Text Analytics Workshop

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

KAPS Group
Knowledge Architecture Professional Services

KAPS Group

Agenda

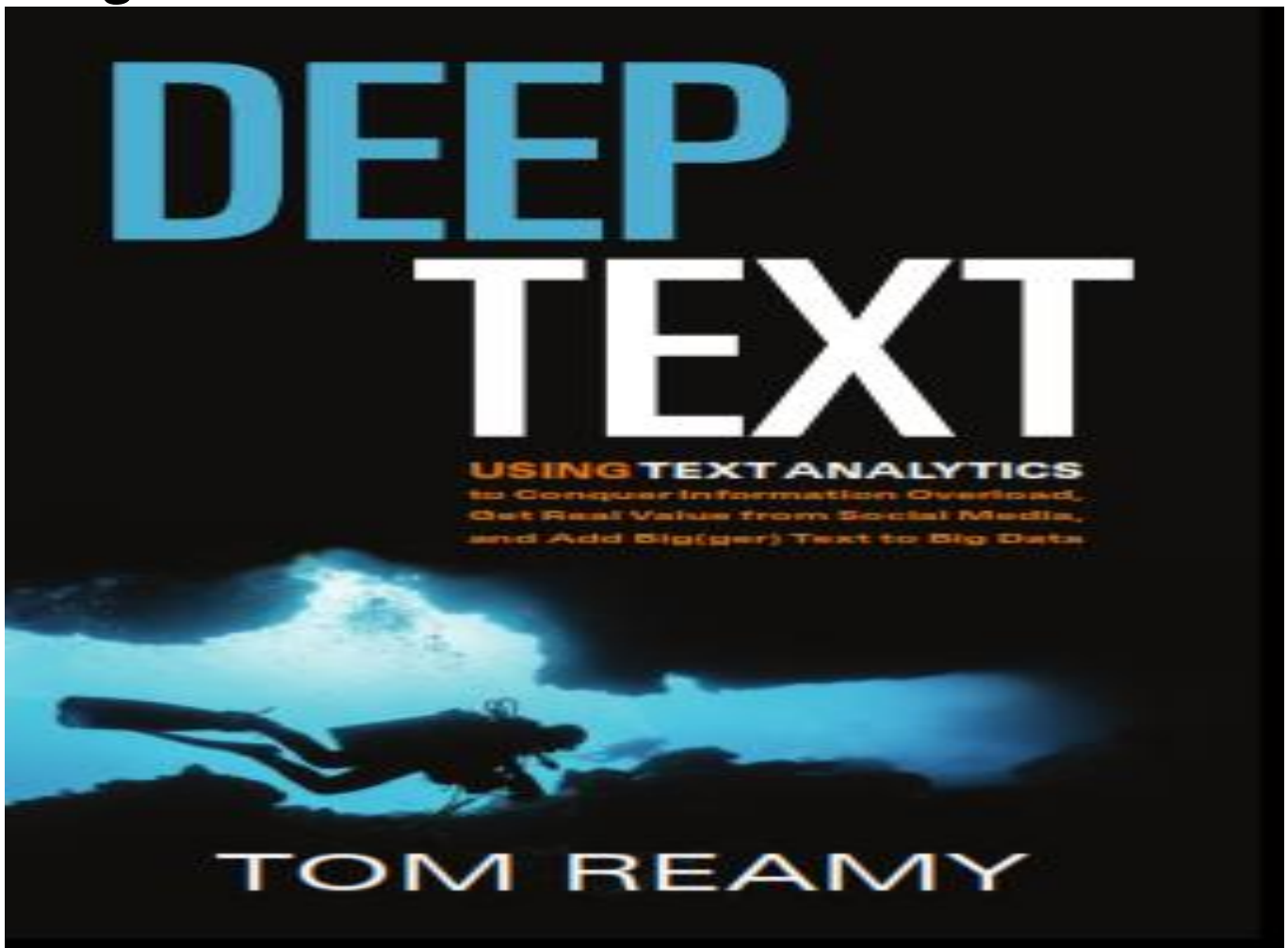
- Introduction – Elements of and State of Text Analytics
- Getting Started with Text Analytics
- Development – Taxonomy, Categorization, Faceted Metadata
  - Mini-POC – Categorization 95% Accuracy
- Text Analytics Applications
- Knowledge Graphs
- Questions / Discussions

Introduction: KAPS Group

- Network of Consultants and Partners - 2002
- Text analytics consulting: Strategy, Development-taxonomy, text analytics foundation & applications
- Mini-Projects – get started or take to next level
  - Strategy, Mini-POC - Categorization
- Partners –Synaptica, SAS, Smart Logic, Expert System, Clarabridge, 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, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Program Chair – Text Analytics Forum – Nov. 17-19 Virtual

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

Book Sign / Meet the Author



Introduction:
Elements of Text Analytics

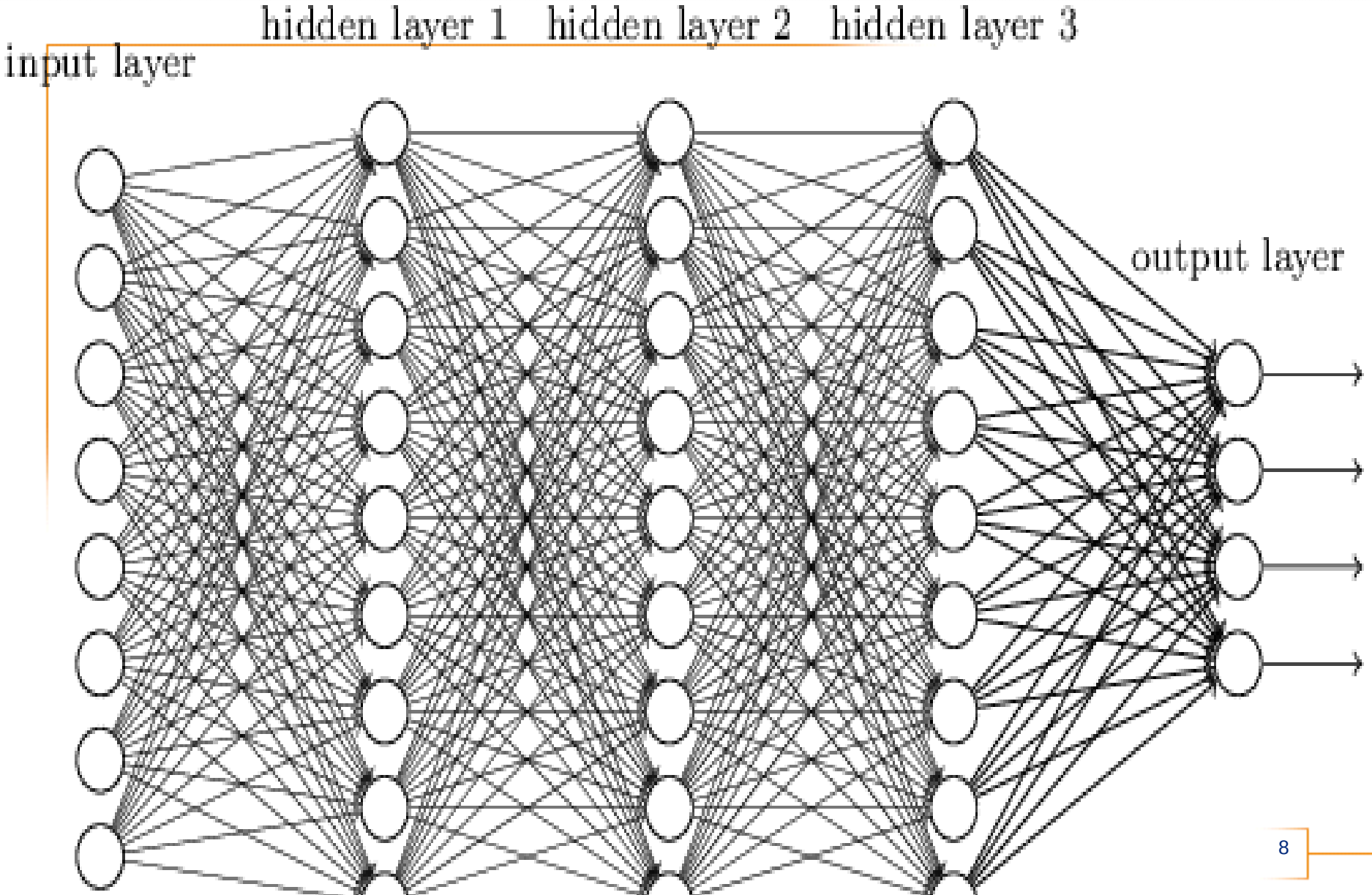
- Text Mining – NLP, statistical, predictive, machine learning
  - Different skills & applications, Math & data not language
- Data Extraction – entities, concepts, events, facts
  - Analytical applications, enhance search - facets
- Sentiment Analysis – positive & negative, attitudes
  - Advanced - racial equality, social media analysis
- Summarization
  - Dynamic – based on a search query term
  - Structured summaries – data & concepts

Introduction:
Elements of Text Analytics

- Auto-categorization
  - Training sets – Bayesian, Vector space
  - Terms – literal strings, stemming, dictionary of related terms
  - Rules – Boolean – AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
  - Templates – rules and content models
- Platform for multiple features – Sentiment, Extraction
  - Disambiguation - Identification of objects, events, context
  - Fact Extraction – context around words, concepts
  - Model more subtle sentiment

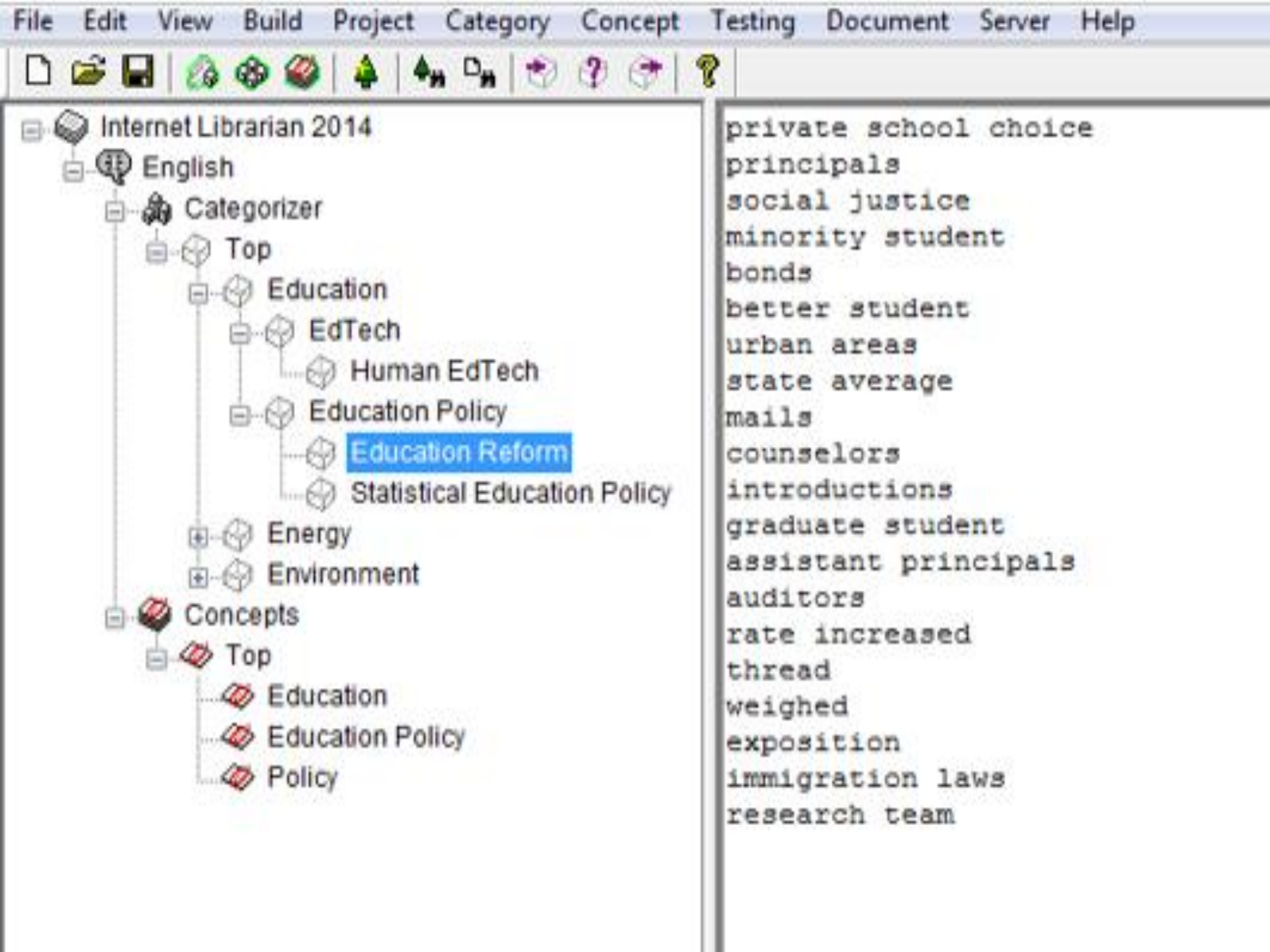
Introduction:
Elements of Text Analytics: Deep Learning

- Neural Networks – from 1980's – bigger & faster
- Strongest in areas like image recognition, pattern
- Weakest – concepts, subjects, deep language, metaphors, etc.
- Black Box – don't know how to improve except indirect manipulation of input
  - Watson – "We don't know how or why it works"
  - Susceptible to bias – hard to fix
- No common sense, rigid – turn object turtle = gun
- Not ready for categorization, OK for data analysis


hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer

D1 | Descriptive Terms

	A	B	C	D
1	#	Percentag	Freq	Descriptive Terms
2	1	34%	766	optimization
3	2	13%	298	+ driver, + device, + mechanism, + layout, + mobile device, + drive force, + lithography, + drive development, hard-drive, + multiprocessor, + fabrication, + parallel, performance analysis, + mobile phone, + hardware platform
4	3	7%	152	+ router, + technology, + memory, + mechanism, + component, hardware, + optimization
5	4	1%	15	dram, + memory, + hardware implementation, + router, hardware, + technology, + component
6	5	15%	344	+ mechanism, + memory, + hardware description language, + hardware optimization, + hardware parameter show, + component, + hardware component, hardware overhead, + keyboard, + hardware system, + drive, + parallel, hardware complexity, performance analysis
7	6	7%	156	+ microprocessor, + pipeline, + firmware, + hardware modification, + hardware trap, hardware-software, device reliability, hardware support, hardware, + hardware implementation, vlsi, + hardware platform, + drive, + drive architecture, + keyboard


File Edit View Build Project Category Concept Testing Document Server Help
Internet Librarian 2014
English
Categorizer
Top
Education
EdTech
Human EdTech
Education Policy
Education Reform
Statistical Education Policy
Energy
Environment
Concepts
Top
Education
Education Policy
Policy
private school choice
principals
social justice
minority student
bonds
better student
urban areas
state average
mails
counselors
introductions
graduate student
assistant principals
auditors
rate increased
thread
weighed
exposition
immigration laws
research team

File Edit View Build Project Category Concept Testing Document Server Help

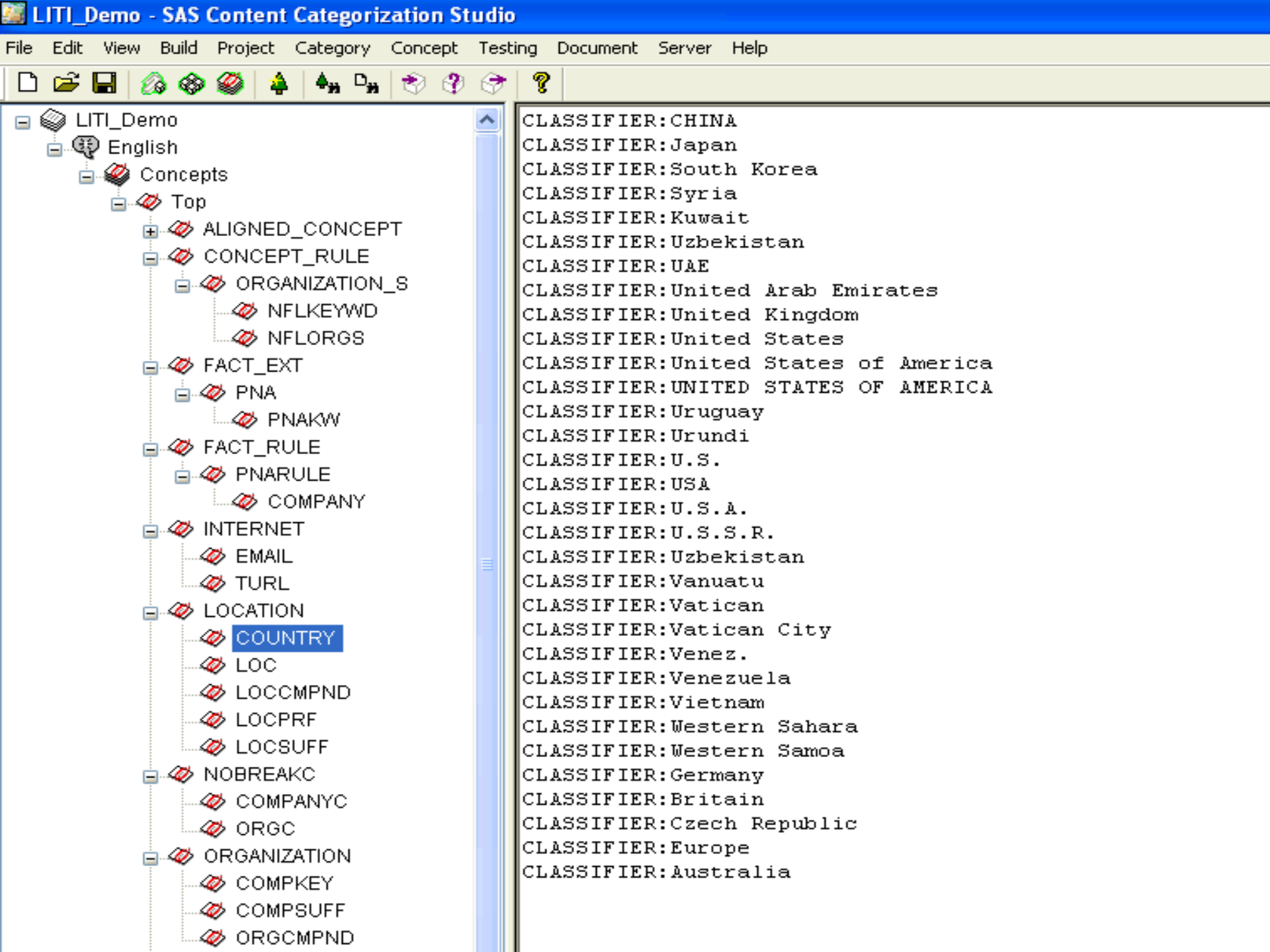
- English
  - Categorizer
    - Top
      - Account Action
      - Bill - Pay
        - CustomerAccountInColl
        - PastDueAmount
        - PayBill
      - Bill Education - Contents
      - Bill Education - Plan
      - Device
        - DeviceActivation
        - DeviceEducation
        - DeviceExchange
        - DeviceLost
        - DeviceResume
      - Device - Service Not Workir
      - Zed - Actions
  - Concepts
    - Top
      - abnormal
      - activate

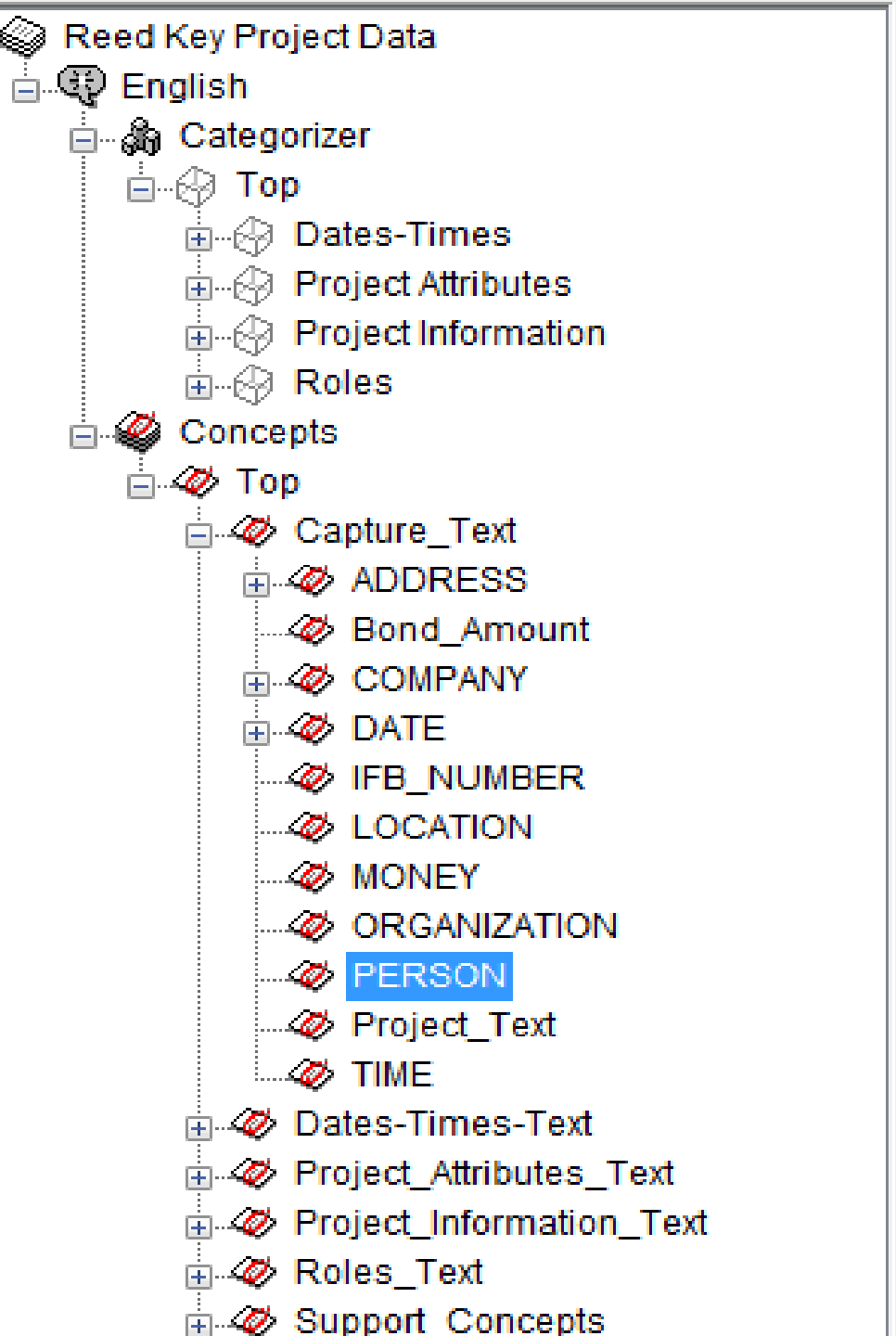
```
(AND,
(OR,
(DIST_5, "[customer]", (AND, "[phone]", "[lost-stolen]")),
(DIST_5, "[called]", (AND, "[phone]", "[lost-stolen]")),
(DIST_5, (AND,"[customer]", "[called]"), "[lost-stolen]")
),
(NOT,
(OR, "[activate]", "[swap]",
(DIST_5, (OR, (OR,"[customer]","[called]"), "[lost-stolen]"), "[restrict]"))
)
)
```

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

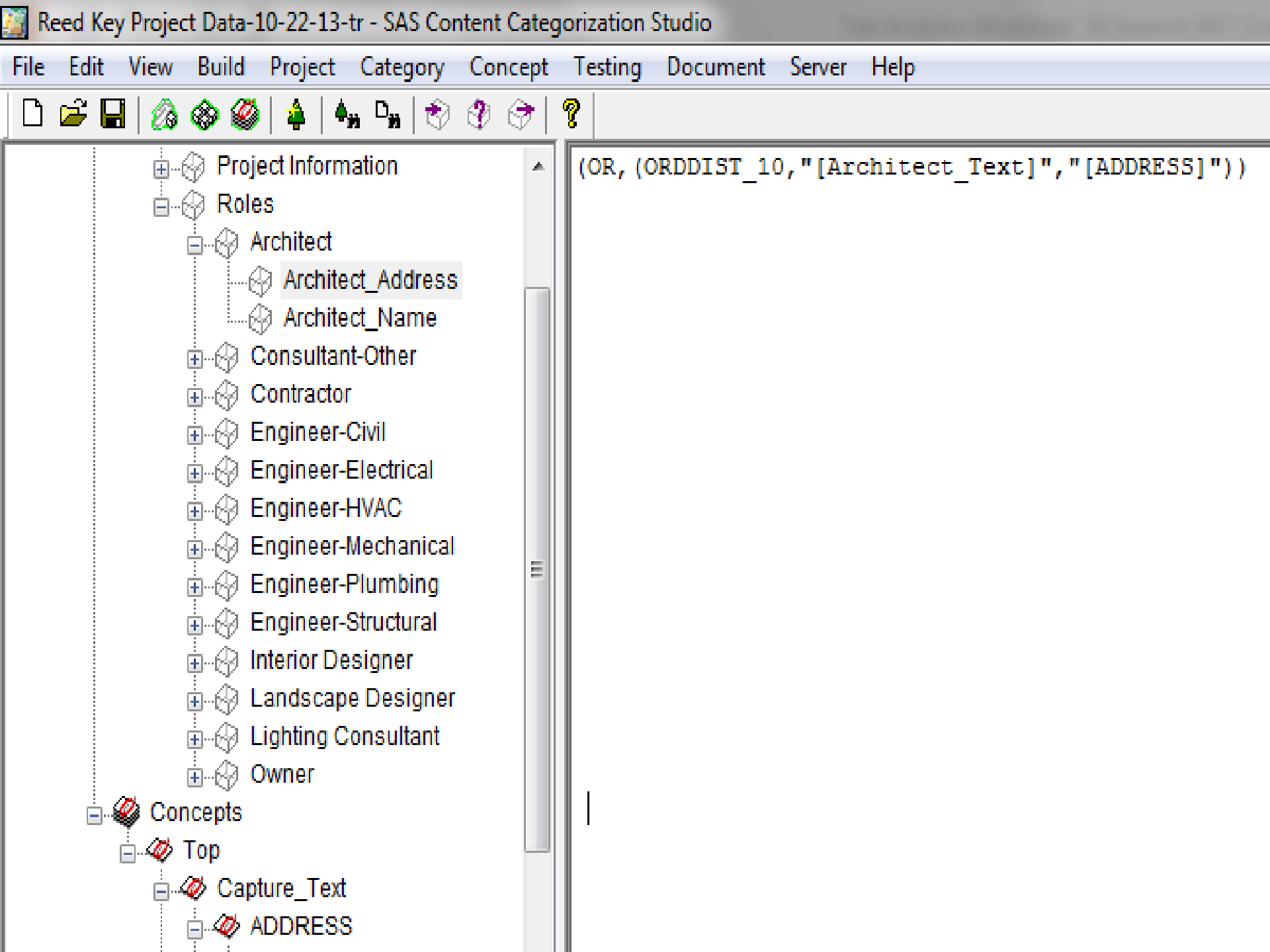
cust,
custeomer,
custeomr,
custeorm,
custimer,
custiomer,
custir,
custm,
custmer,
custmoer,
custmomer,
custmr,
custoemer,
custoemr,
custoemrs,
custoer,
custoerm,
custome,
customeer,
customer,
customera,
customerr,
customers,
customner,
customr,
customre,
customwer,
cu,
cx,
cst,

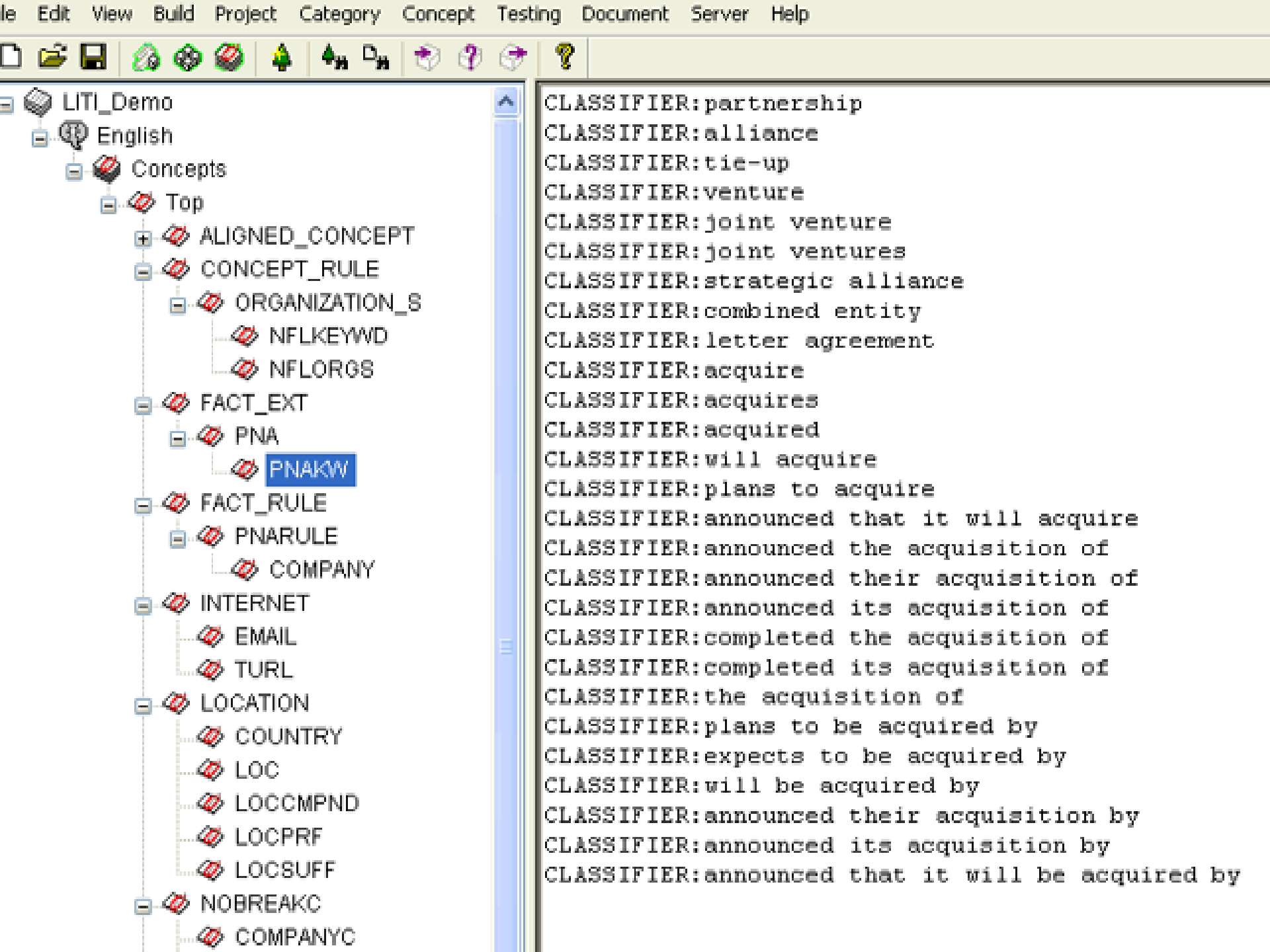

LITI_Demo - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
LITI_Demo
English
Concepts
Top
ALIGNED_CONCEPT
CONCEPT_RULE
ORGANIZATION_S
NFLKEYWD
NFLORGS
FACT_EXT
PNA
PNAKW
FACT_RULE
PNARULE
COMPANY
INTERNET
EMAIL
TURL
LOCATION
COUNTRY
LOC
LOCCMPND
LOCPRF
LOCSUFF
NOBREAKC
COMPANYC
ORGC
ORGANIZATION
COMPKEY
COMPSUFF
ORGCMPND
CLASSIFIER:CHINA
CLASSIFIER:Japan
CLASSIFIER:South Korea
CLASSIFIER:Syria
CLASSIFIER:Kuwait
CLASSIFIER:Uzbekistan
CLASSIFIER:UAE
CLASSIFIER:United Arab Emirates
CLASSIFIER:United Kingdom
CLASSIFIER:United States
CLASSIFIER:United States of America
CLASSIFIER:UNITED STATES OF AMERICA
CLASSIFIER:Uruguay
CLASSIFIER:Urundi
CLASSIFIER:U.S.
CLASSIFIER:USA
CLASSIFIER:U.S.A.
CLASSIFIER:U.S.S.R.
CLASSIFIER:Uzbekistan
CLASSIFIER:Vanuatu
CLASSIFIER:Vatican
CLASSIFIER:Vatican City
CLASSIFIER:Venez.
CLASSIFIER:Venezuela
CLASSIFIER:Vietnam
CLASSIFIER:Western Sahara
CLASSIFIER:Western Samoa
CLASSIFIER:Germany
CLASSIFIER:Britain
CLASSIFIER:Czech Republic
CLASSIFIER:Europe
CLASSIFIER:Australia

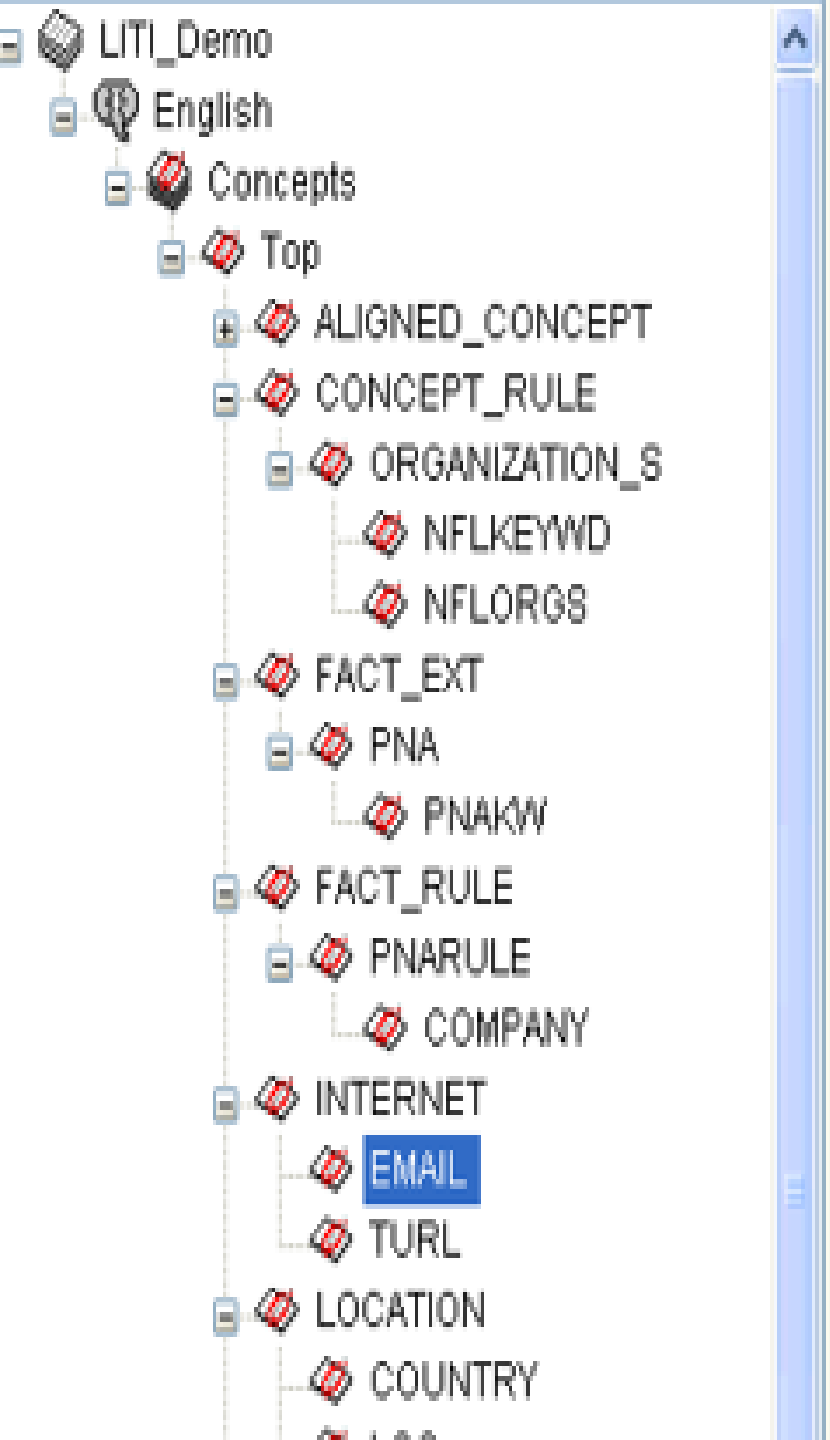



```
# ROOT = *PERSON

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

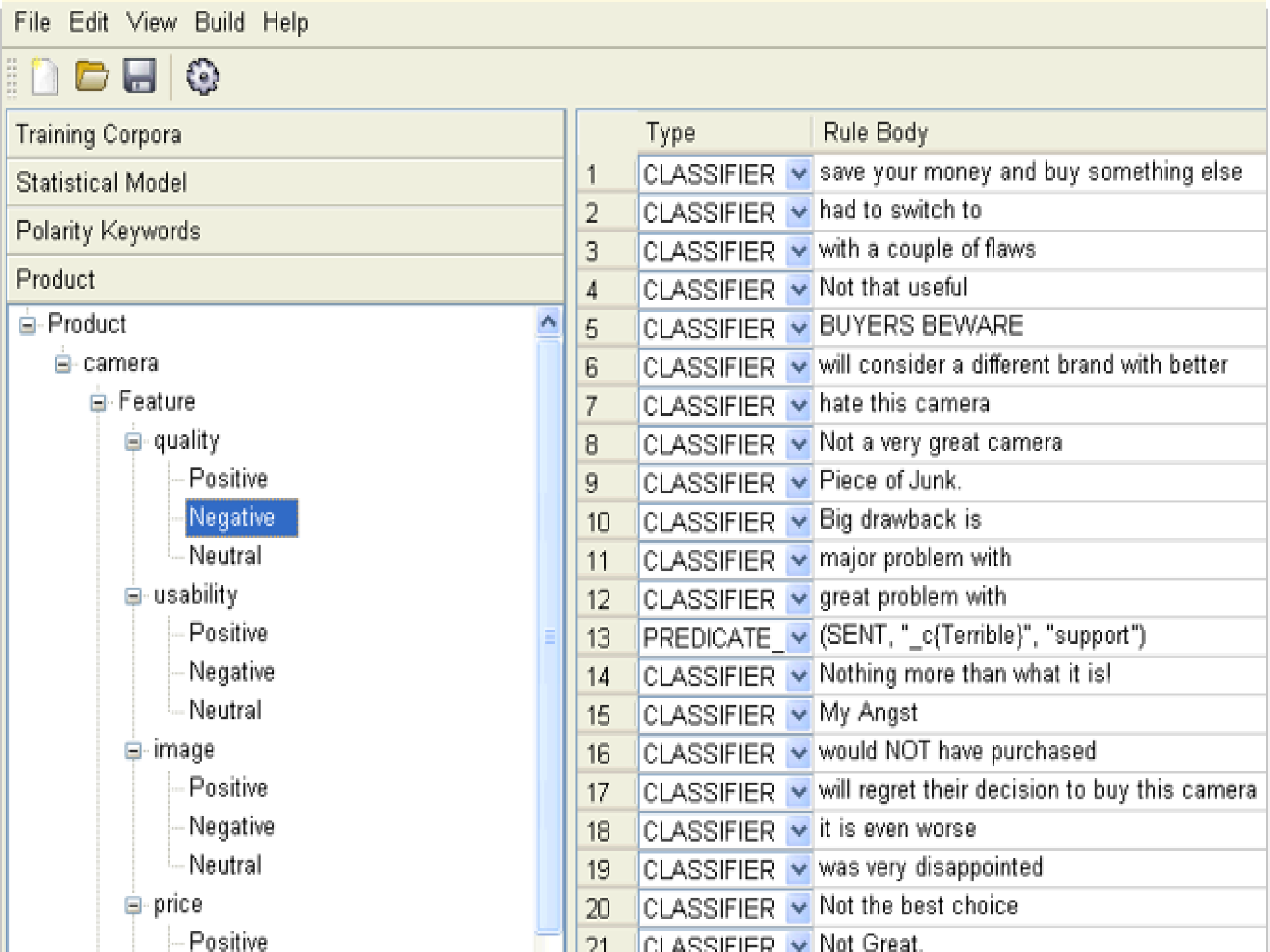

Reed Key Project Data-10-22-13-tr - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Project Information
Roles
Architect
Architect_Address
Architect_Name
Consultant-Other
Contractor
Engineer-Civil
Engineer-Electrical
Engineer-HVAC
Engineer-Mechanical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


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




```
# 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.

Text Analytics Workshop
Field of Text Analytics

- History – academic research, focus on NLP
- Inxight –out of Zerox Parc
  - Moved TA from academic and NLP to auto-categorization, entity extraction, and Search-Meta Data
- Explosion of companies – many based on Inxight extraction with some analytical-visualization front ends
  - Half from 2012 are gone - Lucky ones got bought
- Initial Focus on enterprise text analytics
- Shift to sentiment analysis - easier to do, obvious pay off (customers, not employees)
  - Backlash – Real business value?
- Current – Multiple Applications

Text Analytics Workshop
Introduction: Text Analytics

- Current Market: 2018 – exceed $1 Bil for text analytics (10% of total Analytics)
- Growing 20% a year, search is 33% of total market
- Fragmented market place – full platform, social media, open source, taxonomy management, extraction & analytics, embedded in applications (BI, etc.), CM, Search
- No clear leader.
- Deep Text
  - Linguistic and Cognitive Depth – human-like learning
  - Integration of multiple techniques
  - Infrastructure – Move fast with a stable infrastructure
- AI-Deep Learning – still “Two Years Away”

Text Analytics Workshop
Current State of Text Analytics: Trends

- Market - 3.5 B to 10.5B 2023
- Cloud Technology Growing
  - Real time analytics, text from anywhere
- BOTs in the enterprise
- Conversational Interfaces
- Social – Behavioral Analytics
  - New models of people
  - New Knowledge Organizations
    - K Graphs
    - Emotion, Motivation Taxonomies
- AI and ML – hype? Or ?

Text Analytics Workshop
Current State of Text Analytics: Vendor Space

- Taxonomy Management Plus
- Extraction and Analytics
  - Multiple Dedicated Applications – BI, CI, social media
- Sentiment Analysis
- Open Source, build your own – API's
- Embedded in Content Management, Search, BI, C, etc.
- Full text analytics platforms
- Option: Cloud versions – reduced technical requirements
  - Issue – Security Protocols – external copies of data

Text Analytics Workshop
Benefits of Text Analytics

- What is the ROI of text analytics?
  - Wrong question?
  - What is ROI of organizing your company
- Benefits in 3 areas:
  - Search, Social Media, Multiple Info Apps
- Start with numerical studies
- Stories – Pharma example
- Stories – find own real life stories
- Selling to C Level
  - Different language
  - Need to educate – what it is and why

Text Analytics Workshop
Costs and Benefits - Search

- IDC study – quantify cost of bad search
- Three areas:
  - Time spent searching
  - Recreation of documents
  - Bad decisions / poor quality work
- Costs
  - 50% search time is bad search = $2,500 year per person
  - Recreation of documents = $5,000 year per person
  - Bad quality (harder) = $15,000 year per person
- Per 1,000 people = $ 22.5 million a year
  - 30% improvement = $6.75 million a year
  - Add own stories – especially cost of bad information

Text Analytics Workshop
Benefits – Why Isn't Everyone Doing It?

- Don't know enough about text analytics
- Financial Constraints – too expensive
- Lack of senior management buy-in
- Lack of clarity about value of text analytics
  - Don't believe ROI numbers
- Don't have the necessary in-house expertise
- Other priorities are more important

- Overall: Lack of knowledge and expertise

Text Analytics Workshop
Benefits: Selling the Vision

- All of that is a complex sell – how to do it?
- New Approach – Mini-POC
  - One week
- Elements
  - Taxonomy (Old, one branch) – 10-20 nodes to 100
  - Sample content – 20 documents per node
  - Simple content model – document sections
- Build categorization rules for all nodes
- Demo – Simple search (15%-50%) to 90%+

Benefits: Selling the Vision
Mini-POC

- Something that people can see, touch, play with
- Real application with real content
- See the value of Taxonomy + Text Analytics
- Appeal to all audiences – Librarians to KM to technology geeks to executives
- Option – Comparison with fully automatic clusters
- Start of building a foundation for full enterprise
  - Full POC can build (most of) that foundation

Text Analytics Workshop

- Getting Started with Text Analytics

Text Analytics Workshop
Getting Started with Text Analytics

- Text Analytics is weird, a bit academic, and not very practical
  - It involves language and thinking and really messy stuff
- On the other hand, it is really difficult to do right (Rocket Science)
- Organizations don't know what text analytics is and what it is for
- 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
  - Experience taking taxonomy starters and customizing, rules
- Interdisciplinary team – Need to create

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

- Think Big: Strategic Vision
  - Based on deep understanding of entire information environment – Knowledge Audit – content, technology, people, business activities
  - Establish infrastructure – faster project development
  - Avoid expensive mistakes – dead end technology
- Start Small: Pilot or POC
  - Immediate value ands learn by doing
  - Easier to get management buy-in
- Scale Fast: Multiple Applications
  - Infrastructure – technical and semantic
  - Semantic Resources – catonomies, ontologies
  - First project +10%, Subsequent projects – 50%

KAPS Group

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 – initial selection
- Select 1-2 for a pilot/POC
- POC use cases – basic features needed for initial projects
- Design - Real life scenarios, categorization with your content
- Four week POC – 2 rounds of develop, test, refine / Not OOB
- Majority of time is on auto-categorization
- Option – have software, but stuck or abandoned
  - Train people in good practices

Text Analytics Workshop
POC and Early Development: Risks and Issues

- IT Problem - This is not a regular software process
- 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
- Anyone can do categorization

Quick Start for Text Analytics
Proof of Concept -- Value of POC

- Selection of best product(s)
- Training by doing –SME's learning categorization, Library/taxonomist learning business language
- Understand effort level for categorization, application
- Test suitability of existing taxonomies for range of applications
- Explore application issues – example – how accurate does categorization need to be for that application – 80-90%
- Develop resources – categorization taxonomies, entity extraction catalogs/rules

KAPS Group
Knowledge Architecture Professional Services

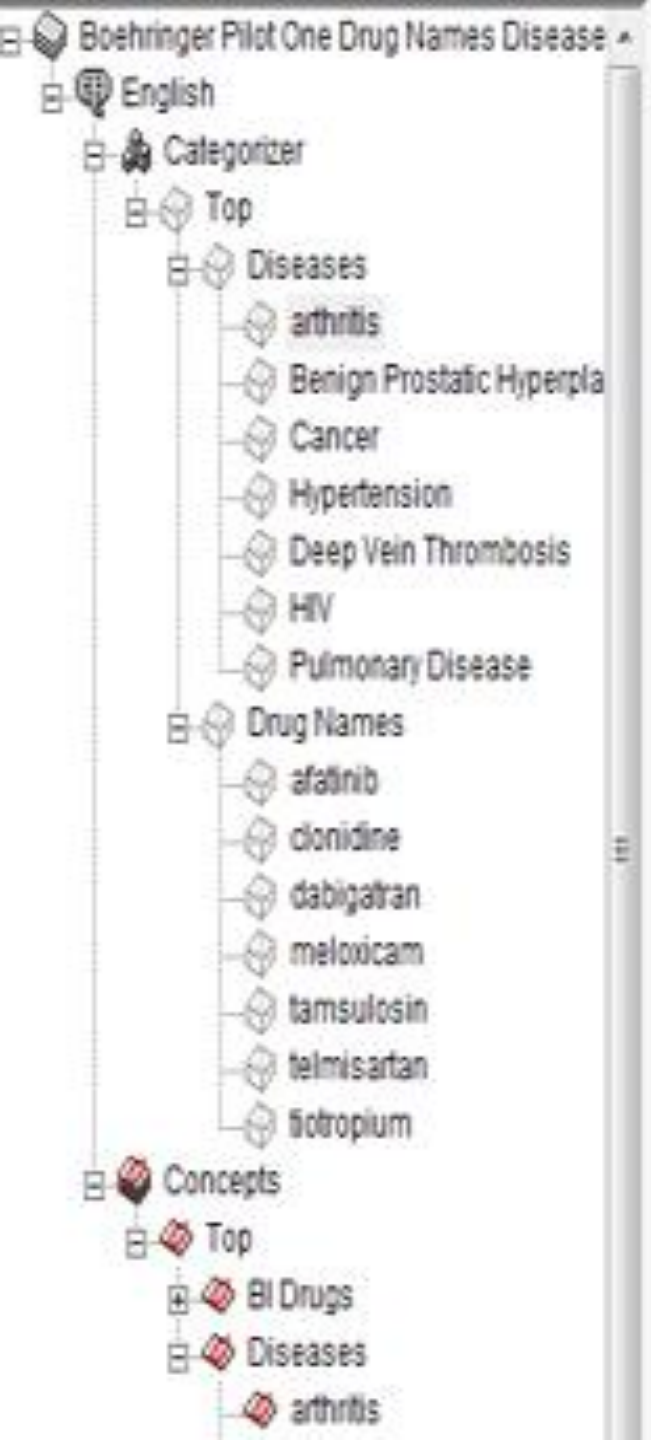
Text Analytics Development

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
- Building blocks
  - Taxonomy, Content, Supplementary Resources

Intel Mini-POC
Categorization Techniques – Three Basic Types

- Statistical – Bayesian, Vector space with machine learning
  - Create a statistical 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
- Templates + Rules
  - Content types + sections
  - Good results – easier to use, less flexible

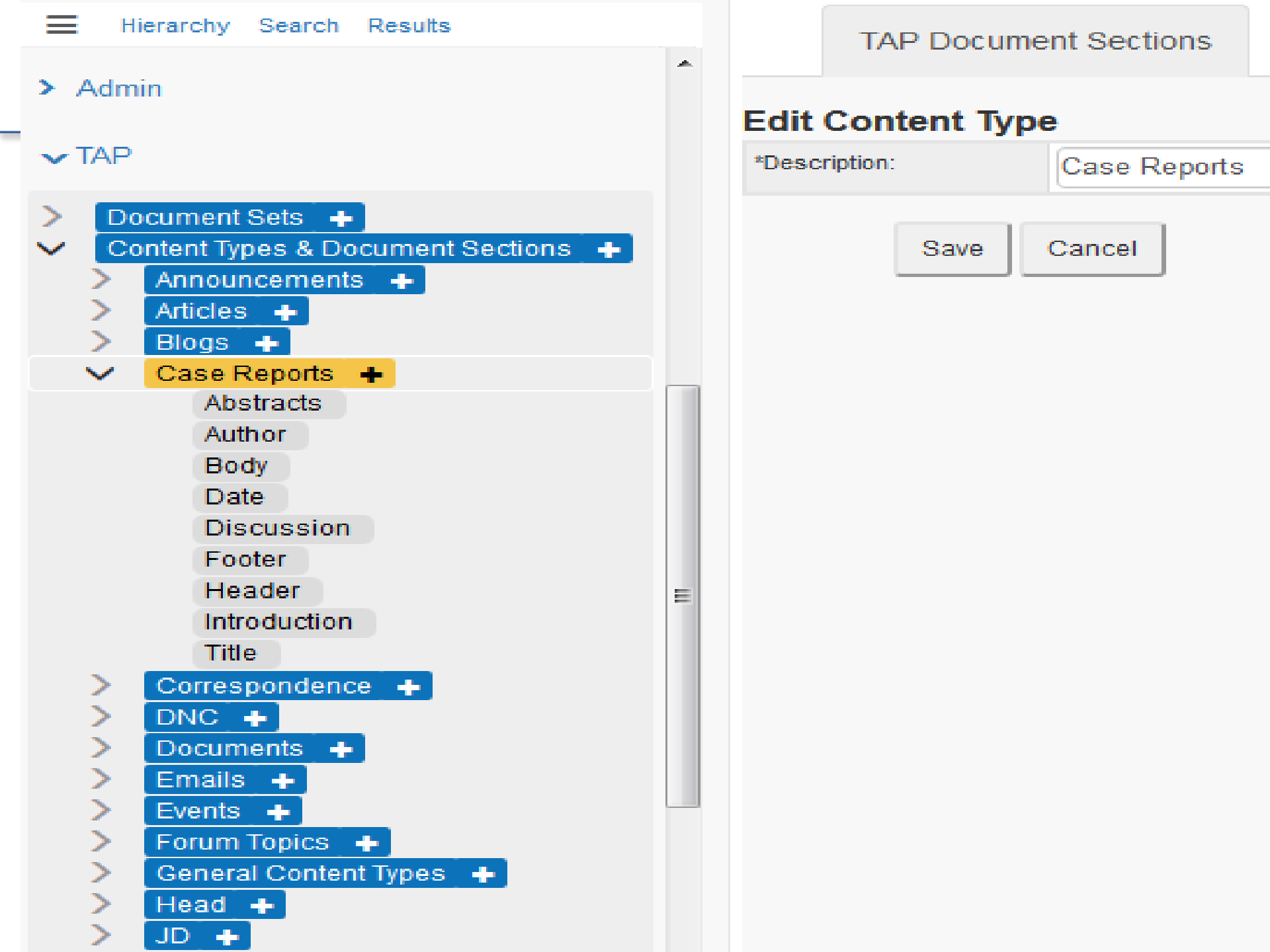



```
(OR,
_/article/title:"[arthritis]",

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

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

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


Hierarchy
Search
Results
Admin
TAP
Document Sets
Content Types & Document Sections
Announcements
Articles
Blogs
Case Reports
Abstracts
Author
Body
Date
Discussion
Footer
Header
Introduction
Title
Correspondence
DNC
Documents
Emails
Events
Forum Topics
General Content Types
Head
JD
TAP Document Sections
Edit Content Type
*Description:
Case Reports
Save
Cancel

Text Analytics Workshop
Statistical vs. Rules

- Current trend – how to combine
- Claim – ML is faster to develop – only if unsupervised – typically bad results
- Selecting documents is time and effort – and difficult
- Start with ML, pass through a taxonomy, human review, custom models for domain-specific annotation
- Pass into an ontology
- Me – one solutions is content model – statistical based on sections – issue = not enough text?

Text Analytics Development: Categorization Process Start with Taxonomy and Content

- Starter Taxonomy
  - If no taxonomy, develop (steal) 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
- External Resources
  - Linked Data – General = DBPedia, Other
  - Linked Data – Specialized – from K Audit

Text Analytics Development: Categorization Process Start with Taxonomy and Content

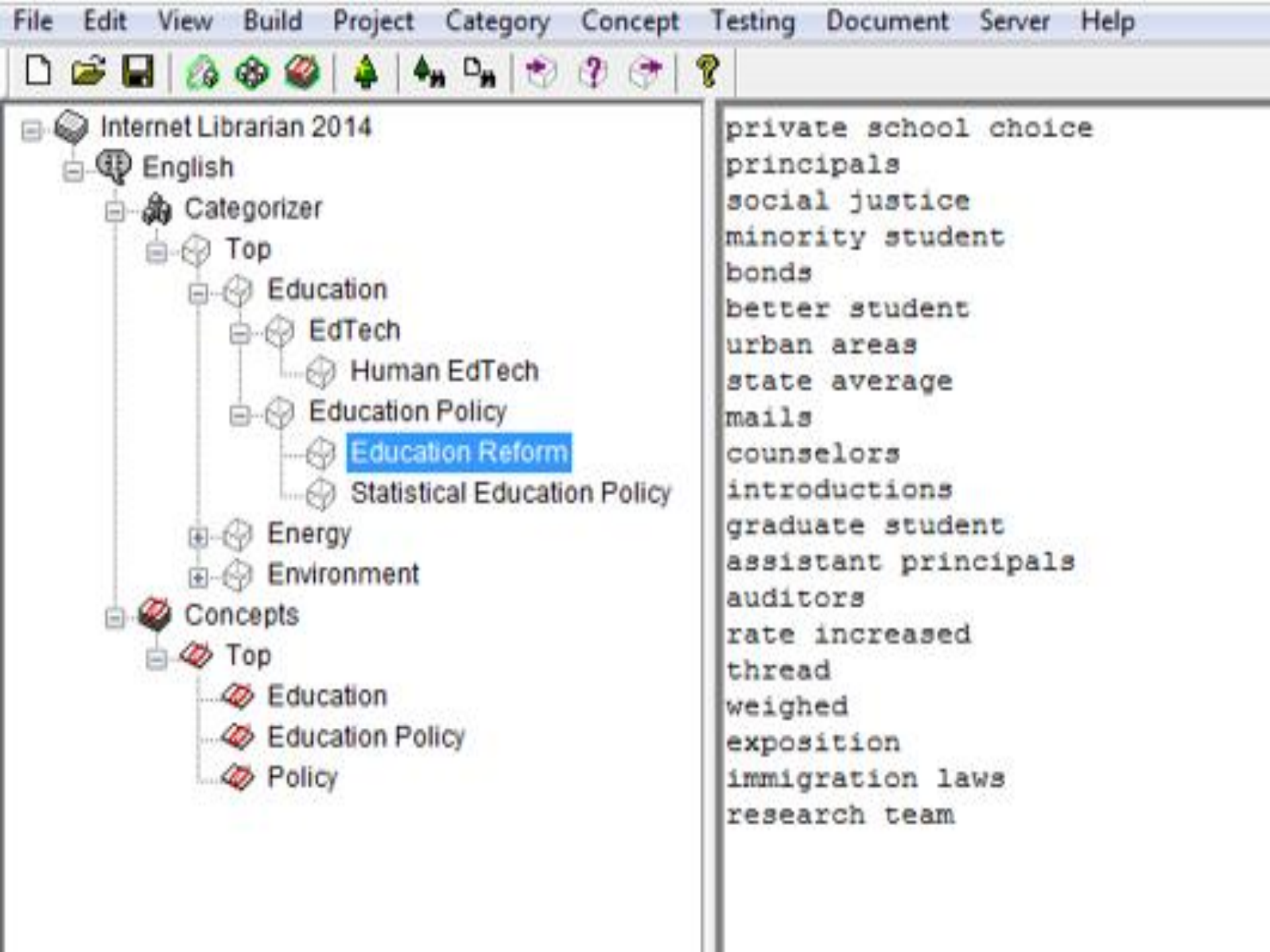
- Content Selection
  - Map of all anticipated content – from K Audit
  - Most common and most important, special cases
  - Map to information needs
- Selection of training sets – 20-50-100 per category
  - SME input, search logs, information interviews
  - Trick – category name in file name
- Automated selection of training sets
  - Taxonomy nodes as first categorization rules
  - Apply and get content
- 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 appear often / important to content (TF/IDF) // Auto-rule
  - Multiple passes – sub-types of content
  - Clustering – word or tag clouds
- Metadata
  - Title, keywords
  - Abstract
  - Special sections – Methods, Objectives, etc.
  - Headings, bold, italics
- Human generated
  - Sections in the text – indicator text
  - Search logs

Text Analytics Workshop
Text Analytics Development: Categorization Process

- 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.) All training sets per category – build recall
  - 4.) Test against all training sets - precision
- Refine rules – patterns in text - break trade off of recall-precision
  - Distance – CLAUSE, SENTENCE, PARAGRAH, DIST
  - Minimum occurrences – only if 2-3+
  - Sections - weights
- Develop templates – separate logic and vocabulary
- Test against more, new content – more terms, refine logic of rules
- Repeat until “done” – 90%?


File Edit View Build Project Category Concept Testing Document Server Help
Internet Librarian 2014
English
Categorizer
Top
Education
EdTech
Human EdTech
Education Policy
Education Reform
Statistical Education Policy
Energy
Environment
Concepts
Top
Education
Education Policy
Policy
private school choice
principals
social justice
minority student
bonds
better student
urban areas
state average
mails
counselors
introductions
graduate student
assistant principals
auditors
rate increased
thread
weighed
exposition
immigration laws
research team

File Edit View Build Project Category Concept Testing Document Server Help

- WBG Taxonomy 1
  - English
    - Categorizer
      - Top
        - Agriculture and Rural Development
        - Debt, Fiscal and Macroeconomic Po
        - Economic Growth
        - Education
          - Abstract
          - Tertiary Education
          - Tertiary Education Old Style
          - Title
        - Energy and Mining
        - Environment
        - Financial and Private Sector
        - Gender
        - Health
        - Information and Communication Tec
        - Poverty
        - Public Administration, Law and Justi
        - Social Development
        - Social Protections and Labor
        - Trade
        - Transport
        - Urban Development
        - Water
    - Concepts
      - Top
        - Education
          - Tertiary Education

```
(START_3000,
(AND,
_/add/doc/docty:"Education Sector Review",

(OR,
_//field[@name='display_title']:"[Tertiary Education]",
_//field[@name='subtopic']:"[Tertiary Education]",
(MINOC_2,_//field[@name='keywd']:"[Tertiary Education]"),
(MINOC_2,_//field[@name='keywordsv2']:"[Tertiary Education]"),
(MINOC_2,_//field[@name='abstracts']:"[Tertiary Education]"),
(MINOC_4,_//field[@name='contentTxt']:"[Tertiary Education]")
)
)
)
```

Mini-POC: The Process Overview

- 40 hours of effort
- Selection of 10 Categories – Subject Taxonomy
  - Basic to search
  - Future – use entity/data taxonomies as facets
- Selection of 10-20 documents per category
  - Good examples of documents about category
- Develop content Structure model rules – document structure
- Develop categorization rules for each category
  - Three rounds of refinement

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

EXECUTIVE SUMMARY

The shortage of nursing faculty in the United States is a critical problem that directly affects the nation's nurse shortage, which is projected to worsen in future years. Short-term interventions to address the nursing shortage are inadequate given the increasing needs of a growing and aging population. A substantial increase in newly educated nurses will be needed to meet future demand; therefore, timely and sustainable interventions to reduce the nursing faculty shortage are required. This paper highlights solutions to the faculty shortage by:

- describing the current faculty shortage in relation to demand, supply, educational preparation and productivity;
- examining the factors that contribute to the faculty shortage;
- reviewing the array of interventions already undertaken; and
- outlining recommendations for further action.

The paper is based on a review of published literature and data, including surveys by government and professional organizations, studies by state task forces addressing the nursing shortage, foundation reports, and reviews by such groups as the National Conference of State Legislatures of activities at the state level, as well as author interviews with leaders in nursing education.

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

```
(OR,(START_100,
(AND,
(OR,"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]")),
(NOT,
(START_100,"[Child & Family Well-being Negatives]")
))),

(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
(OR,
"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]"))),
(NOT,
(AND,

(DIST_25,
(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
(OR,
"[Child & Family Well-being Terms]",
(DIST_7, "[Child & Family Terms]", "[Well-being Terms]"))),
(AND,
 (ORDDIST_500,
"[Document Summary Indicators]",
 "[Child & Family Well-being Negatives]"))
))))))
```

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”

Score with Sections Category	Recall	Total Precision	Top 10 Precision	Notes
Child & Family Well-being	95%	100%	100%	
Childhood Obesity	100%	95%	100%	
Disease Prevention & Health Promotion	90%	85%	90%	
Health Care Coverage & Access	95%	95%	100%	
Nurses & Nursing	95%	95%	100%	
Public & Community Health	95%%	70%	100%	
Coalition & Network Building	93%	93%	100%	
Health Professional	85%	100%	100%	
Immigrant or Migrant	100%	94%	100%	
Policymaker	100%	91%	100%	
Average	**95%**	**92%**	**99%**	

Scores without Sections – Full Text	Recall	Total Precision	Top 10 Precision	Notes
Child & Family Well-being	75%	43%	80%	
Childhood Obesity	100%	67%	70%	
Disease Prevention & Health Promotion	50%	27%	40%	
Health Care Coverage & Access	80%	33%	90%	
Nurses & Nursing	40%	27%	80%	
Public & Community Health	45%%	17%	90%	
Coalition & Network Building	73%	48%	90%	
Health Professional	75%	31%	70%	
Immigrant or Migrant	100%	71%	100%	
Policymaker	75%	50%	100%	
Average	**71%**	**41%**	**81%**	

KAPS Group

RWJF Mini-POC
Overview Average Scores

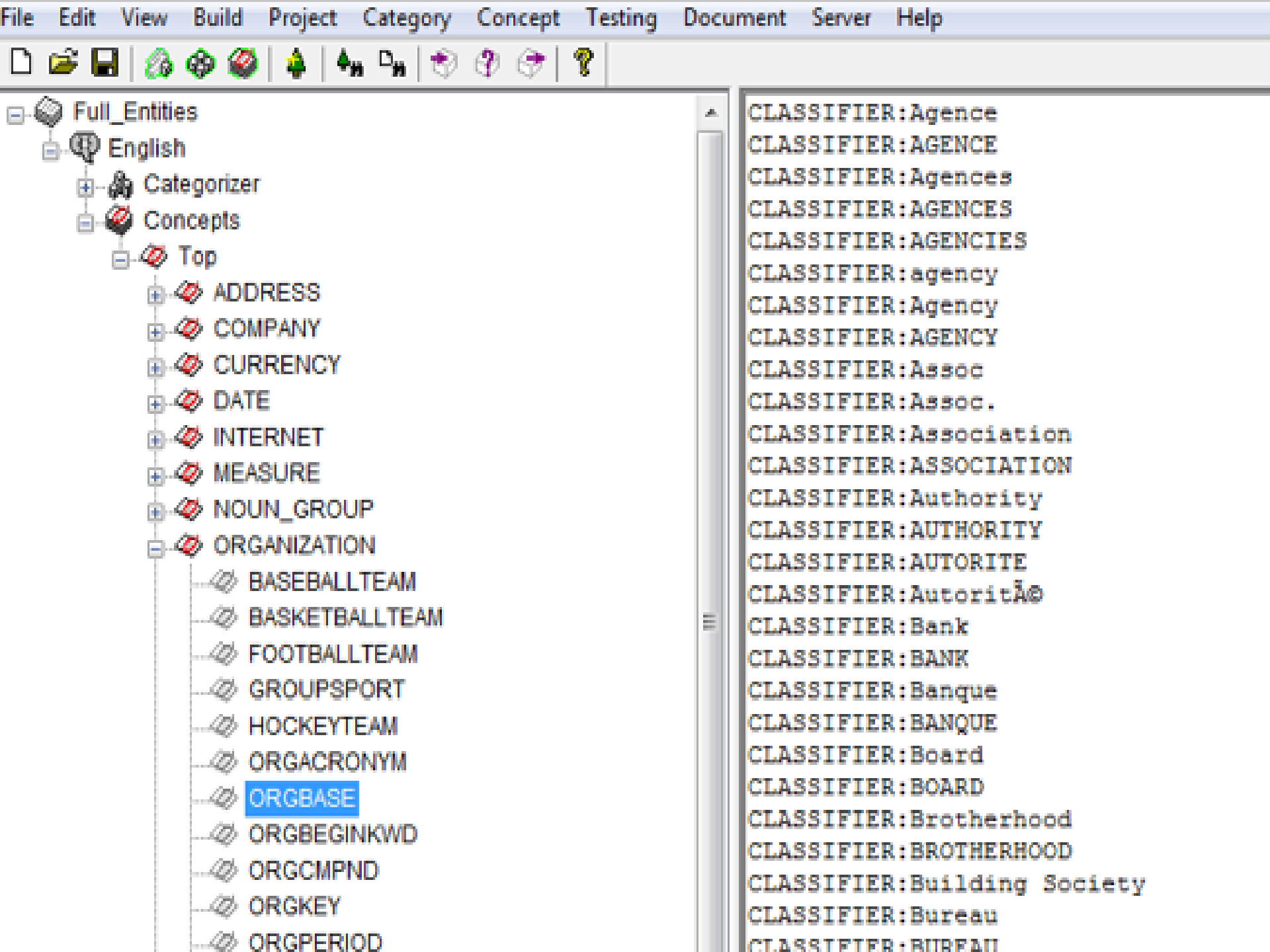
	Recall	Precision	Precision Top 10
With Sections	95%	92%	99%
Full Text	71%	41%	81%
Difference	24%	51%	18%

Content Structure Models
Implications for Taxonomists

- Categorization and data extraction built on taxonomies
  - Bad taxonomies can hurt
- Text analytics can help build good taxonomies
  - Combine conceptual and content analysis
  - Beautiful taxonomy needs to reflect the content
- Taxonomists make the best text analysts
- Added benefit – evaluate taxonomies – against content
  - How orthogonal are facets – effort level, number of terms per rule
  - Also indicator of specificity
  - Very difficult to distinguish 2 categories – rethink?

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")

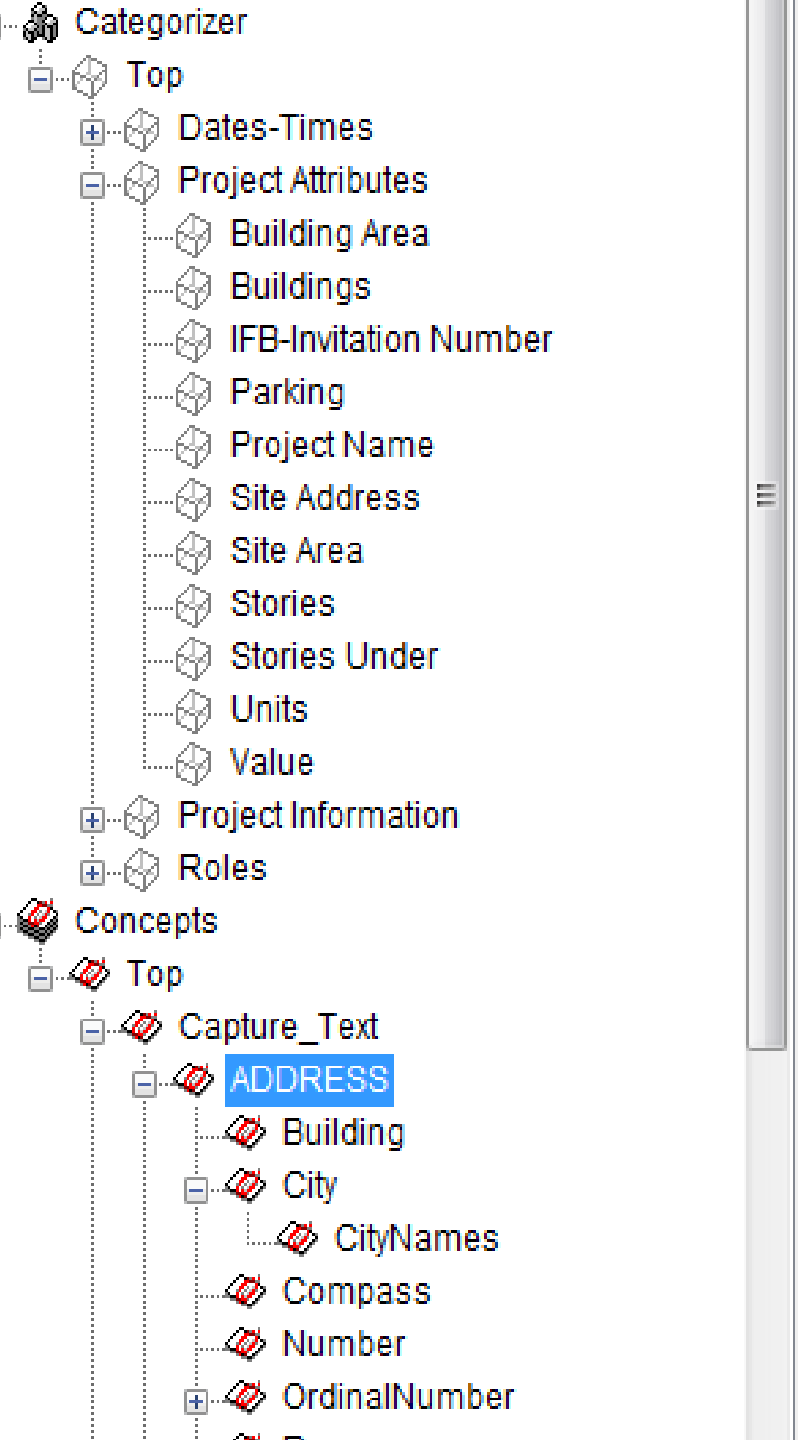
- Content Type
  - Source of Materials
  - DWR,
  - Work Order,
  - Work Order-Related
  - Project Profile
- Project No/Contract No/UPC
- Location: District, Jurisdiction, Route
- Type of Work
- Award Amount
- Manufacturers and Suppliers
- Contractors
- Materials
- Equipment
- Pay Items
- Work Order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

Text Analytics Workshop
Multiple Fact Extraction

- Application – Rich Summary of Key Data in Construction Proposals
- 700, 000 + a year, range in size from 5 pages to 500, 000 pages
- Earlier project – aim was 70% - failed
  - Easy to extract all items, example – dates – couldn't extract specific dates- facts
- 10 week project - new team, new software
- Develop basic methods – fact extraction, automated TOC, accuracy > 80%
- Train client resources to continue and expand capabilities

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)

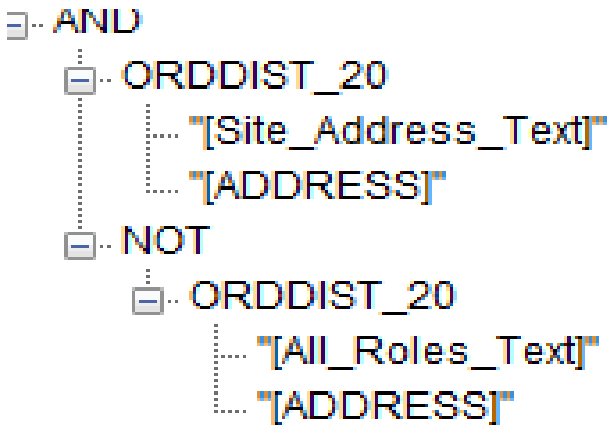



```
*ADDRESS = !City , !StateCode
*ADDRESS = !CityNames , !StateCode
*ADDRESS = !City , !StateName
*ADDRESS = !CityNames , !StateName
*ADDRESS = !City , !StateNameAbbr
*ADDRESS = !CityNames , !StateNameAbbr
*ADDRESS = !City , !StateNameCap
*ADDRESS = !CityNames , !StateNameCap


*ADDRESS = !City , !StateCode !ZipCode
*ADDRESS = !CityNames , !StateCode !ZipCode
*ADDRESS = !City , !StateName !ZipCode
*ADDRESS = !CityNames , !StateName !ZipCode
*ADDRESS = !City , !StateNameAbbr !ZipCode
*ADDRESS = !CityNames , !StateNameAbbr !ZipCode
*ADDRESS = !City , !StateNameCap !ZipCode
*ADDRESS = !CityNames , !StateNameCap !ZipCode

*ADDRESS = !City , !StateCode , !ZipCode
*ADDRESS = !CityNames , !StateCode , !ZipCode
*ADDRESS = !City , !StateName , !ZipCode
*ADDRESS = !CityNames , !StateName , !ZipCode
*ADDRESS = !City , !StateNameAbbr , !ZipCode
*ADDRESS = !CityNames , !StateNameAbbr , !ZipCode
*ADDRESS = !City , !StateNameCap , !ZipCode
*ADDRESS = !CityNames , !StateNameCap , !ZipCode

*ADDRESS = !StreetAddress !City , !StateCode
*ADDRESS = !StreetAddress !CityNames , !StateCode
```


AND
ORDDIST_20
"[Site_Address_Text]"
"[ADDRESS]"
NOT
ORDDIST_20
"[All_Roles_Text]"
"[ADDRESS]"

Text Analytics Workshop Applications

Text Analytics Workshop Applications: Survey Results – 2017

- Important Areas:
  - Business Intelligence – 87%
  - Decision Support - 83%
  - Financial Intelligence – 81%
  - KM-Productivity – 80%
  - Search – Search Apps – 78%
  - Security – 77%
  - Compliance – 76%
  - Voice of Customer – 73%
  - Social Media Analysis – 69%

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: Information Environment Metadata – Tagging – Mind the Gap

- Tagging documents with taxonomy nodes is tough
  - And expensive – central or distributed
- Library staff –experts in categorization not subject matter
  - Too limited, narrow bottleneck
  - Often don't understand business processes and uses
- 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
Information Platform: Content Management

- Hybrid Model – Internal 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
Enterprise Info Apps

- Focus on business value, cost cutting, new revenues
- Business Intelligence – integrate data (what) and text (why)
- Financial Services – risk and fraud
- Customer Relationship Management
- eDiscovery – legal, research – grant proposals
- Structured Summaries – extract key data and facts
- KM – expertise location, community portals, ESN
- News aggregator

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 Characteristics

- Scale = Huge! 100's of Millions / Billions
- Poor Quality of the Text
- Conversations, not stand alone documents
  - Issues of co-reference, who is speaking
- Direct Business Value
  - Customers, competitors, fix products, new products
- Document Level Sentiment – too broad, too complex
- From direct monitoring (surveys) to Indirect (Twitter)
- Add depth with more sophisticated text analytics

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

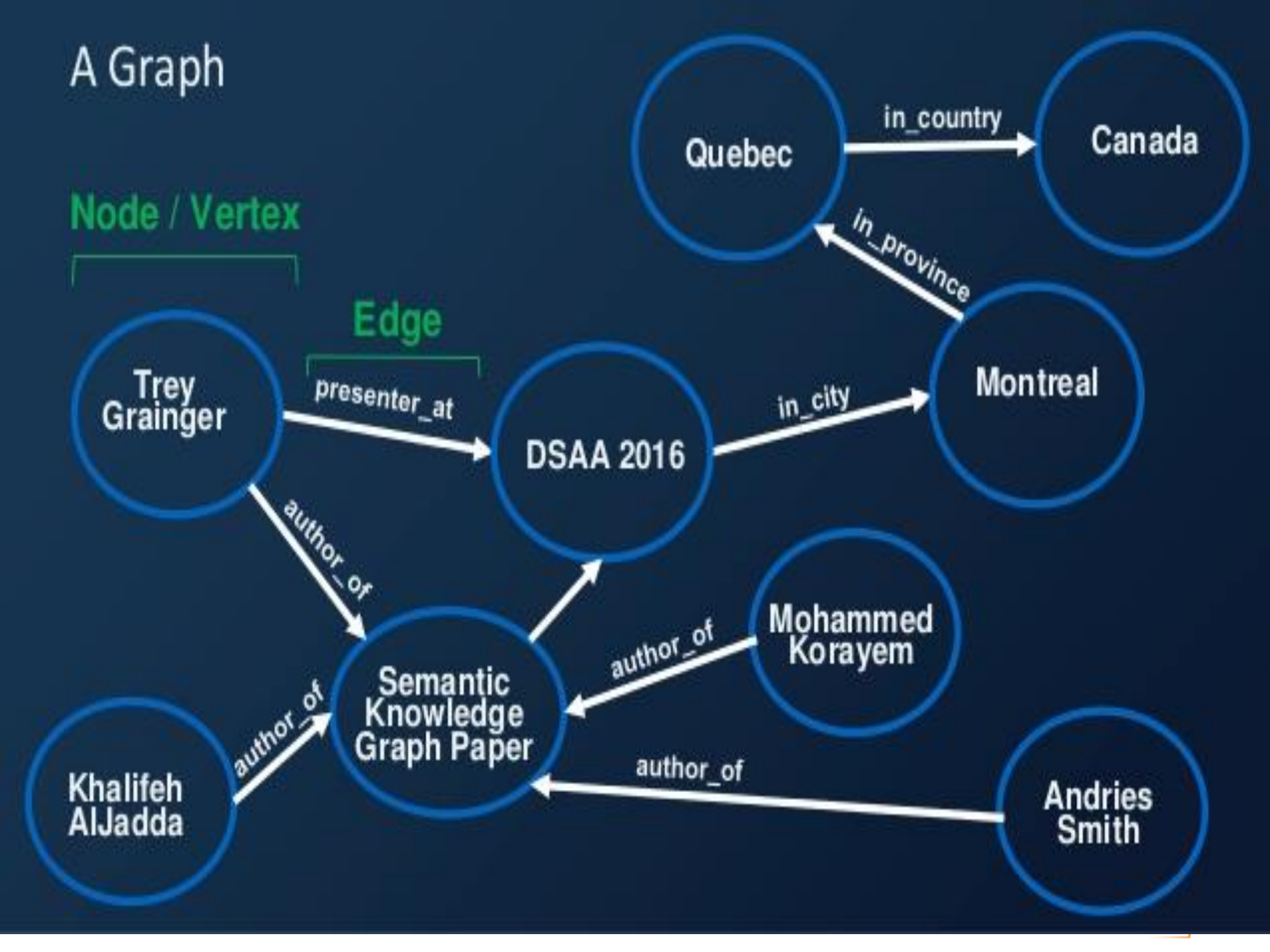
Social Media Applications
Pronoun Analysis: Fraud Detection; Enron Emails

- 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

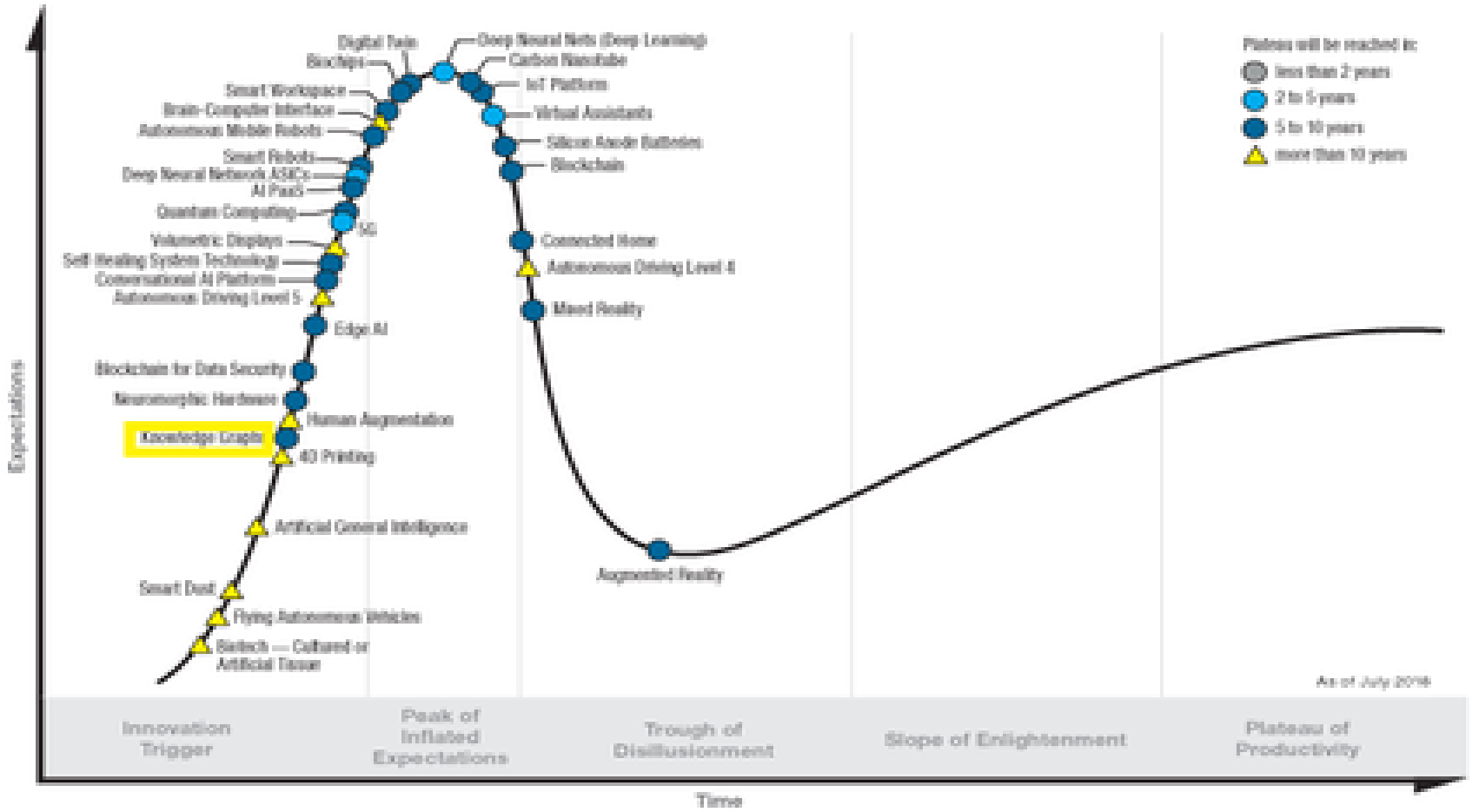
Text Analytics Workshop
Knowledge Graphs

Text Analytics Workshop
Knowledge Graphs – What are K Graphs?

- Entities (Nodes) and their relationships (Properties)
- Collections of RDF Triples stored in a graph database
- Graph databases
  - Data structure = edges and nodes
- Amazon Neptune, Microsoft Azure, Neo4j
- Franz, Stardog
- Marklogic
- GraphDB


A Graph
Node / Vertex
Edge
Quebec
in_country
Canada
in_province
Trey Grainger
presenter_at
DSAA 2016
in_city
Montreal
author_of
Mohammed Korayem
author_of
Semantic Knowledge Graph Paper
author_of
Khalifeh AlJadda
author_of
Andries Smith

Hype Cycle for Emerging Technologies, 2018



gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

Gartner.

Text Analytics Workshop
Knowledge Graphs - Advantages

- Store semantics in a structured format easily used by computers
- Fast, scalable, consistent data – avoid data migration
- Emerging standards & resources
  - FIBO – Financial ontology
- Integration of data silos – no coding
  - Enterprise K Graph, structured and unstructured
  - External data – linked data resources
  - Financial data and 3rd party data
- Computer reasoning – relationships
- Combine with text analytics- disambiguation can be done with rules and/or ML

Text Analytics Workshop
Knowledge Graphs - Applications

- Search – facets, linked data – people, organizations,
  - Can build answering apps, not just links
- Match customer queries to official terminology
- Fraud detection, Insider threats, recommendations, CRM, digital assistants, Regulatory reporting
- Accelerate data prep for AI, ML
- Social graphs
- Voice assistants
- Taskonomy – taxonomy + K Graph

Text Analytics Workshop
Knowledge Graphs - Development

- Start small – high impact use cases
- Statistics only = noisy, need extraction combined with Text Analytics
- Entity types – stored in an ontology
- Combine multiple taxonomies into an ontology
- Salience (Discovery) vs. aboutness (Search)
- Primary limits
  - Data not concepts, things not strings
  - No SQL (standard language)

Text Analytics Workshop Conclusions

- Partnership – Taxonomy/Ontology and Text Analytics
- Taxonomy - structure for text analytics
- Text Analytics – applies the taxonomy and helps develop
  - Evaluation tools
  - Categorization rules always change/deepen the taxonomy
- Taxonomists make the best text analysts
- Text analytics best approached as infrastructure – platform for multiple applications
- Categorization is the brains of the outfit
  - Smart applications, makes everything else smarter
- AI, K Graphs need concepts – Text Analytics & Taxonomy

Questions?

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

KAPS Group
Knowledge Architecture Professional Services