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 – State of Text Analytics
  - Elements of Text Analytics
  - Current State of Text Analytics & Future Trends
  - Value of Text Analytics
- Getting Started with Text Analytics
- Development – Taxonomy, Categorization, Faceted Metadata
- Text Analytics Applications
  - Platform for Information Applications
  - Integration with Search and ECM
  - Multiple Applications
- Questions / Discussions

Introduction: KAPS Group

- Network of Consultants and Partners
- Services:
- Strategy – IM & KM - Text Analytics, Social Media, Integration
- Taxonomy/Text Analytics, Social Media development, consulting
- Text Analytics Smart Start – Audit, Evaluation, Pilot
- Partners – Smart Logic, Expert Systems, SAS, SAP, IBM, FAST, Concept Searching, Clarabridge, Lexalytics
- Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, Dept. of Transportation, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Book –Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data

Introduction:
Elements of Text Analytics

- Text Mining – NLP, statistical, predictive, machine learning
  - Different skills, mind set, Math not language
- Semantic Technology – ontology, fact extraction
- Extraction – entities – known and unknown, concepts, events
  - Catalogs with variants, rule based
- Sentiment Analysis
  - Objects and phrases – statistics & rules – Positive and Negative
- Summarization
  - Dynamic – based on a search query term
  - Generic – based on primary topics, position in document

Introduction:
Elements of Text Analytics

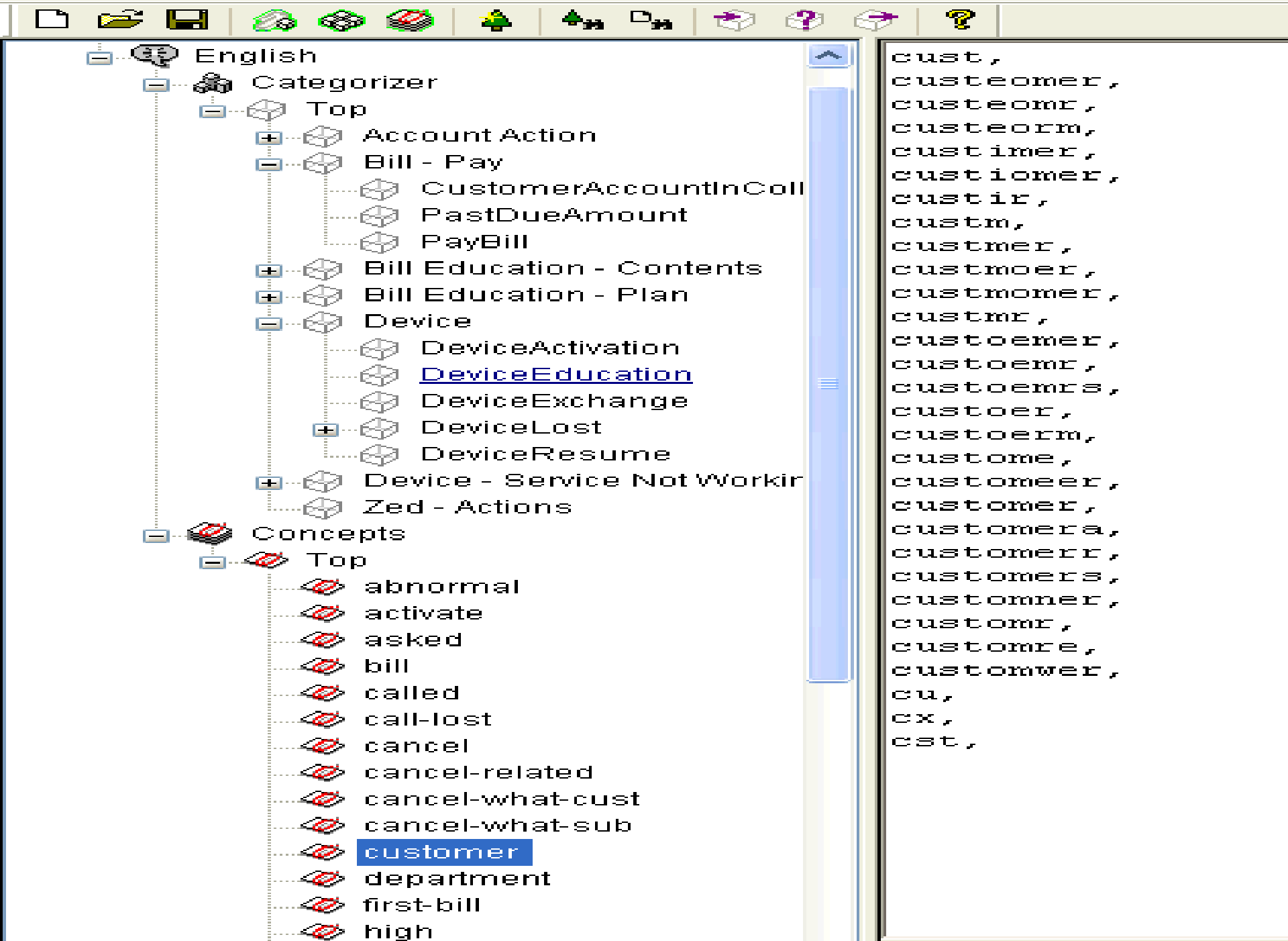
- Auto-categorization
  - Training sets – Bayesian, Vector space
  - Terms – literal strings, stemming, dictionary of related terms
  - Rules – simple – position in text (Title, body, url)
  - Semantic Network – Predefined relationships, sets of rules
  - Boolean– Full search syntax – AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
- Platform for multiple features – Sentiment, Extraction
  - Disambiguation - Identification of objects, events, context
  - Distinguish Major-Minor mentions
  - Model more subtle sentiment

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
      - asked
      - bill
      - called
      - call-lost
      - cancel
      - cancel-related
      - cancel-what-cust
      - cancel-what-sub
      - customer
      - department
      - first-bill

```
(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



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
```

File Edit View Build Project Category Concept Testing Document Server Help

- LITI_Demo
  - English
    - Concepts
      - Top
        - ALIGNED_CONCEPT
        - CONCEPT_RULE
        - FACT_EXT
        - FACT_RULE
        - INTERNET
        - LOCATION
          - COUNTRY
          - LOC
          - LOCCMPND
          - LOCPRF
          - LOCSUFF
        - NOBREAKC
        - ORGANIZATION
        - PERSON
          - FIRST
          - FULLNAME
          - LAST
          - NATIONMOD
          - TITLE
        - PERSON_EXPORT
          - SFN
          - SLN
        - PRIORITY_CONCEPT

```
# We learn person last name from context, e.g. capital word followed by "said".
# NOTE: ">" (called export) indicates that last name learnt once with "said" will
#       also be matched everywhere else without "said"

C_CONCEPT:_c{ _cap }> said

# A first name is a capital word that appears before a Lastname (SLN)

C_CONCEPT:_c{ _cap } SLN

# Rule indicates that anything matched by PERSON_EXPORT is more reliable than "_cap _cap _w" match

C_CONCEPT:PRIORITY=40:_c{ PERSON_EXPORT }
CONCEPT:_cap _cap  _w
```

Syntax Check | Classifier | Filename | Grammar | LITI | Load Text...

Taxonomy | Dependencies | Definition | Testing | Data | Document

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

```
SEQUENCE:(keyw,org1,org2):_keyw{PNAKW} between _org1{ ORGANIZATION } and _org2{ ORGANIZATION }
```

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: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
```

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

```
# Example of regular expression matching for identifying URLs

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
        - Negative
        - Neutral
      - size
        - Positive
        - Negative
        - Neutral

	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.
22	CLASSIFIER	but unfortunately
23	CLASSIFIER	Don't Buy This Camera
24	CLASSIFIER	little outdated
25	PREDICATE_	(SENT, "_a{stuck}", "_b{error}")
26	CLASSIFIER	am disgusted with
27	CLASSIFIER	save your self some trouble

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
8	7	11%	245	hardware, + hardware unit, + drive resource management issue, hardware availability, hardware development, hardware precision, + hardware basic, hardware design, + hardware resource, hardware acceleration, + hardware configuration
9	8	10%	217	+ component, + technology, + mechanism, + parallel, + optimization
10	9	4%	87	+ equipment, hardware cache due, + router, hardware, + memory, + device, + component, + technology, + mechanism, + optimization
11				

Text Analytics Workshop
Introduction: 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 2008 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 is growing – time for a jump?

Text Analytics Workshop
Current State of Text Analytics

- Current Market: 2013 – exceed $1 Bil for text analytics (10% of total Analytics)
- Growing 20% a year
- Search is 33% of total market
- Other major areas:
  - Sentiment and Social Media Analysis, Customer Intelligence
  - Business Intelligence, Range of text based applications
- Fragmented market place – full platform, low level, specialty
  - Embedded in content management, search, No clear leader.

Interviews with Leading Vendors, Analysts: Current Trends

- From Mundane to Advanced – reducing manual labor to "Cognitive Computing"
- Enterprise – Shift from Information to Business – cost cutting, new applications rather than productivity gains
- Growth of social media & sentiment – cash and customers
- Deep Text
  - Linguistic and cognitive depth – human-like learning
  - Integration of multiple techniques and modules
  - Infrastructure – Move fast with a stable infrastructure
- Big Data – more focus on extraction (where it began) but categorization adds depth and sophistication

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
Current State of Text Analytics: Vendor Space

- Taxonomy Management
- Extraction and Analytics
  - Multiple Dedicated Applications – BI, CI, social media
- Sentiment Analysis
- Open Source, build your own – API's
- Embedded in Content Management, Search
- Full text analytics platforms

Text Analytics Workshop
Future Directions: Survey Results

- Important Areas:
  - Predictive Analytics & text mining – 90%
  - Search & Search-based Apps – 86%
  - Business Intelligence – 84%
  - Voice of the Customer – 82%, Social Media – 75%
  - Decision Support, KM – 81%
  - Big Data- other – 70%, Finance – 61%
  - Call Center, Tech Support – 63%
  - Risk, Compliance, Governance – 61%
  - Security, Fraud Detection-54%

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

Text Analytics Workshop
Benefits of Text Analytics: Search

- Why Text Analytics?
  - Enterprise search has failed to live up to its potential
  - Enterprise Content management has failed to live up to its potential
  - Taxonomy has failed to live up to its potential
  - Adding metadata, especially keywords has not worked
- What is missing?
  - Intelligence – human level categorization, conceptualization
  - Infrastructure – Integrated solutions not technology, software
- Text Analytics can be the foundation that (finally) drives success – search, content management, KM, and much more

Text Analytics Workshop
Costs and Benefits

- 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 – Social Media

- Understand what customers are saying – satisfaction
  - Customer management effectiveness
  - Insight into customers mind
- Early warning of issues with products
- Lead generation
- Managing brand perception
- Product design insight
- Marketing campaign effectiveness
- Lead generation
- Attrition rate management / reduction

Text Analytics Workshop
Benefits – Info Apps

- Multiple applications
- Email audit – find money owed
- Products – summary of 700K documents
- Customer support – head off cancelations
- Reduce fraud
- Improve customer support – trends, issues, etc.
- Range of applications – almost unlimited

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

- Don't know what text analytics is
- Actually are but don't know it
- Don't do text analytics – not part of culture
- IT doesn't understand the value – yet
- Don't believe ROI calculations
- It's too complex
- It's too expensive

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
Selling the Benefits

- 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
- Don’t oversell – not another revolution?

Future of Text Analytics
Primary Obstacle: Complexity

- Usability of software is one element
- More important is difficulty of conceptual-document models
  - Language is easy to learn , hard to understand and model
- Need to add more intelligence (semantic networks) and ways for the system to learn – social feedback
- Customization – Text Analytics– heavily context dependent
  - Content, Questions, Taxonomy-Ontology
  - Level of specificity – Telecommunications
  - Specialized vocabularies, acronyms

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 experience putting together

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

- Think Big: Strategic Vision
  - Based on deep understanding of entire information environment
  - Establish infrastructure – faster project development
  - Avoid expensive mistakes – dead end technology, etc.
- Start Small: Pilot or POC
  - Immediate value and 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%

Text Analytics Workshop
Smart Start Step One- Knowledge Audit

- Info Problems – what, how severe
- Formal Process – Knowledge Audit
  - Contextual & Information interviews, content analysis, surveys, focus groups, ethnographic studies, Text Mining
- Informal for smaller organizations, specific application
- Category modeling – Cognitive Science – how people think
  - Panda, Monkey, Banana
- Natural level categories mapped to communities, activities
    - Novice prefer higher levels
    - Balance of informative and distinctiveness
- Strategic Vision – Text Analytics and Information/Knowledge Environment

KAPS Group
Knowledge Architecture Professional Services

Smart Start Step Two - Software Evaluation Different Kind of software evaluation

- Traditional Software Evaluation - Start
  - Filter One- Ask Experts - reputation, research – Gartner, etc.
    - Market strength of vendor, platforms, etc.
    - Feature scorecard – minimum, must have, filter to top 6
  - Filter Two – Technology Filter – match to your overall scope and capabilities – Filter not a focus
  - Filter Three – In-Depth Demo – 3-6 vendors
- Reduce to 1-3 vendors
- Vendors have different strengths in multiple environments
  - Millions of short, badly typed documents, Build application
  - Library 200 page PDF, enterprise & public search

Smart Start Step Three – Proof of Concept / Pilot Project

- POC use cases – basic features needed for initial projects
- Design - Real life scenarios, categorization with your content
- Preparation:
  - Preliminary analysis of content and users information needs
    - Training & test sets of content, search terms & scenarios
  - Train taxonomist(s) on software(s)
  - Develop taxonomy if none available
- Four week POC – 2 rounds of develop, test, refine / Not OOB
- Need SME's as test evaluators – also to do an initial categorization of content
- Majority of time is on auto-categorization

Text Analytics Workshop
POC and Early Development: Risks and Issues

- CTO 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
  - Librarians often overdo, SME's often get lost (keywords)
- Meta-language issues – understanding the results
  - Need to educate IT and business in their language

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

- Selection of best product(s)
- Identification and development of infrastructure elements – taxonomies, metadata – standards and publishing process
- 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

Development

Text Analytics Development: Categorization Basics

- Representation of Domain knowledge – taxonomy, ontology
- Categorization – Know What
  - Most basic to human cognition
  - Basic level categories
  - Most difficult to do with software
- Beyond Categorization – making everything else smarter
- No single correct categorization
  - Women, Fire, and Dangerous Things
- Sentiment Analysis to Expertise Analysis(KnowHow)
  - Know How, skills, “tacit” knowledge

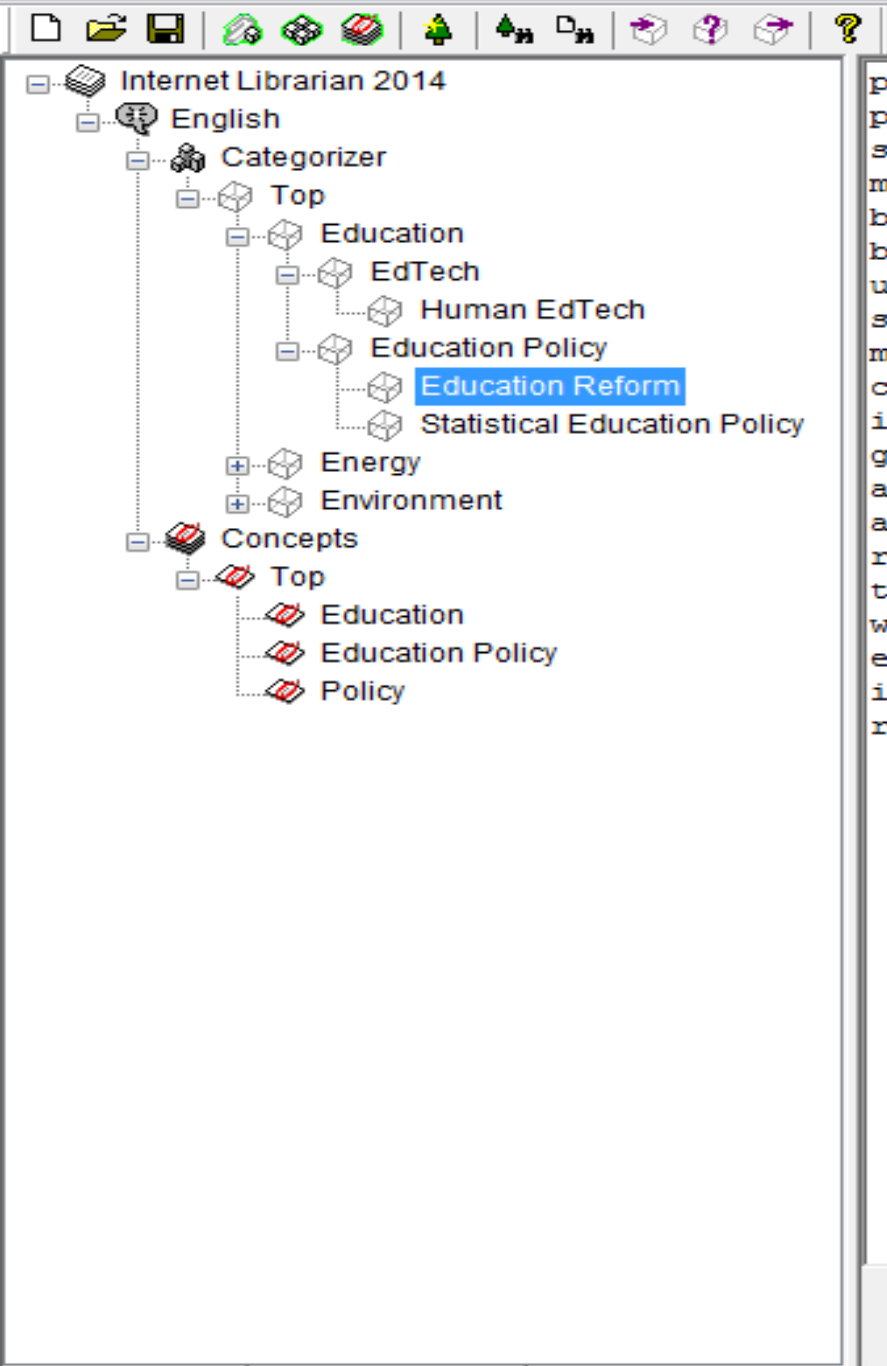
Text Analytics Development: Categorization Process Start with Taxonomy and Content

- 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
- Content Selection
  - Map of all anticipated content
  - Selection of training sets – if possible
  - Automated selection of training sets – taxonomy nodes as first categorization rules – apply and get content

Text Analytics Workshop
Text Analytics Development: Categorization Process

- Start: Term building – from content – basic set of terms that appear often / important to content
  - Auto-suggested and/or human generated
- Add terms to rule, get 90%+ recall
- Apply to broader set of content, build back up to 90%+
- Apply to new types of content – build precision -- Rules
- Repeat, refine, repeat, refine, repeat
- Develop logic templates
- Test against more, new content – add more terms, refine logic of rules
- Repeat until “done” – 90%?

File Edit View Build Project Category Concept Testing Document Server Help



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

Syntax Check | Indent | Text View | Tree View | Load Text...

Rules | Testing | Data | Document

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
        - Health

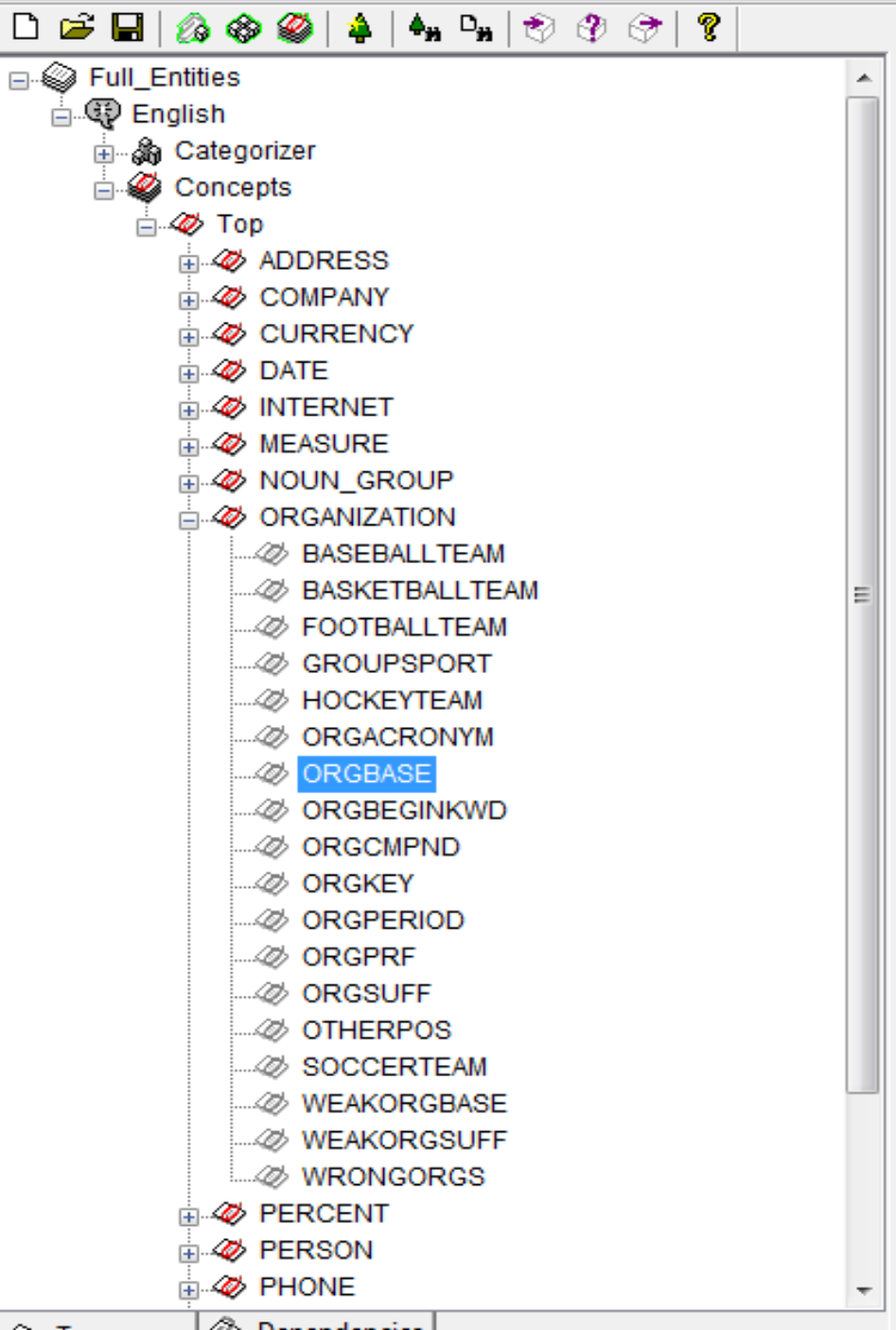
```
(START_3000,
(AND,
_/add/doc/docty:"Education Sector Review",

(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]")
)
)
)
```

Syntax Check | Indent | Text View / Tree View | Load Text... | Expand Forms | Server Que

Text Analytics Workshop
Development: Entity Extraction Process

- Facet Design – from Knowledge Audit, K Map
- Find and Convert catalogs:
  - 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




```
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
CLASSIFIER:BUREAU
CLASSIFIER:Caisse
CLASSIFIER:CAISSE
CLASSIFIER:Cathedral
CLASSIFIER:CATHEDRAL
CLASSIFIER:Center
CLASSIFIER:CENTER
CLASSIFIER:Centre
CLASSIFIER:CENTRE
```

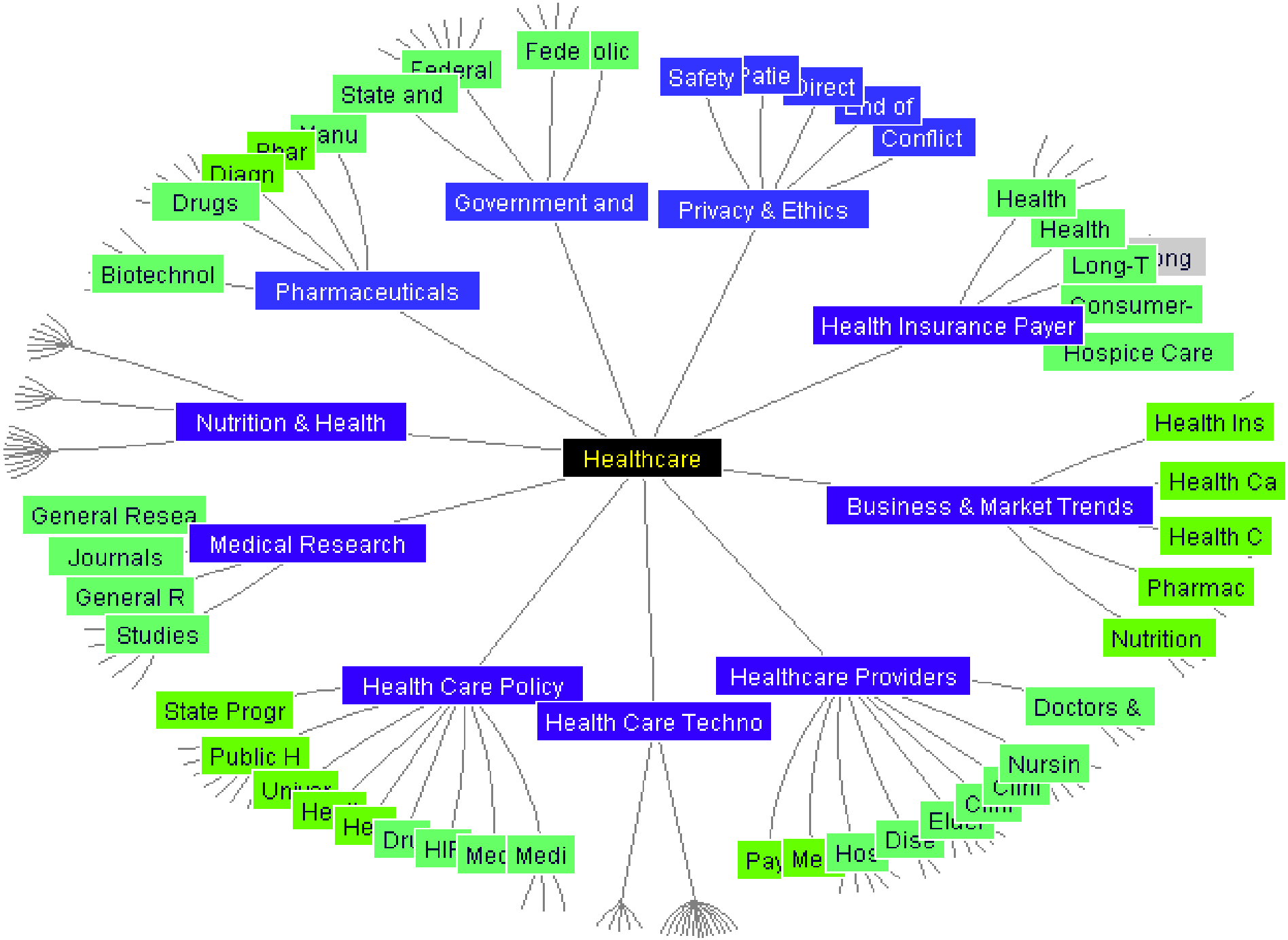
Syntax Check | Classifier | Filename | Grammar | LITI | Load Text...

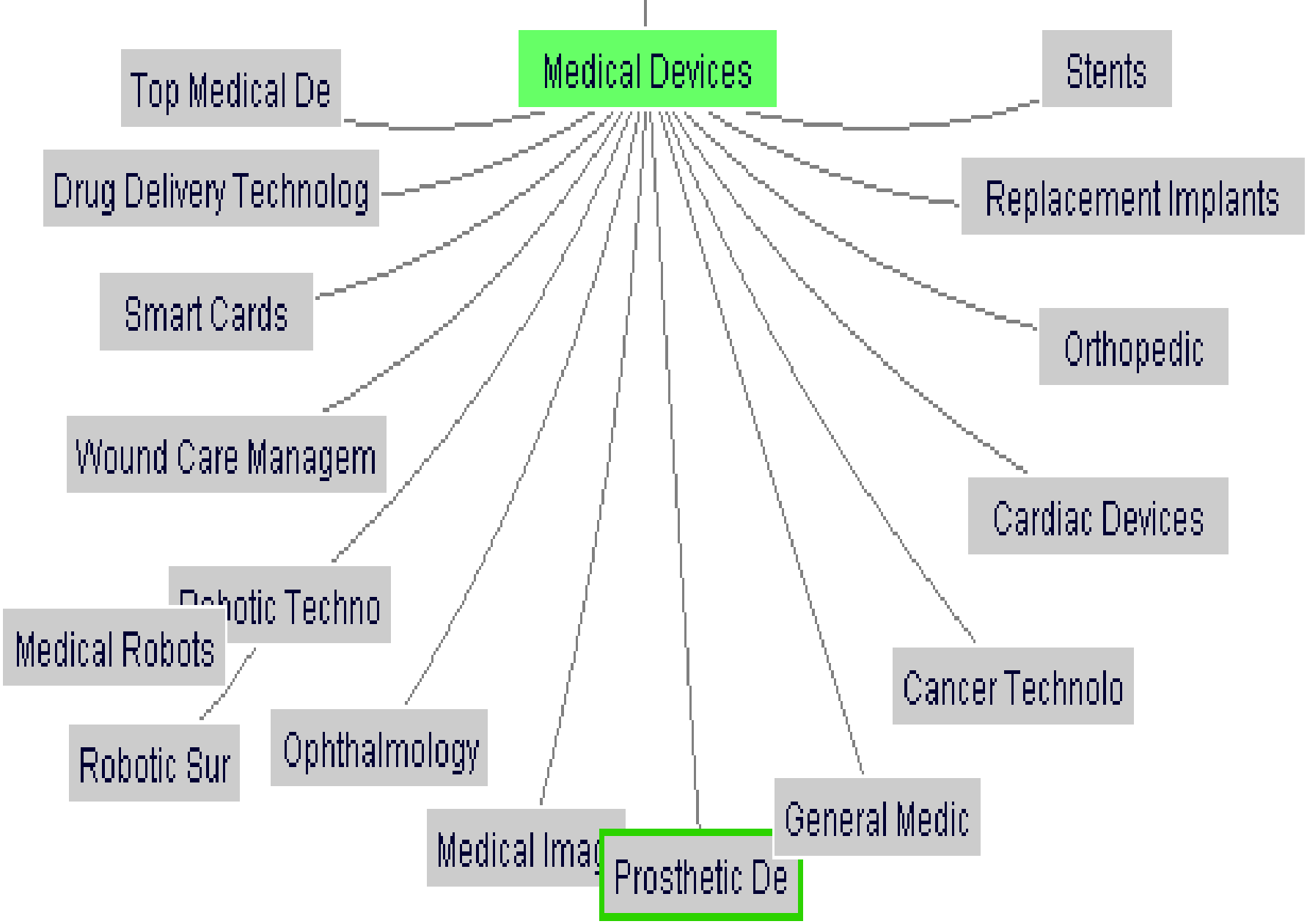
Text Analytics Workshop
Case Study - Background

- Inxight Smart Discovery
- Multiple Taxonomies
  - Healthcare – first target
  - Travel, Media, Education, Business, Consumer Goods,
- Content – 800+ Internet news sources
  - 5,000 stories a day
- Application – Newsletters
  - Editors using categorized results
  - Easier than full automation

Text Analytics Workshop
Case Study - Approach

- Initial High Level Taxonomy
  - Auto generation – very strange – not usable
  - Editors High Level – sections of newsletters
  - Editors & Taxonomy Pro's - Broad categories & refine
- Develop Categorization Rules
  - Multiple Test collections
  - Good stories, bad stories – close misses - terms
- Recall and Precision Cycles
  - Refine and test – taxonomists – many rounds
  - Review – editors – 2-3 rounds
- Repeat – about 4 weeks


Healthcare
Government and
Fede olic
Federal
State and
Safety Patie
Direct
End of
Conflict
Privacy & Ethics
Pharmaceuticals
Drugs
Biotechnol
Health Insurance Payer
Health
Health
Long-T
Consumer-
Hospice Care
Nutrition & Health
Business & Market Trends
Health Ins
Health Ca
Health C
Pharmac
Nutrition
Medical Research
General Resea
Journals
General R
Studies
Health Care Policy
Health Care Techno
Healthcare Providers
State Progr
Public H
Doctors &
Nursin
Medi


Medical Devices
Top Medical De
Stents
Drug Delivery Technolog
Replacement Implants
Smart Cards
Orthopedic
Wound Care Managem
Cardiac Devices
Robotic Techno
Medical Robots
Cancer Technolo
Robotic Sur
Ophthalmology
General Medic
Medical Imag
Prosthetic De

Category Properties

General | Definition | Exceptions

Category Threshold: 0.01

Terms | XDOCs | Rules

☑ Use Advanced Rules

```
"pipeline" AND NOT SENTENCE("oil" OR "gas")
```

Filter results from above

Must Include | Must not include

☑ Apply filter:

☐ Must include all filters

Filter	Case Sensitive
New England Journal ...	☐
Journal of the America...	☐
The Lancet American	☐
Journal of Medicine	☐
British Medical Journal	☐
Nature	☐

Text Analytics Workshop
Case Study – Issues & Lessons

- Taxonomy Structure: Aggregate vs. independent nodes
  - Children Nodes – subset – rare
- Trade-off of depth of taxonomy and complexity of rules
- No best answer – taxonomy structure, format of rules
  - Need custom development
  - Recall more important than precision – editors role
- Combination of SME and Taxonomy pros
  - Combination of Features – Entity extraction, terms, Boolean, filters, facts
- Training sets and find similar are weakest
- Plan for ongoing refinement

Text Analytics Workshop
Enterprise Environment – Case Studies

- A Tale of Two Catonomies
  - It was the best of times, it was the worst of times
- Basic Approach
  - Initial meetings – project planning
  - High level K map – content, people, technology
  - Contextual and Information Interviews
  - Content Analysis
  - Draft Taxonomy – validation interviews, refine
  - Categorization and entity extraction development
  - Integration and Governance Plans

Text Analytics Workshop
Enterprise – Case One – Taxonomy, 7 facets

- Taxonomy of Subjects / Disciplines:
  - Science > Marine Science > Marine microbiology > Marine toxins
- Facets:
  - Organization > Division > Group
  - Clients > Federal > EPA
  - Facilities > Division > Location > Building X
  - Content Type – Knowledge Asset > Proposals
  - Instruments > Environmental Testing > Ocean Analysis > Vehicle
  - Methods > Social > Population Study
  - Materials > Compounds > Chemicals

KAPS Group
Knowledge Architecture Professional Services

Text Analytics Workshop
Enterprise – Case One – Taxonomy, 7 facets

- Project Owner – KM department – included RM, business process
- Involvement of library - critical
- Realistic budget, flexible project plan
- Successful interviews – build on context
  - Overall information strategy – where taxonomy fits
- Good Draft taxonomy and extended refinement
  - Software, process, team – train library staff
  - Good selection and number of facets
- Developed broad categorization and one deep-Chemistry
- Final plans and hand off to client

Text Analytics Workshop
Enterprise – Case Two – Taxonomy, 4 facets

- Taxonomy of Subjects / Disciplines:
  - Geology > Petrology
- Facets:
  - Organization > Division > Group
  - Process > Drill a Well > File Test Plan
  - Assets > Platforms > Platform A
  - Content Type > Communication > Presentations

Enterprise – Case Two – Taxonomy, 4 facets Environment & Project Issues

- Value of taxonomy understood, but not the complexity and scope
  - Under budget, under staffed
- Location – RM software -Solution looking for the right problem
- No library involvement
- Project mind set – not infrastructure
  - Rushing to meet deadlines doesn't work with semantics
- Not enough research – and wrong people
- Not enough facets, wrong set – business not information
  - Ill-defined facets – too complex internal structure

Text Analytics Workshop: Applications

- 3 Main Types:
  - Search – An Enterprise Platform
  - Info Apps – Unstructured Text is Everywhere
  - Social Media – Fastest Growing Area

Text Analytics and Search
What's Wrong With Search?

- Search Engines are Stupid!
  - (and people have better things to do)
- Documents deal in language BUT it's all chicken scratches to Search
- Relevance – requires meaning
  - Imagine trying to understand what a document is about in a language you don't know
- Mzndin agenpfre napae ponaoen afpenafpenae timtnoe.
  - Dictionary of chicken scratches (variants, related)
  - Count the number of chicken scratches = relevance - Not
- Google = popularity of web sites and Best Bets
  - For documents in an enterprise – Counting and Weighting

Text Analytics and Search
Multi-dimensional and Smart

- 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
- People – tagging, evaluating tags, fine tune rules and taxonomy

Delve for the Web: The Front Page of Knowledge Management

Enterprise Knowledge Man

delvenews.com/news

Apps | Add to Delicious | Twitter | NPR News Apps | NP | Delve | News Roundups | Tools | DIY | Fantasy Baseball new | Notable sites | Other Bookmarks

DELVE | Search for Authors, Sources, and Topics | Home | Organization | Profile

DELVE **RECOMMENDATIONS** FOR YOU

ADJUST YOUR FEED

ELECTRIC VEHICLES, PACIFIC GAS AND ELECTRIC COMPANY...

SolarCity accuses utilities of slowing home-battery project

by DAVID R. BAKER in SAN FRANCISCO CHRONICLE | 7 HOURS AGO

The San Mateo company has installed battery packs in more than 100 houses throughout California, each pack linked to rooftop solar panels. The lithium-ion

SHARE

Who posted it:

@edgunther (Ed Gunther)

@scotthomasson (Scott Thomasson)

@katiefehren (Katie Fehrenbacher)

Users follow topics, people, and companies selected from Delve taxonomies.

BUSINESS & FINANCE, ELECTRIC VEHICLES...

Tesla Fights for a Place to Park

1

by JAD MOUAWAD in THE NEW YORK TIMES | 6 HOURS AGO

New Jersey is the latest state to bar Tesla's stores for its electric vehicles under laws that favor independent dealerships.

SHARE

Who posted it:

@greenwombat (Todd Woody)

ENERGY, ENERGY EFFICIENCY...

#AskEnergySaver: Answering Your Home-Energy Saving Questions

by U.S. DEPARTMENT in BREAKING ENERGY | 9 HOURS AGO

In 2012, the average American family spent 2.7 percent of their household income on home energy bills. While this might not sound like a lot, it adds up

Who posted it:

@BreakingEnergy (Breaking Energy)

INVITE COWORKERS

PASTE A LINK TO SHARE!

www.example.com/article

Tesla Fights for a Place to Park

SHARE

Social media data from Twitter powers recommendation algorithms.

ACTIVITY

Harrison Wells read

Tesla Fights for a Place to Park

Harrison Wells read

Montreal's 'Tube That Jean Built'

Shared by James Wyatt

Harrison Wells read

Renewable Solar & Wind Energy Produced As Much As 60% Of Germany's Electricity October 3rd

Shared by James Wyatt

Harrison Wells read

Enterprise Knowledge Man

getwiser.com/news#

WISER | Search for Topics, Sources, and Keywords | My Feed | Organization | Profile

Sandra Wagner

Paste a link

www.example

I'm Following

- Woods & Co.
- James Wyatt
- Affordable Care Act
- Antibacterials
- Clinical Care
- Dual Eligible Seniors
- Electronic Health Records
- Employee Health Insurance
- Evidence Based Medicine
- Global Health

Expand

Trending at Woods & Co.

Shared by Wen Lee • 4 • 2 • 9

HOSPITALS' PURCHASE OF DOCTORS LEADS TO HIGHER PRICES, SPENDING, STUDY FINDS

Kaiser Health News • *Phil Galewitz*

Shared by Wen Lee • 3 • 6

HOSPITALS OVERCHARGE MED RECORDS BY $7M

Healthcare IT News • *Erin McCann*

Shared by Wen Lee • 2 • 1 • 3

BAYER AND MERCK IN $14.2B CONSUMER DEAL

CNBC

Most Recent Activity

You shared

And now I have to defend doctors' salaries, a

We Recommend

Shared by You •

HEALTH ECONOMICS

AND NOW I HAVE TO DEFEND DOCTORS' SALARIES, A LITTLE

by *Aaron Carroll* in *The Incidental Economist* • a day ago

Sarah Kliff has a post up with some up-to-the-minute data on how much doctors make. The first chart, showing doctors' starting salaries, is the kind of graphic that's going to make a lot of doctors get quickly defensive:

SHARED

Leave a comment

COMMENT

Outside signals:

@sarahkliff (Sarah Kliff)

@Health Finance | ADD TAGS

HEALTH ECONOMICS

HOSPITAL READMISSIONS ARE DOWN, BUT ARE THEY APPROPRIATELY MEASURED?

by *Austin Frakt* in *The Incidental Economist* • a day ago

The Department of Health and Human Services (HHS) released some news that suggests patients are receiving better care from hospitals:* The data in this report shows a substantial nine percent decrease in harms

SHARE

Outside signals:

@jordanrau (Jordan Rau)

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 is the answer(s)!

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
  - Target usually more general

Text Analytics Workshop
Adding Structure to Unstructured Content

- Beyond Documents – categorization by corpus, by page, sections or even sentence or phrase
- Documents are not unstructured – variety of structures
  - Sections – Specific - “Abstract” to Function “Evidence”
  - Corpus – document types/purpose
  - Textual complexity, level of generality
- Need to develop flexible categorization and taxonomy – tweets to 200 page PDF
- Applications require sophisticated rules, not just categorization by similarity

- 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
          - BPH
          - Cancer
          - Clondine Disease
          - HIV
          - Pulmonary
          - Thrombosis

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

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

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

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

Syntax Check | Indent | Text View / Tree View | Load Text... | Expand Forms | Server Query...

Taxonomy | Dependencies | Rules | Testing | Data | Document

Text Analytics Workshop
Document Type Rules

- (START_2000, (AND, (OR, _/article:"[Abstract]", _/article:"[Methods]“), (OR,_/article:"clinical trial*", _/article:"humans",
- (NOT, (DIST_5, (OR,_/article:"approved", _/article:"safe", _/article:"use", _/article:"animals"),
- If the article has sections like Abstract or Methods
- AND has phrases around “clinical trials / Humans” and not words like “animals” within 5 words of “clinical trial” words – count it and add up a relevancy score
- Primary issue – major mentions, not every mention
  - Combination of noun phrase extraction and categorization
  - Results – virtually 100%

Text Analytics Workshop
Enterprise Info Apps

- Focus on business value, cost cutting, new revenues
- Business Intelligence
  - It is a growing field with revenues of $13.1 billion in 2012.
  - Early identification of product issues
  - What are competitors doing
  - Integrate data and text
- Financial Services
  - Trend – using text analytics with predictive analytics – risk and fraud
  - Combine unstructured text (why) and structured transaction data (what)
  - Customer Relationship Management, Fraud Detection
  - Stock Market Prediction – Twitter, impact articles

Text Analytics Workshop
Enterprise Info Apps

- eDiscovery,
  - Collect all documents about a particular situation (Search)
  - Reduce human effort, add intelligence to selection
  - Payoff is big – One firm with 1.6 M docs – saved $2M
- Text Analytics Assisted Review
  - Scan millions of documents for indications of revenue
- AI Headhunters
  - Scan resumes for red and green flags
- Automatic Summaries
  - Extract key data – disambiguation, co-reference
  - Create story summaries – baseball game, finance

Text Analytics Workshop
Applications: KM

- Knowledge Management as if knowledge matters
  - Not sterile DIKW or Tacit debates, but practical
- Past – taxonomy and metadata have failed
  - Too rigid, too dumb, too much effort
- Text Analytics changes that
  - Flexible, smart, support human efforts
- KM and Social Media
  - Text Analytics adds a new dimension to expertise location Communities of Practice, collaboration

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

Text Analytics Workshop: Applications Expertise Analysis

- Mid-level in a taxonomy / hierarchy
- Short and easy words
- Maximum distinctness and expressiveness
- First level named and understood by children
- Level at which most of our knowledge is organized
- Levels: Superordinate – Basic – Subordinate
  - Mammal – Dog – Golden Retriever
  - Furniture – chair – kitchen chair

Text Analytics Workshop: Applications
Expertise Analysis

- Experts prefer lower, subordinate levels
  - In their domain, (almost) never used superordinate
- Novice prefer higher, superordinate levels
- General Populace prefers basic level
- Not just individuals but whole societies / communities differ in their preferred levels
- Issue – artificial languages – ex. Science discipline
- Issue – difference of child and adult learning – adults start with high level

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
Deep Learning

- Neural Networks – from 1980's
- New = size and speed
- Larger Networks = can learn better and faster
- Multiple networks = more automatic – networks learn from other networks
- Strongest in areas like image recognition
- Next is entity / fact extraction & discovering relationships
- Weakest – concepts, subjects, deep language, metaphors, etc.

Social Media Applications Beyond Simple Sentiment

- Beyond Good and Evil (positive and negative)
  - Degrees of intensity, complexity of emotions and documents
- Importance of Context – around positive and negative words
  - Rhetorical reversals – “I was expecting to love it”
  - Issues of sarcasm, (“Really Great Product”), slanguage
- Essential – need full categorization and concept extraction
- New Taxonomies – Appraisal Groups – “not very good”
  - Supports more subtle distinctions than positive or negative
- Emotion taxonomies - Joy, Sadness, Fear, Anger, Surprise, Disgust
  - New Complex – pride, shame, confusion, skepticism
- New conceptual models, models of users, communities

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.
  - Used at a high rate, short and hard to detect, very social, processed in the brain differently than content words
- 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

Text Analytics Workshop Conclusions

- Text Analytics needs strategic vision and quick start
  - But also concrete and quick application to drive acceptance
- Text Analytics is the mechanism to finally fix search
  - And get value from taxonomies, metadata, content management, etc.
- Two major techniques
  - Deep Text – depth and intelligence
  - Deep Learning – power and scale, learning
- Integration of the two = the future
- Future – Text Analytics and Cognitive Science = Metaphor Analysis, deep language understanding, AI, common sense?

Questions?

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

KAPS Group
Knowledge Architecture Professional Services

Resources

- Books
  - Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data
    - Tom Reamy
  - Women, Fire, and Dangerous Things
    - George Lakoff
  - Knowledge, Concepts, and Categories
    - Koen Lamberts and David Shanks
  - Thinking Fast and Slow
    - Daniel Kahneman
  - Any cognitive science book written after 2010

Resources

- Conferences – Web Sites
  - Text Analytics World - All aspects of text analytics
    - http://www.textanalyticsworld.com
  - Text Analytics Summit
    - http://www.textanalyticsnews.com
  - Semtech
    - http://www.semanticweb.com
  - Sentiment Analysis Symposium
    - www.sentimentsymposium.com
  - New Text Analytics Conference - 2017

Resources

- LinkedIn Groups:
  - Text Analytics, Text Analytics World
  - Taxonomy Community of Practice
  - Sentiment Analysis
  - Text and Social Analytics
  - Metadata Management
  - Semantic Technologies, Semantic Web
  - Association for Information Science & Technology
- Journals
  - Academic – Cognitive Science, Linguistics, NLP
  - Applied – Scientific American Mind, New Scientist

Text Analytics Workshop

The start and foundation: Knowledge Architecture Audit

- Knowledge Map - Understand what you have, what you are, what you want
  - The foundation of the foundation
- Contextual interviews, content analysis, surveys, focus groups, ethnographic studies, Text Mining
- Category modeling – “Intertwingledness” -learning new categories influenced by other, related categories
  - Monkey, Panda, Banana
- Natural level categories mapped to communities, activities
  - Novice prefer higher levels
  - Balance of informative and distinctiveness
- 4 Dimensions – Content, People, Technology, Activities

Text Analytics Workshop
Knowledge Audit: Contextual Interviews

- Organizational Context – Free Form
  - Management, enterprise wide function
  - What is the size and makeup of the organizational units that will be impacted by this project?
  - Are there special constituencies that have to be taken into account?
  - What is the level of political support for this project? Any opposition?
  - What are your major information or knowledge access issues?
- These determine approach and effort for each area

Text Analytics Workshop
Knowledge Audit: Information Interviews

- Structured, feed survey – list options
  - Could you describe the kinds of information activities that you and your group engage in? (types of content, search, write proposals, research?) How often?
  - How do they carry out these activities?
- Qualitative Research
  - What are your major information or knowledge access issues -- examples?
  - In an ideal world, how would information access work at your organization?
  - What is right and what's wrong with today's methods
- Output = map of information communities, activities

Text Analytics Workshop
Knowledge Audit: Map of Information Technology

- Content Management – ability to integrate text analytics
- Search – Integration of text analytics – Beyond XML
  - Metadata – facets
- Existing Text Analytics – Underutilization?
  - Text Mining – often separate silo, how integrate?
- Taxonomy Management, Databases, portals
  - Semantic Technologies, Wiki's
- Visualization software
  - Applications – business intelligence, customer support, etc.
- Map- often reveals multiple redundancies, technology silos

Text Analytics Workshop
Knowledge Audit: Content Analysis

- Content Map – size, format, audience, purpose, priority, special features, data and text, etc.
- Content Creation – content management workflow and real life workflow, publishing process – policy
  - Integrate external content – little control, massive scale
- Content Structure –taxonomies, vocabularies, metadata standards
- Drill Down, theme discovery
  - Search log analysis
  - Folksonomy if available
  - Text Mining, categorization exploration, clustering

Text Analytics Workshop
Knowledge Audit- Output

- Strategic Vision and Change Management
  - Format – reports, enterprise ontology
  - Political/ People and technology requirements
- Business Benefits and ROI
  - Enterprise Text Analytics- information overload – IDC study:
    - Per 1,000 people = $ 22.5 million a year
    - 30% improvement = $6.75 million a year
  - Add own stories – especially cost of bad information, cost cutting
- Strategic Project Plan and Road Map
  - Text Analytics support requirements –taxonomies, resources
  - Map of Initial Projects – and selection criteria