Deep Text Social Media Analysis A Text Analytics Foundation

Tom Reamy
Chief Knowledge Architect

KAPS Group

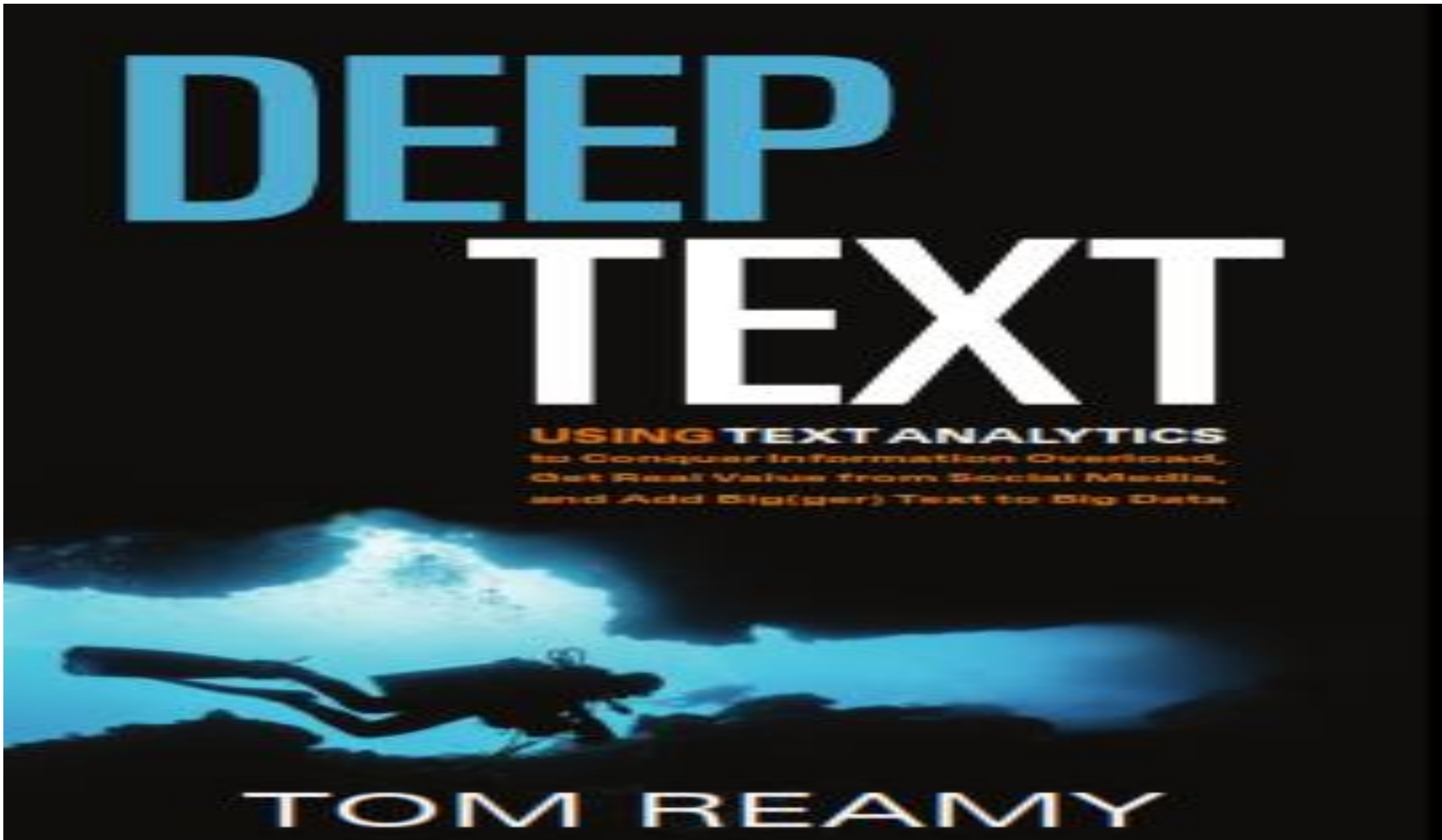
http://www.kapsgroup.com

Author: Deep Text

Agenda

- Introduction
  - What is Deep Text Analytics – Definition & Elements
- Smart Social Media Analysis
- Key Concepts
  - Basic Level Categories and Expertise Analysis
- Development - Approaches
- Applications
- Getting Started with Text Analytics
- Questions / Discussions

A treasure trove of technical detail, likely to become a definitive source on text analytics – *Kirkus Reviews (available 30% off)*



Introduction:
Deep Text: The Book

- The only book on text analytics
- 5 sections, 3 chapters each
  - Text Analytics Basics
  - Getting Started in Text Analytics (Smart Start)
  - Text Analytics Development
  - Text Analytics Applications
  - ETA – Enterprise Text Analytics as a Platform
- A treasure trove of technical detail, likely to become a definitive source on text analytics. – Kirkus Reviews
- This book will give you all the answers and is the definitive book on the business possibilities of the technology**.** - Martin White

Introduction:
Deep Text: The Book – Who Am I?

- Professional student / independent consultant – all but 6 years
- History of Ideas to Programmer – AI (Only 2 years away)
- Games – Galactic Gladiators/Adventures – still available
- KAPS Group – 13 years, Network of consultants
  - Taxonomy to text analytics
  - Consulting, development – platform and applications
  - Strategy, Smart Start, Search, Smart Social Media
  - Partners – SAS, IBM, Synaptica, Expert System, Smartlogic, etc.
  - 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

KAPS Group
Knowledge Architecture Professional Services

Introduction: What is Text Analytics?

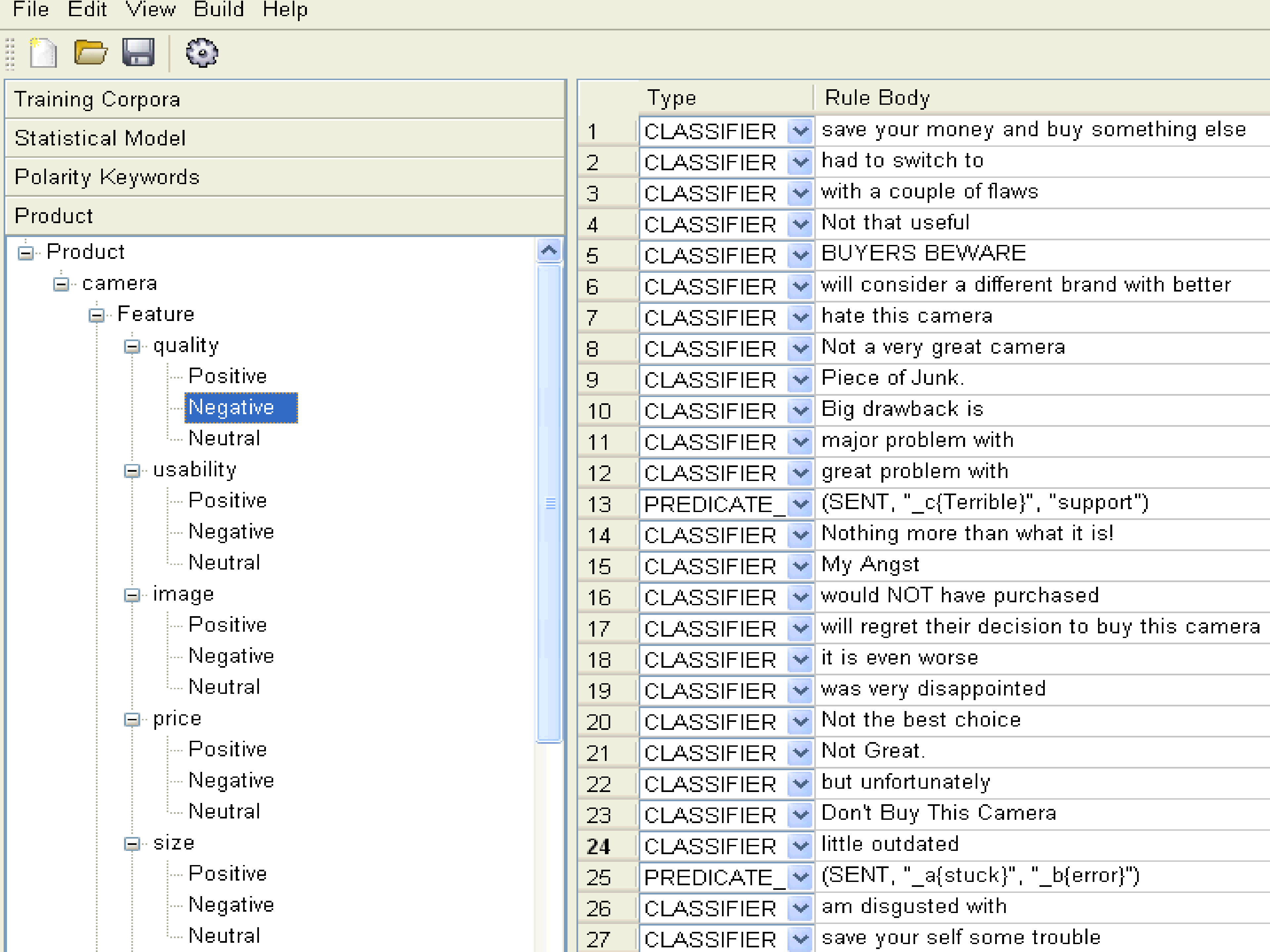
- Text analytics is the use of software and knowledge models to analyze and add structure to unstructured text.
- Text Mining – NLP, statistical, predictive, machine learning
  - Different skills, mind set, Math & data not language
- Annotation/Extraction – entities and facts – known and unknown, concepts, events - catalogs with variants, rule based
- Sentiment Analysis
  - Entities and sentiment words – statistics & rules
- Summarization
  - Dynamic – based on a search query term
  - Document – based on primary topics, position in document

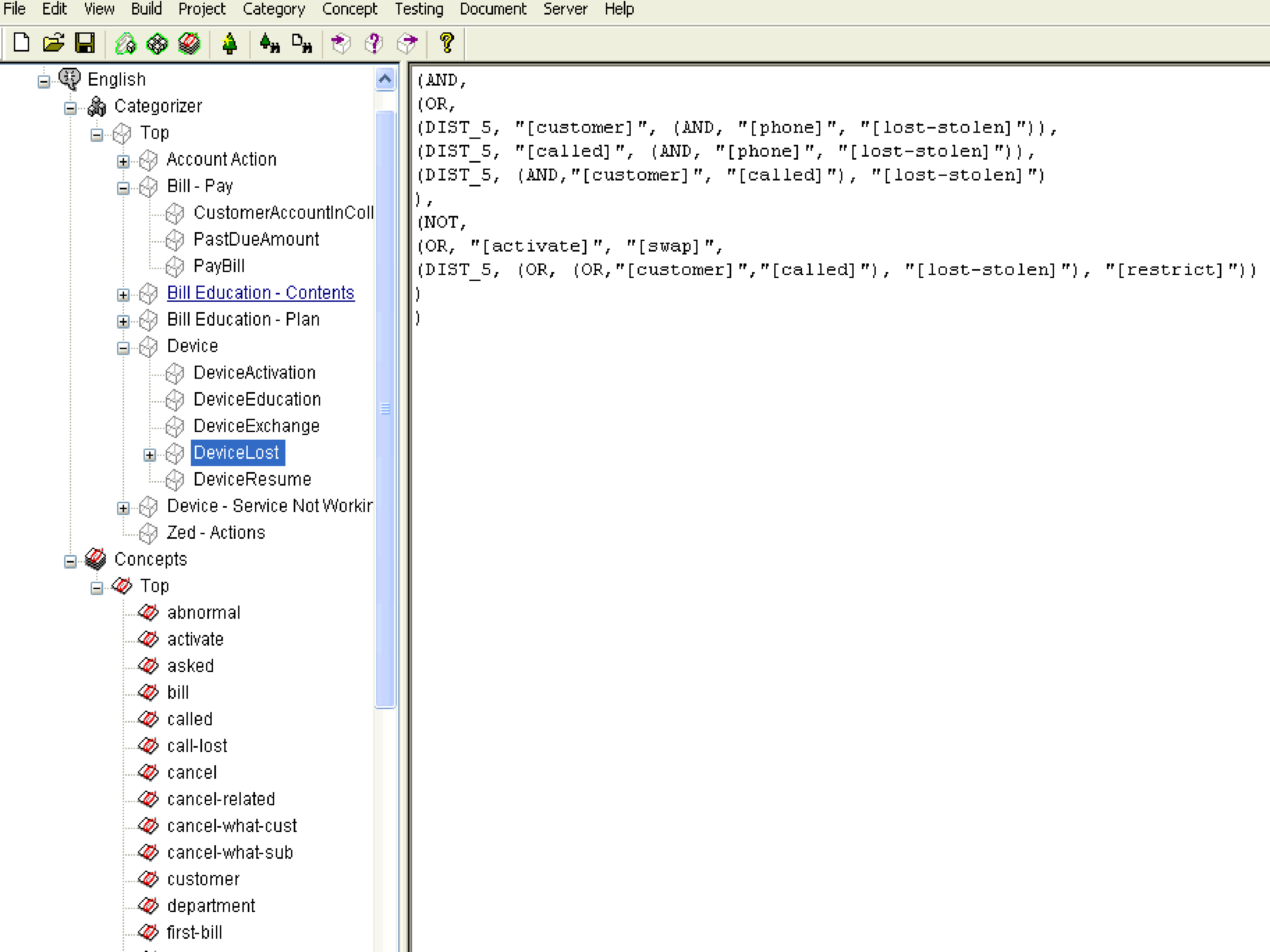
Introduction:
What is Text Analytics?

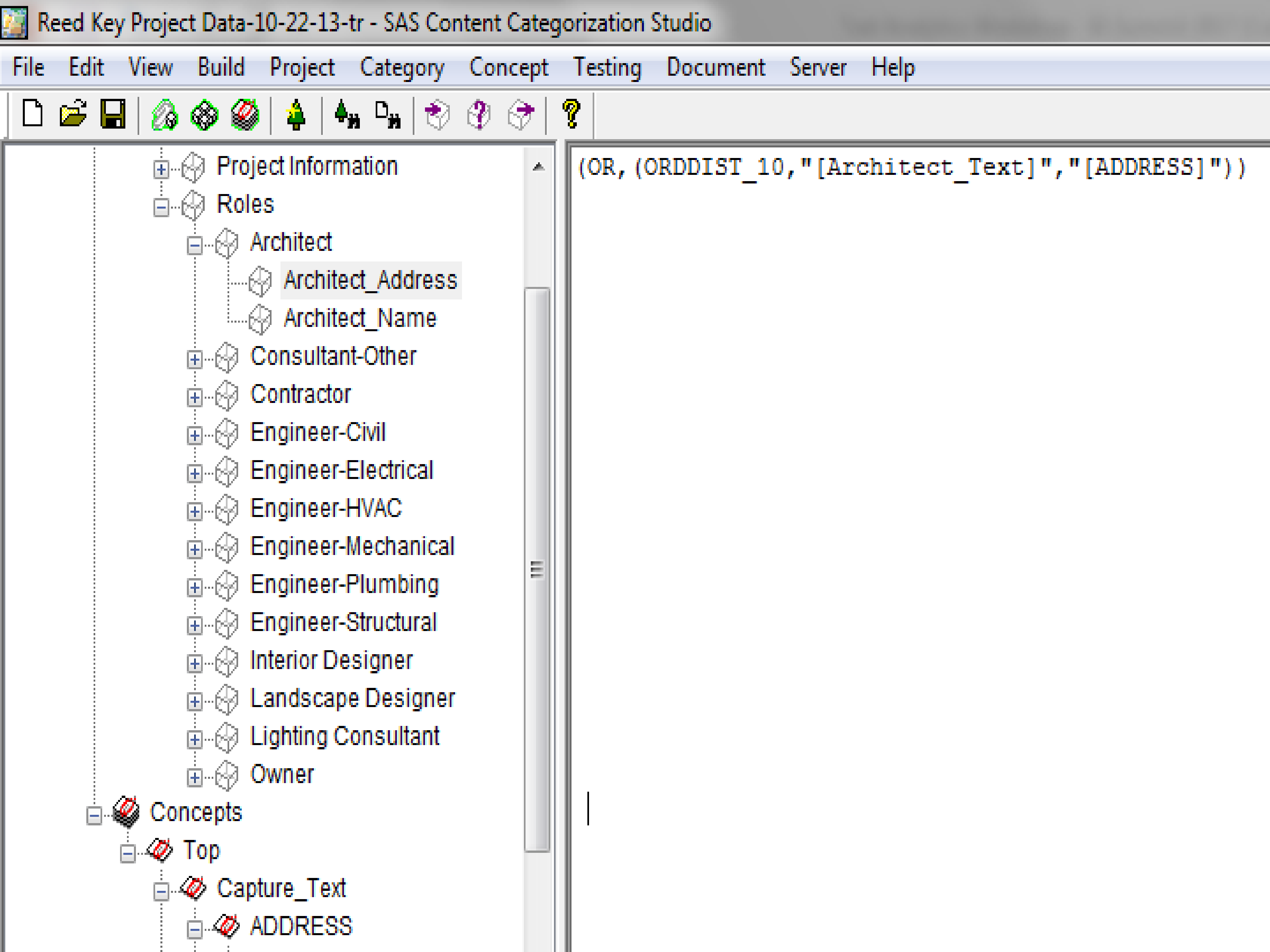
- Auto-categorization = the brains of the outfit
  - Training sets – Bayesian, Vector space
  - Terms – literal strings, stemming, dictionary of related terms
  - Rules – simple – position in text (Title, body, url)
  - Boolean– Full search syntax – AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
- Auto-categorization = smart social media analysis

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				


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


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]"))
)
)


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]"))

Introduction: What is Text Analytics?

- History – Inxight - Moved TA from academic and NLP to enterprise - auto-categorization, entity extraction, and Search-Meta Data
- Shift to sentiment analysis - easier to do, obvious pay off
  - Backlash – Real business value?
- Current Market: 2016 – 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.

Smart Sentiment Analysis

Smart Sentiment Analysis
Sentiment & Categorization

- Beyond Good and Evil (positive and negative)
  - Taxonomy of Objects and Features to taxonomy of emotions
  - Addition of focus on behaviors – why someone calls a support center – and likely outcomes
- Emphasis on context around positive and negative words
  - Issue of sarcasm, slanguage – “Really great product”
  - Rhetorical reversals – “I was expecting to love it”
- Limited value of Positive and Negative
  - Degrees of intensity, complexity of emotions and documents
  - Granularity of Application – early categorization

Smart Sentiment Analysis
Sentiment & Categorization

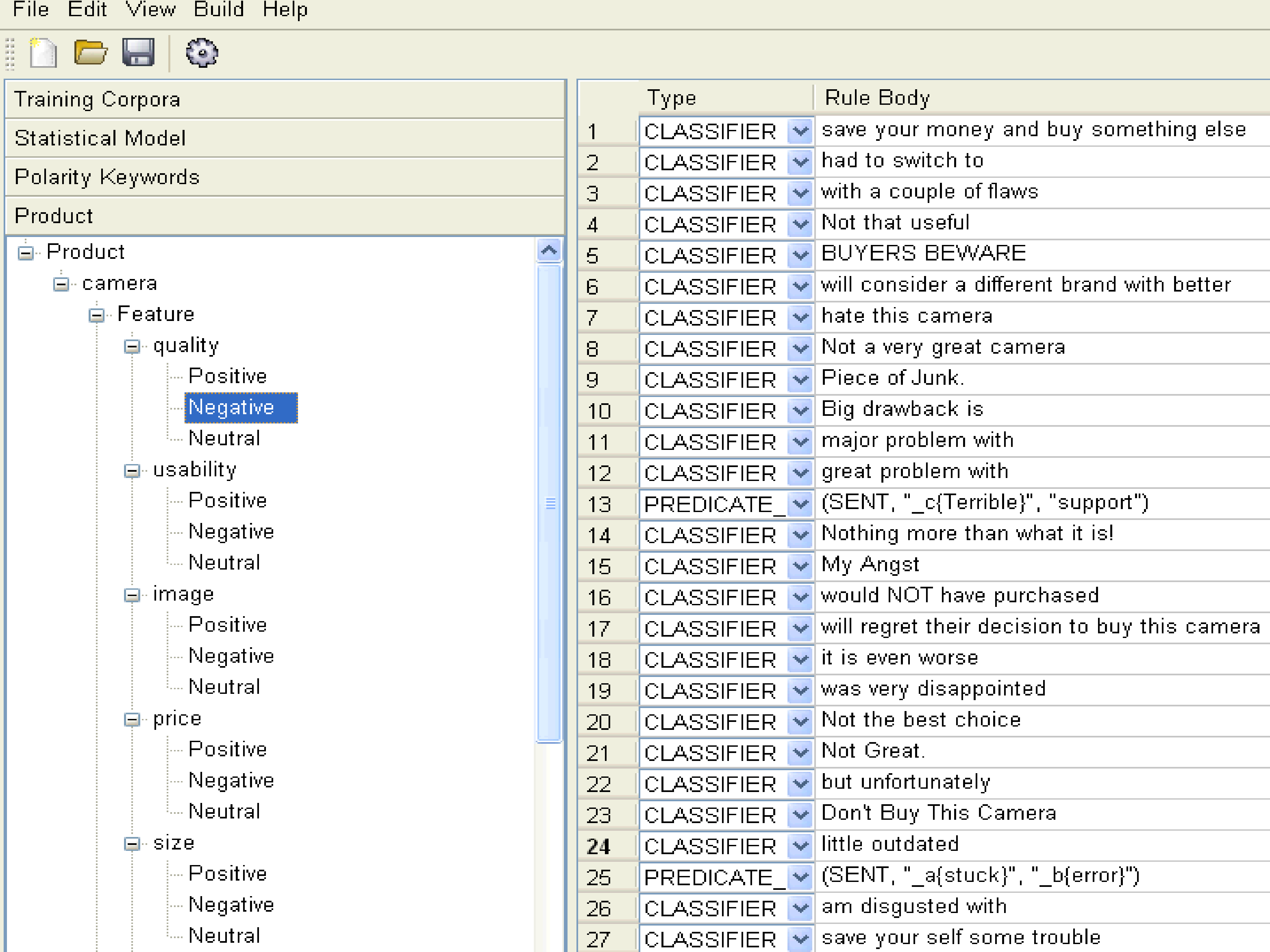
- Two flawed approaches: Lack of Accuracy, Depth
  - Statistical Signature of Bag of Words
  - Dictionary of positive & negative words
- Essential – need full categorization and concept extraction to do sentiment analysis well
- Categorization
  - Most basic to human cognition
  - Most difficult to do with software
- No single correct categorization
  - Women, Fire, and Dangerous Things

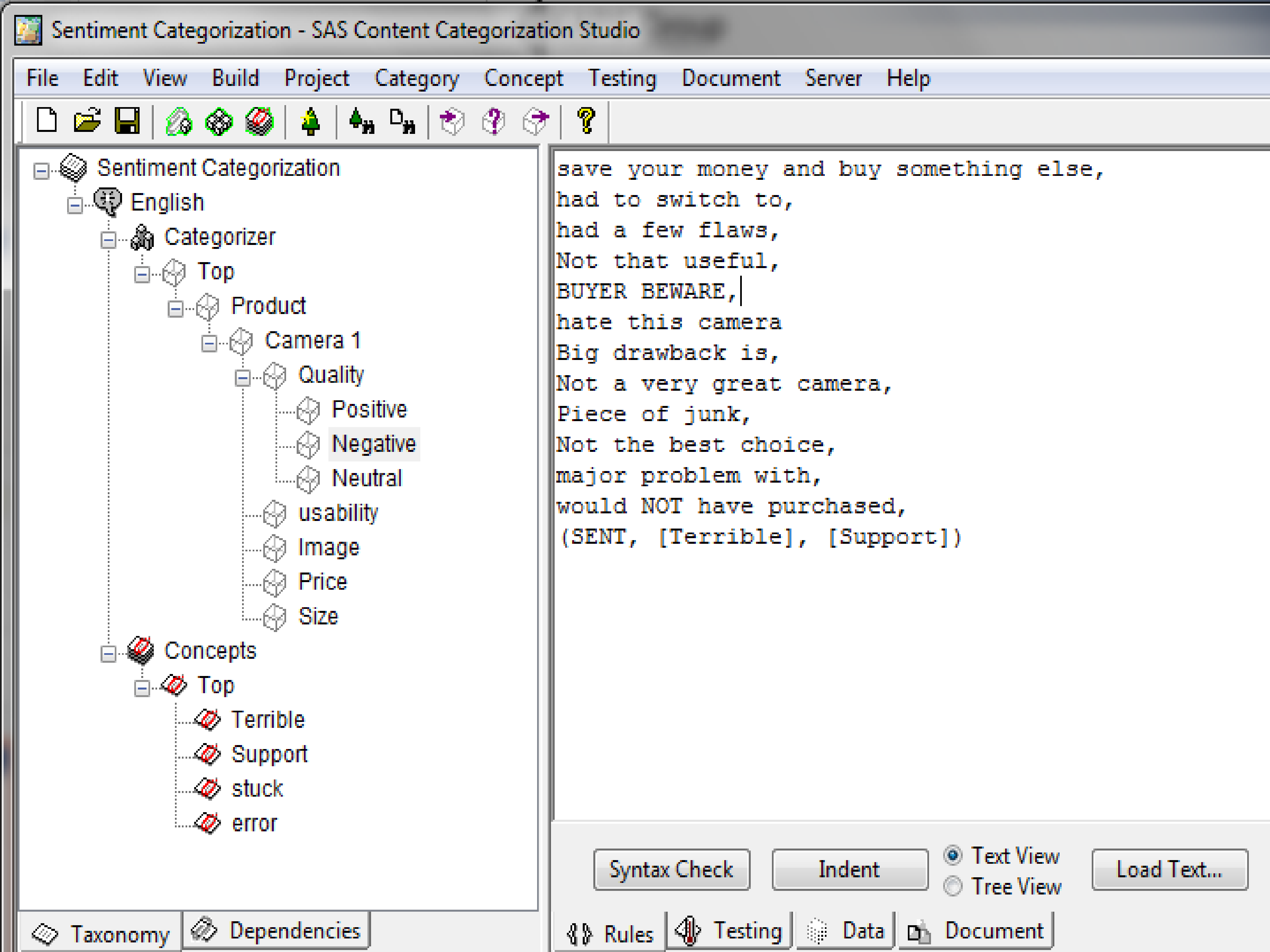
Smart Sentiment Analysis
Sentiment & Categorization

- Borges – Celestial Emporium of Benevolent Knowledge
  - Those that belong to the Emperor
  - Embalmed ones
  - Those that are trained
  - Suckling pigs
  - Mermaids
  - Fabulous ones
  - Stray dogs
  - Those that are included in this classification
  - Those that tremble as if they were mad
  - Innumerable ones
  - Other

Smart Sentiment Analysis
New Taxonomies

- New Taxonomies – Appraisal
  - Appraisal Groups – Adjective and modifiers – “not very good”
  - Four types – Attitude, Orientation, Graduation, Polarity
  - Supports more subtle distinctions than positive or negative
- Emotion taxonomies
  - Joy, Sadness, Fear, Anger, Surprise, Disgust
  - New Complex – pride, shame, embarrassment, love, awe
  - New situational/transient – confusion, concentration, skepticism
- Beyond Keywords
  - Analysis of phrases, multiple contexts – conditionals, oblique
  - Analysis of conversations – dynamic of exchange, private language


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


Sentiment Categorization - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Sentiment Categorization
English
Categorizer
Top
Product
Camera 1
Quality
Positive
Negative
Neutral
usability
Image
Price
Size
Concepts
Top
Terrible
Support
stuck
error
Taxonomy
Dependencies
save your money and buy something else,
had to switch to,
had a few flaws,
Not that useful,
BUYER BEWARE,
hate this camera
Big drawback is,
Not a very great camera,
Piece of junk,
Not the best choice,
major problem with,
would NOT have purchased,
(SENT, [Terrible], [Support])
Syntax Check
Indent
Text View
Tree View
Load Text...
Rules
Testing
Data
Document

Deep Text Social Media Analysis
Key Concepts

– Basic level categories
– Expertise

Deep Text Social Media Analysis
Basic Level Categories

- Mid-level in a taxonomy / hierarchy
- Levels: Superordinate – Basic – Subordinate
  - Mammal – Dog – Golden Retriever
  - Furniture – chair – kitchen chair
- Basic in 4 dimensions
  - Perception – overall perceived shape, single mental image, fast identification
  - Function – general motor program
  - Communication – shortest, most commonly used, neutral, first learned by children
  - Knowledge Organization – most attributes are stored at this level - Maximum distinctness and expressiveness

Deep Text Social Media Analysis
How recognize Basic level

- Cue Validity – probability that a particular object belongs to some category given that it has a particular feature (cue)
  - X has wings – bird
- Short words – noun phrase
- Kinds of attributes
  - Superordinate – functional (keeps you warm, sit on it)
  - Basic – Noun and adjectives – legs, belt loops, cloth
  - Subordinate – adjectives – blue, tall
- More complex for non-object domains
- Issue – what is basic level is context dependent

Deep Text Social Media Analysis
Other levels

- Subordinate – more informative but less distinctive
  - Basic shape and function with additional details
    - Ex – Chair – office chair, armchair
  - Convention – people name objects by their basic category label, unless extra information in subordinate is useful
- Superordinate – Less informative but more distinctive
  - All refer to varied collections – furniture
  - Often mass nouns, not count nouns
  - List abstract / functional properties
  - Very hard for children to learn

Deep Text Social Media Analysis
Basic Level Categories and Expertise

- Experts prefer lower, subordinate levels
  - Novice prefer higher, superordinate levels
  - General Populace prefers basic level
- Expertise Characterization for individuals, communities, documents, and sets of documents
- Experts chunk series of actions, ideas, etc.
  - Novice – high level only
  - Intermediate – steps in the series
  - Expert – special language – based on deep connections

Deep Text Social Media Analysis
Expertise Analysis: Analytical Techniques

- Corpus context dependent
  - Author748 – is general in scientific health care context, advanced in news health care context
- Need to generate overall expertise level for a corpus
- Develop expertise rules – similar to categorization rules
  - Use basic level for subject
  - Superordinate for general, subordinate for expert
- Also contextual rules
  - “Tests” is general, high level
  - “Predictive value of tests” is lower, more expert
  - Not counting “big words” – essay evaluation debacle

Deep Text Social Media Analysis Development

Deep Text Social Media Analysis Development: Deep Text vs. Deep Learning

- Two Schools – Language Rules vs. Math / Patterns
  - Depth & Intelligence vs. Speed & Power
  - Two systems in the brain – system 1 and 2
- Deep Learning
  - Neural Networks – from 1980’s, new = size and speed
  - Strongest in areas like image recognition, fact lookup
  - Weakest – concepts, subjects, deep language, metaphors, etc.
- Deep Text – Language, concepts, symbols
  - Categorization & Generalization – most basic to human cognition
    - Beyond Categorization – making everything else smarter
  - Rules = higher accuracy – 98% - Rules brittle?

- Boehringer Pilot One Drug Names Diseases
  - English
    - Categorizer
      - Top
        - Diseases
          - arthritis
          - Benign Prostatic Hyperplasia
          - Cancer
          - Deep Vein Thrombosis
          - HIV
          - Hypertension
          - Pulmonary Disease
        - Drug Names
          - afatinib

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

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

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

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

Deep Text Social Media Analysis
Deep Text vs. Deep Learning

- Deep Learning is a Dead End - accuracy – 60-70%
  - Black Box – don't know how to improve except indirect manipulation of input – "We don't know how or why it works"
  - Domain Specific, tricks not deep understanding
  - No common sense and no strategy to get there
  - Major – loss of quality – who is training who?
  - Need millions of examples – not human-like learning
- AI should be AAI – Artificial Animal Intelligence – patterns not concepts
- Extra Benefits of a Deep Text Approach – Multiple InfoApps
- Future = Interpenetration of Opposites
  - Make Deep Learning smarter, add learning to Deep Text

Text Analytics Development: Categorization Process
Start with Taxonomy and Content

- Starter Taxonomy
  - If no taxonomy, develop (steal) initial high level
  - Library of semantic resources – templates, catalogs, data
- 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
- Start: taxonomy as initial categorization
- Term building – from content – basic set of terms that appear often / important to content
  - Auto-suggested and/or human generated
- Cycles: test set, recall, precision -> more content
- Rule templates, sectionize documents

Deep Text Social Media Analysis Development: Entity Extraction Process

- Entity Categories – from Knowledge Audit, K Map
- Find and Convert catalogs:
  - Organization – internal resources
  - People – corporate yellow pages, HR
  - Linked Data – more data, less control
  - Scripts to convert catalogs – programming resource
- Build initial rules – follow categorization process
  - Differences – scale, threshold – application dependent
  - Recall – Precision – balance set by application
  - Issue – disambiguation – Ford company, person, car
- Unknown entities – NLP rules – “cap cap said”

KAPS Group
Knowledge Architecture Professional Services

Deep Text Social Media Analysis Applications

KAPS Group
Knowledge Architecture Professional Services

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
- New techniques beyond counting pos. & neg.
  - Context, intensity, new models of emotions
  - New conceptual models, models of users, communities

Deep Text Social Media Analysis
Expertise Analysis: Application areas

- Business & Customer intelligence
  - General – characterize people's expertise to add to evaluation of their comments
- Combine with VOC & sentiment analysis – finer evaluation – what are experts saying, what are novices saying
- Social Media - Community of Practice
  - Characterize the level of expertise in the community
- Expertise location
  - Generate automatic expertise characterization based on authored documents

Deep Text Social Media Analysis Enterprise Info Apps

- Focus on business value-new revenues not cost cutting
- Business Intelligence
  - Early identification of product issues
  - What are competitors doing
  - Integrate data and text
- Financial Services
  - 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

Deep Text Social Media Analysis
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
- Enterprise Social Networks – growing
- Automatic Summaries
  - Extract key data – disambiguation, co-reference
  - Create story summaries – baseball game, finance

Social Media Applications
Voice of the Customer / Voter / Employee

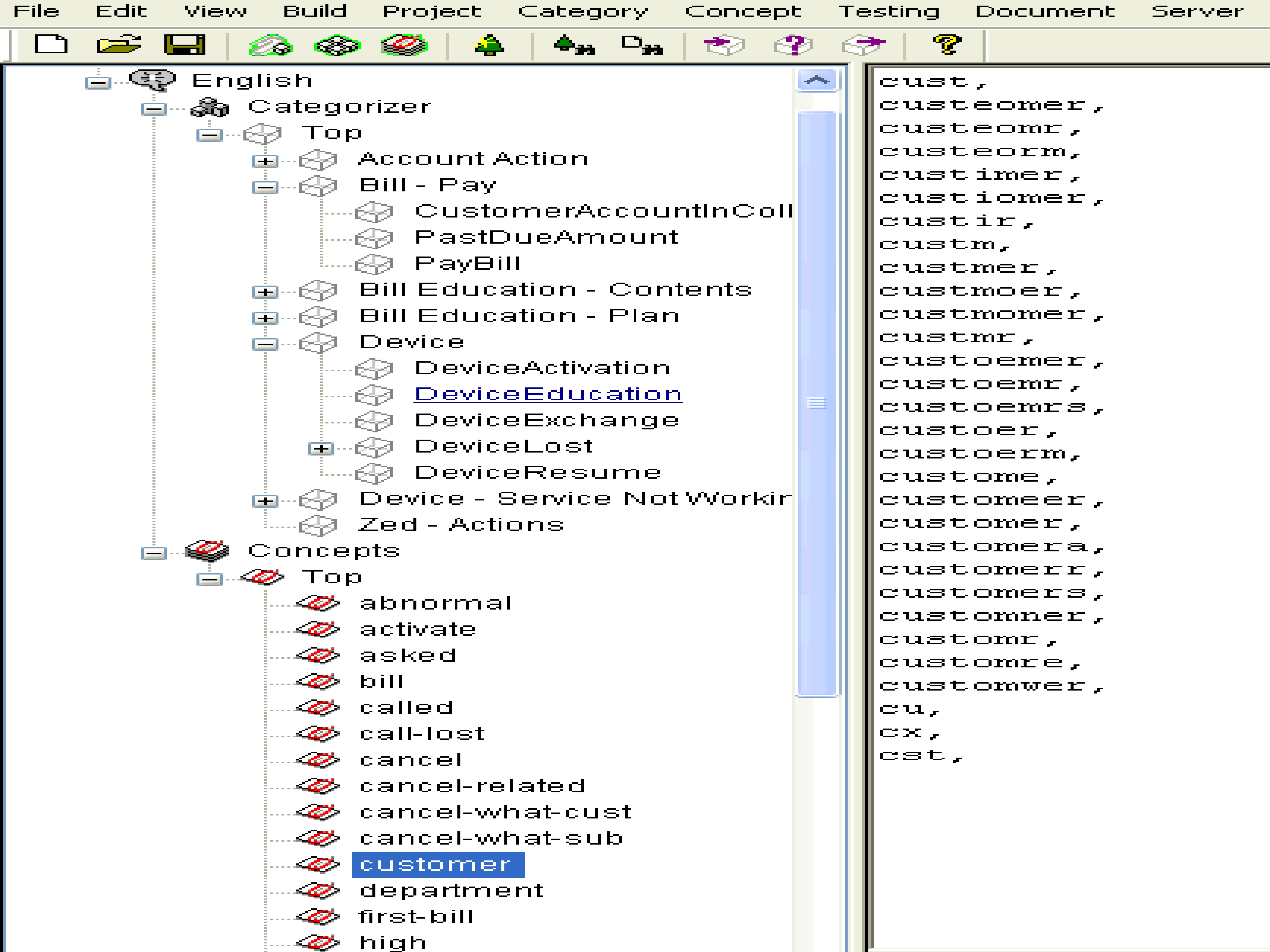
- Detection of a recurring problem categorized by subject, customer, client, product, parts, or by representative.
- Analytics to evaluate and track the effectiveness:
  - Representatives, policies, programs, actions
- Competitive intelligence – calls to switch from brand X to Y in a particular region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

New Applications in Social Media
Behavior Prediction – Telecom Customer Service

- Problem – distinguish customers likely to cancel from mere threats
- Analyze customer support notes
- General issues – creative spelling, second hand reports
- Develop categorization rules
  - First – distinguish cancellation calls – not simple
  - Second - distinguish cancel what – one line or all
  - Third – distinguish real threats

New Applications in Social Media
Behavior Prediction – Telecom Customer Service

- 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.
  - cci and **is upset that he has the asl charge** and **wants it off** **or** her is going to **cancel** his act
- Combine sophisticated rules with sentiment statistical training and Predictive Analytics and behavior monitoring


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,

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”
- Current research – 76% accuracy in some contexts

Deep Text Social Media Analysis
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
- Companies get stuck – know the software but not how to really use it well, leads to abandoned projects
- Best way to start?
- Buy and read Deep Text, of course

Deep Text Social Media Analysis
Smart Start: Think Big, Start Small, Scale Fast

- Think Big: Strategic Vision
  - K Audit – content, people, technology, KOS
  - Establish Infrastructure – stabile foundation
  - Avoid expensive mistakes – dead end approach, technology, etc.
- Start Small: Pilot or POC
  - Immediate value and learn by doing
  - Easier to get management buy-in
- Scale Fast: multiple applications
  - Set of semantic resources and skill to implement
  - First project + 10%, subsequent projects – 50%

Deep Text Social Media Analysis
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
- Strategic Vision – Text Analytics and Information/Knowledge Environment

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, 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

Deep Text Social Media Analysis
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

Deep Text Social Media Analysis Conclusions

- Deep text analytics adds depth and intelligence, context and categorization to social media analysis
- Deep text analytics Is an infrastructure platform technology – Enterprise & Social
- Needs a strategic vision
  - But also concrete and quick application to drive acceptance
- Future is
  - Deep Text and Deep Learning integration – modules
  - Text + Data, Language + Math, Social + Enterprise, psychology + cognitive science
  - Combination of sophisticated marketing and equally sophisticated text analytics

Questions?

Learn More:

- Taxonomy Boot Camp – 10/17-18-London
- Internet Librarian – 10/22-25-Monterey
- Taxonomy Boot Camp – 11/6-7 - DC
- Text Analytics Forum – 11/8-9 –DC

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
  - Don't Think of an Elephant
    - 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 Forum - All aspects of text analytics
    - http://www.textanalyticsforum.com
  - Semtech
    - http://www.semanticweb.com
  - Dataversity Conferences
  - http://www.dataversity.net/
  - Sentiment Analysis Symposium
    - www.sentimentsymposium.com

KAPS Group
Knowledge Architecture Professional Services

Resources

- LinkedIn Groups:
  - Text Analytics
  - Text Analytics Forum
  - Taxonomy Community of Practice
  - Sentiment Analysis
  - Text and Social Analytics
  - Metadata Management
  - Semantic Technologies, Semantic Web
  - Association for Information Science & Technology