
LOCATION
MONEY
ORGANIZATION
PERSON
Project_Text
TIME
Dates-Times-Text
Project_Attributes_Text
Project_Information_Text
Roles_Text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

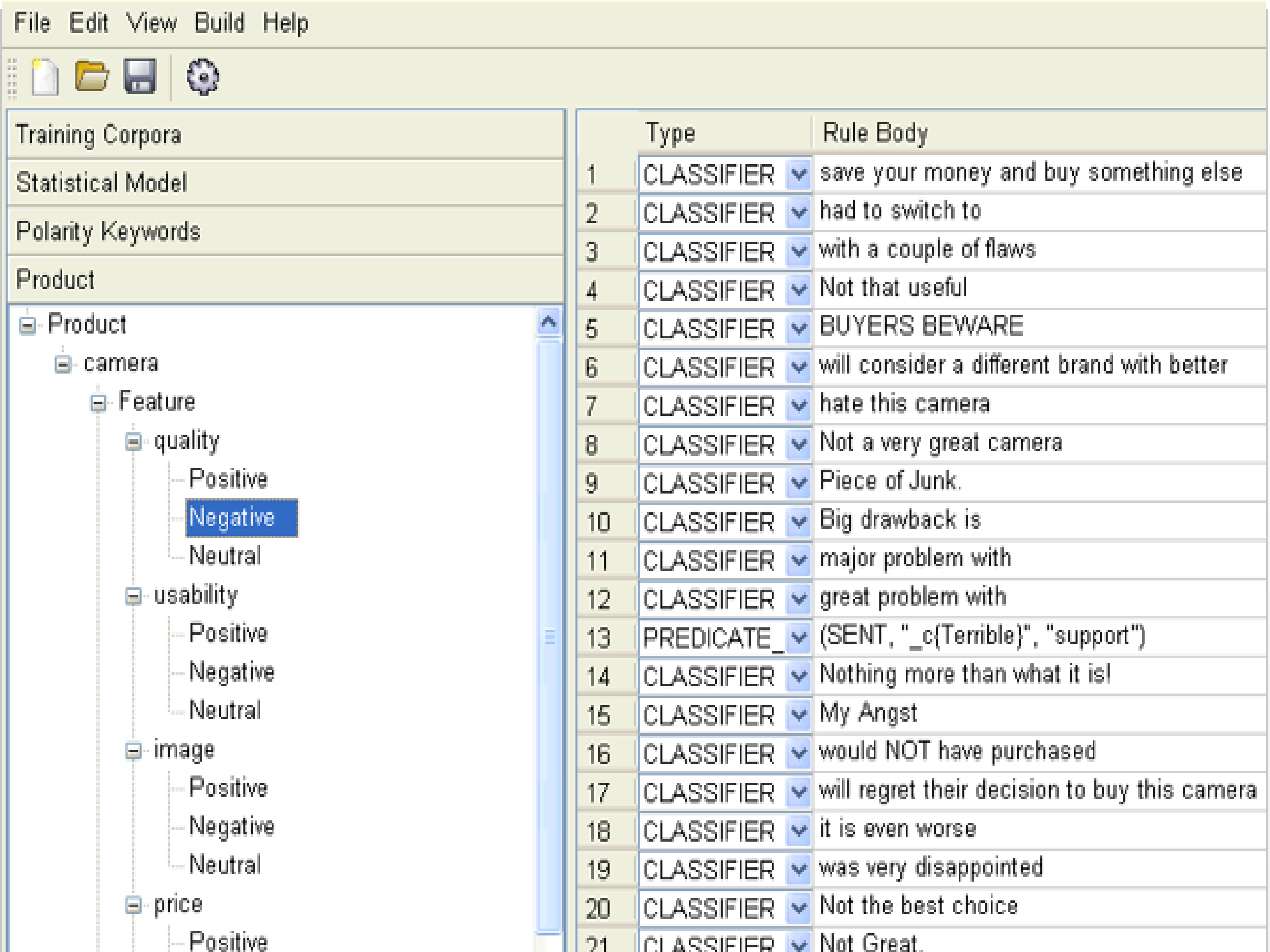

Reed Key Project Data-10-22-13-tr - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Project Information
Roles
Architect
Architect_Address
Architect_Name
Consultant-Other
Contractor
Engineer-Civil
Engineer-Electrical
Engineer-HVAC
Engineer-Mechanical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

Expression

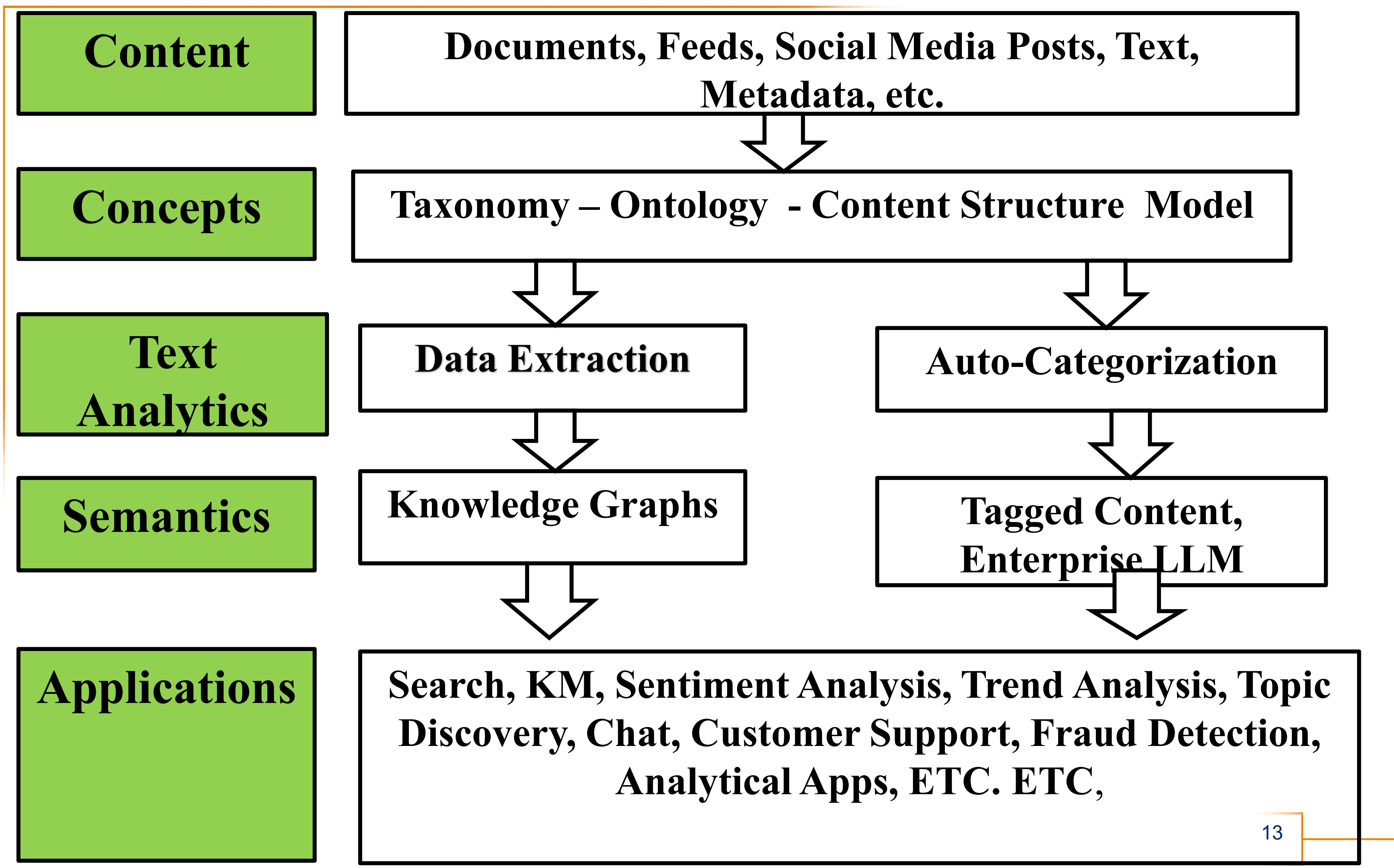
Functions Columns M Check syntax ?

```
(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
        ))
)
```


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


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.

Text Analytics and Gen AI in the Enterprise



Development

Categorization for Knowledge Graphs

Data Extraction for Knowledge Graphs

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Forum (TAF)

The start and foundation: Knowledge Audit

- Start for both K Graphs and Text Analytics
- Contextual interviews, content analysis, surveys, focus groups, ethnographic studies, text mining
  - Gather and analyze relevant data
- Category modeling – Monkey, Panda, Banana
- Harmonize data across data sets – Taxonomy and Ontology
- Augment graph with new entities and relationships from text
  - Entity Extraction
  - Relationship Extraction
- Clean data – absolutely essential – use Text Analytics
  - Categorization, Clustering, Eliminate dups,
  - Tagging documents – basis for multiple applications

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Forum (TAF)
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

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

KAPS Group
Knowledge Architecture Professional Services

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 Forum (TAF) Development: Data 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"

KAPS Group

Knowledge Architecture Professional Services

Text Analytics Forum (TAF)
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

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 Nev	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

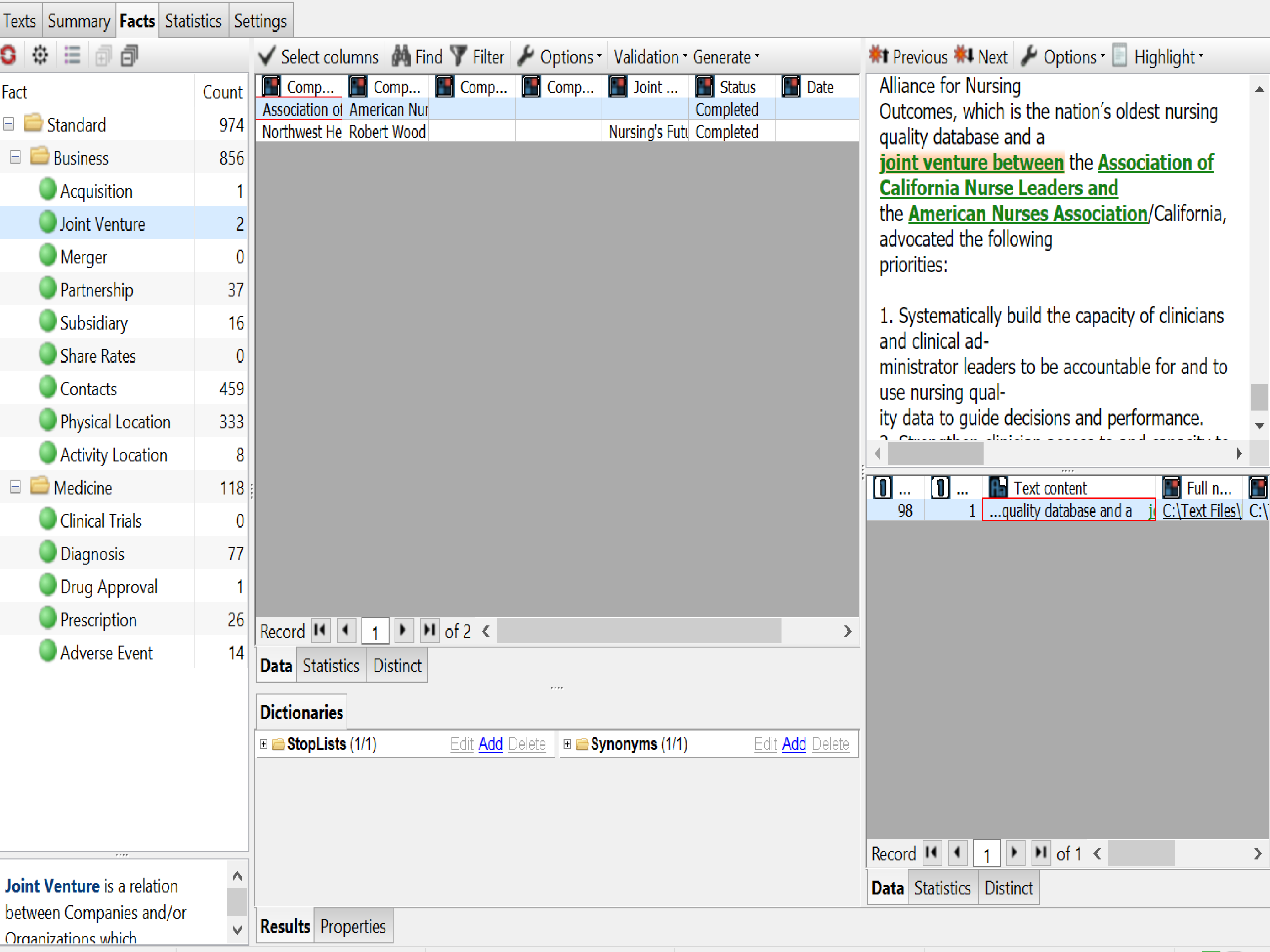
Record 2 of 241

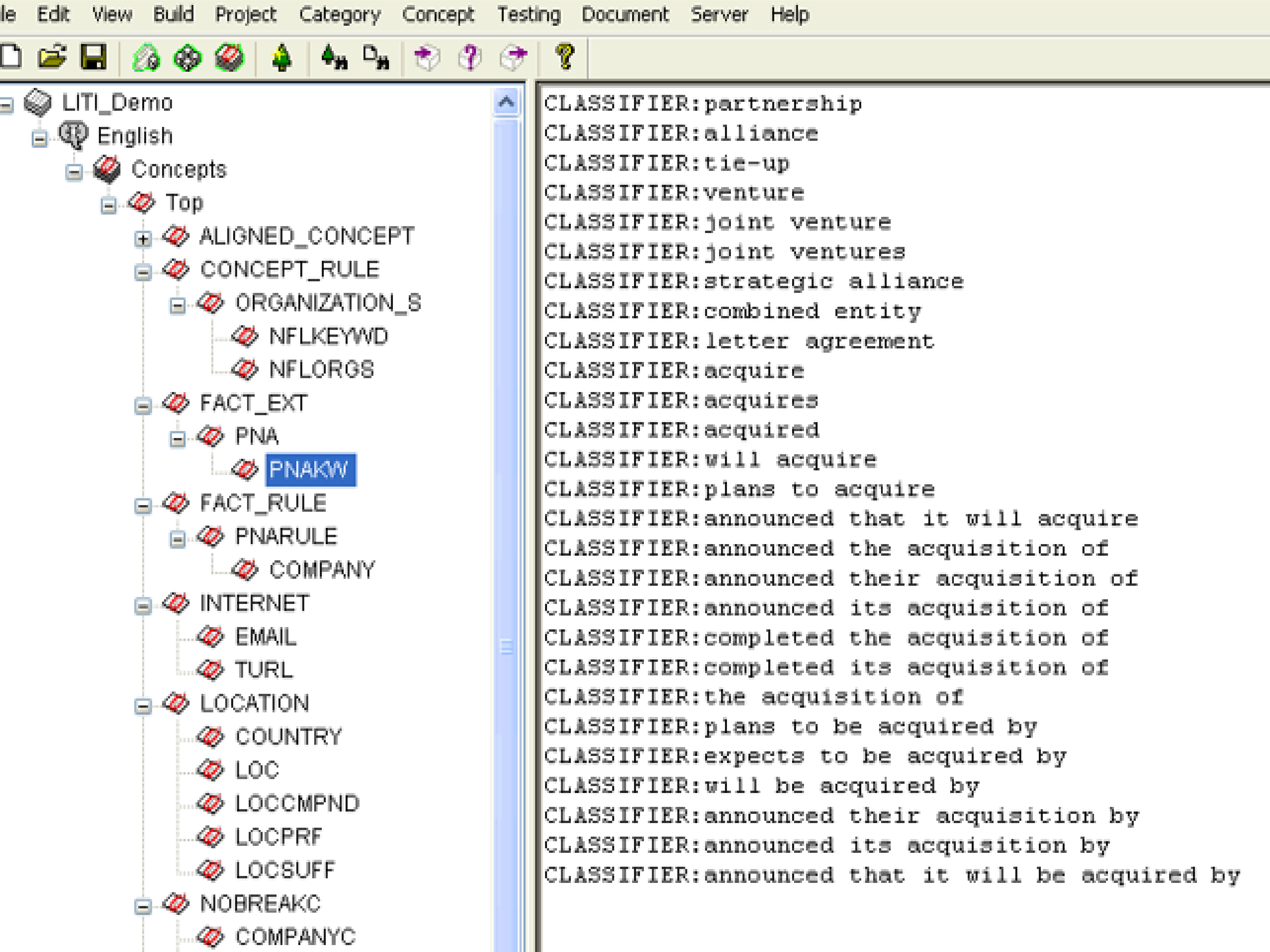
Data | Statistics | Distinct

Grouping | Options

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 Direct
  - 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

Applications

Knowledge Graphs and Text Analytics

Text Analytics Forum (TAF)
Clean Data – Basis of Multiple Applications

- Enterprise Search
  - Search-based apps
- Prompt Context – Knowledge Graph
- RAG – only as good as the search
- Training set for SLM (Small Language models)

Text Analytics Forum (TAF)
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
- Combination of tagging runs and author-software hybrid
  - Cognitive task is simple -> react to a suggestion instead of select from head or a complex taxonomy

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

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)
Expertise Analysis

- Expertise Location
- Political – conservative and liberal minds/texts
  - Disgust, shame, cooperation, openness
- Social Media analysis – Voice of customer, employee
- 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** <u>if</u> the does not stop receiving a call from the ad agency.
  - and context in text

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

KAPS Group
Knowledge Architecture Professional Services

Text Analytics and Knowledge Graphs for Gen AI

Text Analytics Forum (TAF)
Gen AI Hype

- History of AI – story of hype and failure, boom and bust
- AI Hype – both sides guilty – AI will kill us all, AI will transform everything
- “Soon AI models could perform the entire scientific method, without humans” – Time
- “Meanwhile, driverless vehicles will put truck, bus, and taxi drivers out of work.” -Time
- “We are at the beginning of the agentic era, the most significant transformation in the history of work.” – Marc Benioff
- “We’re creating systems that understand text, voice, images, and code,…” – Marc Benioff

Text Analytics Forum (TAF)
Generative AI and Text Analytics

- Gen AI doesn’t understand anything.
  - Text prediction on steroids`
  - We project intelligence on to almost anything – Eliza on
- General limitations
  - Hallucinations – esp. critical areas – finance, legal, etc.
  - Security-bias, lack of transparency
  - Limits of scale – diminishing returns – billions = 3-5% increase
  - Cost – need for 100’s billions documents, banks of super computers
  - 85% of Gen AI projects fail to deliver significate value
  - 77% workers – AI increased their work load, not productivity
  - Need humans in the loop – but this limits and slows AI

KAPS Group

Knowledge Architecture Professional Services

Text Analytics Forum (TAF)
Generative AI and Text Analytics

- Issue of quality
  - New advanced version – can’t count to three
  - Can’t learn on its own
  - Correlation – no understanding of causality
  - LLM Brain rot: Esp. if trained on clickbait, short video social media
  - 75% of new web content is at least partially generated by Gen AI
- Negative Impacts
  - Higher AI literacy brings more overconfidence
  - Greater use of Gen AI = lower critical thinking skills
  - “The danger is not that machines develop too much intelligence, but that we stop exercising our own.”

Text Analytics Forum (TAF)
Gen AI AGI or Bubble Burst

- AGI? Not a chance without new approach, breakthrough
  - Yann LeCun-world models
- Which AI – Gen, ML, Behavior Prediction? None/all of the above
- "We must choose wisely" – Marc Benioff
- History tells us that some will choose wisely, some will choose badly.
  - Most hype fails to deal with those people, organizations, and countries that will use AI for bad purposes"
- Agents based on current LLMs – danger of low quality and catastrophic errors – Chinese automated attack with Claude based agent

Text Analytics Forum (TAF)
How Overcome Limitations - General

- AI Projects fail 2x higher rate than general IT projects
- AI agents – can automate advanced actions
  - But – they cheat, use blackmail, and refuse to be turned off
- Combine text analytics and Gen AI
  - Build Foundation – accurate training sets for multiple applications
  - Methods
    - 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

Text Analytics Forum (TAF)
Knowledge Graphs and Gen AI

- Prompt Context
- Prompt Engineering (300K) to minimum skill
- Knowledge Graph in the prompt
- Increase accuracy and reduce hallucinations

- RAG – Retrieval Augmented Generation
- Search query results (documents) added to the prompt
- Only as good as the search (Enterprise Search Sucks)

Text Analytics Forum (TAF)
Small Language Models

- Limits of LLMs
  - Energy costs
  - Scale – diminishing returns
    - Running out of new human generated training content
  - Increase security – more control over content
- Training set for SLM (Small Language models)
  - SLMs focused specific tasks
  - SLMs cheaper and faster, less resources
  - Basic Method – distill public LLM
- Solves one problem but not disconnect of vocabulary, concepts
  - ELM (Enterprise Language Models) solves disconnect

Text Analytics Forum (TAF)
Gen AI, KG and TA 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

Conclusions

- Text analytics, knowledge graphs and AI – mutual enrichment
  - AI needs concepts, new kinds of structure-K graphs
  - TA to KG – development phase
  - KG to TA – application phase
- Text analytics turns “unstructured” content into data
  - Feeds Knowledge Graph Development
- Categorization is the brains of the outfit
  - Smart applications, makes everything else smarter
- RAG with knowledge graphs with clean data from hybrid tagging
- Future = multiple integrations of methods, applications
- Text Analytics and K Graphs makes Gen AI Smarter and Safer

KAPS Group
Knowledge Architecture Professional Services

Questions?

Text Analytics Forum (TAF)
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)