

Deep Text

New Approaches in Text Analytics and Knowledge Organization

Tom Reamy
Chief Knowledge Architect
KAPS Group

<http://www.kapsgroup.com>

Author: Deep Text

Agenda

- Introduction:
 - What is Text Analytics?
- Text Analytics and Knowledge Organization Today
- New Approaches - Fast and Slow (Thinking)
 - Deep Learning and Deep Text
- New Methods for Text Analytics
 - Unstructured Text Isn't
 - Beyond Simple Sentiment
- Applications
 - Pronoun Patterns
 - Behavior Prediction
- Context and Integration
- Questions

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:

What is Text Analytics?

- Text Mining – NLP, statistical, predictive, machine learning
- Extraction – entities – known and unknown, concepts, events
 - Catalogs with variants, rule based
- Sentiment Analysis
 - Positive and Negative expressions – catalogs & rules
- Auto-categorization
 - Training sets, Terms, Semantic Networks
 - Rules: Boolean - AND, OR, NOT
 - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
 - Disambiguation - Identification of objects, events, context
 - Build rules based, not simply Bag of Individual Words
- It is metadata! Meta NOT About

Text Analytics and Knowledge Organization Tools for Building KO

- Add powerful bottom up techniques to taxonomy building
- 250,000+ documents, search logs, emails, etc.
- Text mining = most frequent terms + multiple analysis of output
 - By date, publisher/authors, document type, etc.
 - Future = personality type, learning styles, political, etc.
- Clustering – co-occurring terms – suggest categories
- Entity extraction – existing entities, discover new ones, facets
- Simple categorization – use single or simple set of words as rules
- Text analytics provides speed and depth of coverage
- Additional methods for exploring content – supplement the experience of SME's and taxonomists

Text Analytics and Knowledge Organization Tools for Applying KO

- 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 and Knowledge Organization Tools for Applying KO - Hybrid

- 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
 - Facets – Requires a lot of Metadata - Entity Extraction feeds facets
- 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 and Knowledge Organization Tools for Applying KO - Hybrid

- Taxonomy provides a consistent and common vocabulary
 - Enterprise resource – integrated not centralized
- Text Analytics provides a consistent tagging
 - Human indexing is subject to inter and intra individual variation
- Taxonomy provides the basic structure for categorization
 - And candidates terms
- Text Analytics provides the power to apply the taxonomy
 - And metadata of all kinds
- Text Analytics and Taxonomy Together – Platform
 - Consistent in every dimension
 - Powerful and economic

New Approaches in Deep Text Analytics

Thinking Fast and Slow – Daniel Kahneman

- Brain: System 1 and System 2
- System 1 – fast and automatic – little conscious control
- Represents categories as prototypes – stereotypes
 - Norms for immediate detection of anomalies – distinguish the surprising from the normal
 - fast detection of simple differences, detect hostility in a voice, find best chess move (if a master)
 - Priming / Anchoring – susceptible to systemic errors
 - Biased to believe and confirm
 - Focuses on existing evidence (ignores missing – WYSIATI).

New Approaches in Deep Text Analytics

Thinking Fast and Slow

- System 2 – Complex, effortful judgments and calculations
 - System 2 is the only one that can follow rules, compare objects on several attributes, and make deliberate choices
 - Understand complex sentences
 - Check the validity of a complex logical argument
 - Focus attention – can make people blind to all else – Invisible Gorilla
- Similar to traditional dichotomies – Tacit – Explicit, etc
- Basic Design – System 1 is basic to most experiences, and System 2 takes over when things get difficult – conscious control
- Text Analysis and Text Mining
- Categorization by example and categorization by rules

New Approaches in Deep Text Analytics

Deep Learning – System 1

- 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.

New Approaches in Deep Text Analytics

Applied Watson – System 1 & 2

- Key concept is that multiple approaches are required – and a way to combine them – confidence score
- Multiple sources – taxonomies, ontologies, etc.
- Special modules – temporal and spatial reasoning – anomalies
- Taxonomic, Geospatial, Temporal, Source Reliability, Gender, Name Consistency, Relational, Passage Support, Theory Consistency, etc.
- Massive parallelism, many experts, pervasive confidence estimation, integration of shallow and deep knowledge
- Key step – fast filtering to get to top 100 (System 1)
- Then – intense analysis to evaluate (System 2) – multiple scoring

New Approaches in Deep Text Analytics

Categorization Rules – System 2

- Representation of Domain knowledge – taxonomy, ontology
- Catonomies – taxonomy + categorization rules
- Categorization – deep analysis
 - Most basic to human cognition
 - Basic level categories
- Beyond Categorization – making everything else smarter,
- No single correct categorization
 - Women, Fire, and Dangerous Things
- Sentiment Analysis to Expertise Analysis(Know How)
 - Know How, skills, “tacit” knowledge

New Approaches in Deep Text Analytics

Adding Structure to Unstructured Content– System 2

- Documents are not unstructured – variety of structures
 - Sections – Specific - “Abstract” to Function “Evidence”
 - Content Type – library of sections per type, department
 - Textual complexity, level of generality
- Beyond Documents – categorization by corpus, by page, sections or even sentence or phrase
- Applications require sophisticated rules, not just categorization by similarity
- Model of human learning – over-generalization, not pattern analysis of a few million examples

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]"))
)
```

Text View
 Tree View

New Approaches in Text Analytics

Document Type Rules

- Look at first 2,000 words (and last 2,000) - most important
- Take existing sections – Title – assign strongest relevancy score
- Dynamic define multiple sections – Abstract, Methods, etc.
- “[Abstract]” – multiple words – Summary, Overview, etc.
- Minimum of 2 Phrases “[arthritis]” within 7 words of “[drugs]”
- NOT words like “[Animals]”
- Assign relevancy score
- Primary issue – major mentions, not every mention
 - Combination of noun phrase extraction and categorization
 - Results – virtually 100%

New Approaches in Deep Text Analytics 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”), slang language
- 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
- Essential – need full categorization and concept extraction

New Approaches in Deep Text Analytics

Social Media: Beyond Simple Sentiment

- Analysis of Conversations- Higher level context
- Techniques: self-revelation, humor, sharing of secrets, establishment of informal agreements, private language
- Quality of communication (strength of social ties, extent of private language, amount and nature of epistemic emotions – confusion+)
- Personality types – aggressive, political – disgust = conservative

Deep Text Analytics 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 of:
 - Representatives, policies, programs, actions, etc.
- 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

Deep Text Analytics Applications

Expertise Analysis

- Expertise Characterization for individuals, communities, documents, and sets of documents
- Experts prefer lower, subordinate levels
 - Novice & General – high and basic level
- Experts language structure is different
 - Focus on procedures over content
- Applications:
 - Business & Customer intelligence – add expertise to sentiment
 - Deeper research into communities, customers
 - Expertise location- Generate automatic expertise characterization based on documents

Deep Text Analytics 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 **cancel** 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

Deep Text Analytics 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

Context and Integrated Solutions System 1 & 2 – and Text Analytics Approaches

- “Automatic Categorization” – System 1 prototypes
 - Limited value -- only works in simple environments
 - Shallow categories with large differences
 - Not open to conscious control [Black Box]
- System 2 – categories – complex, minute differences, deep categories
- Together:
 - Choose one or other for some contexts
 - Combine both – need to develop new kinds of categories and/or new ways to combine?

Context and Integrated Solutions

Deep Learning and Deep Text

- Text Analytics and Big Data enrich each other
 - Data tells you what people did, TA tells you why
- Text Analytics – pre-processing for Text Mining
 - Discover additional structure in unstructured text
 - Behavior Prediction – adding depth in individual documents
 - New variables for Predictive Analytics, Social Media Analytics
 - New dimensions – 90% of information, 50% using Twitter analysis
- Text Mining for TA– Semi-automated taxonomy development
 - Apply data methods, predictive analytics to unstructured text
 - New Models – Watson ensemble methods, reasoning apps
- Extraction – smarter extraction – sections of documents, Boolean, advanced rules – drug names, adverse events – major mention

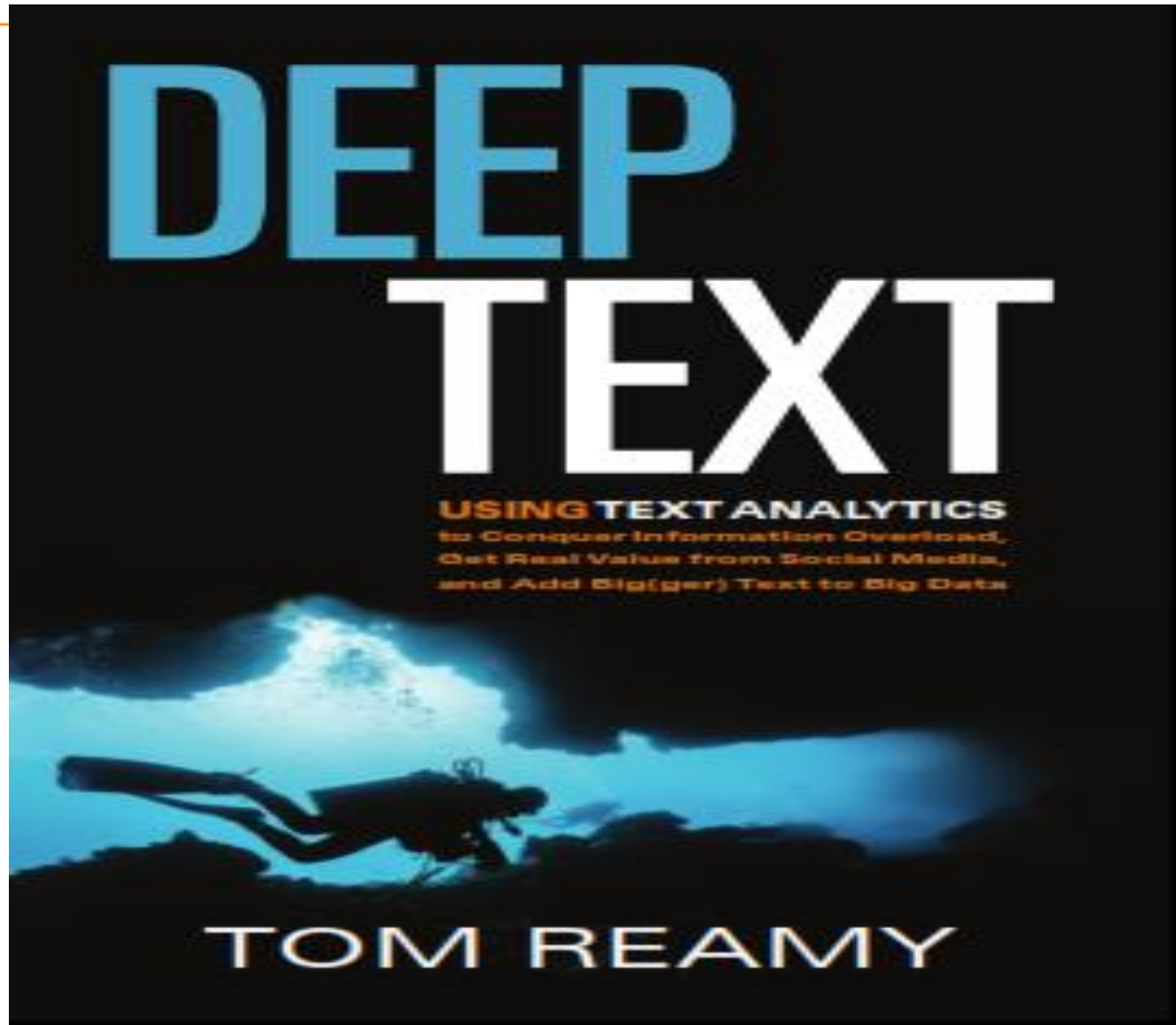
Context and Integrated Solutions

Integration of Text and Data Analytics

- Expertise Location: Case Study: Data and Text
- Data Sources:
 - HR Information: Geography, Title-Grade, years of experience, education, projects worked on, hours logged, etc.
- Text Sources:
 - Document authored (major and minor authors) – data and/or text
 - Documents associated (teams, themes) – categorized to a taxonomy
 - Experience description – extract concepts, entities
- Self-reported expertise – requires normalization, quality control
- Complex judgments:
 - Faceted application
 - Ensemble methods – combine evaluations

Deep Text: New Approaches Conclusions

- Two major techniques changing the world of text
 - Deep Text – depth and intelligence
 - Deep Learning – power and scale, machine learning
- Present applications – fix search (finally), smart Info Apps-eDiscovery, fraud detection, BI, CI, social media, smart summaries, etc., etc.
- Full Integration of the two - ongoing
 - Understand black box / learning for rules systems
- Make automated smarter, make manual scale
- Make humans work smarter – assistant reader – add depth to reading - like hybrid tagging taken to next level



Questions?

Tom Reamy
tomr@kapsgroup.com

KAPS Group

<http://www.kapsgroup.com>

Upcoming: ASIS&T – Copenhagen Oct 14

Taxonomy Boot Camp – London Oct 18-19

Taxonomy Boot Camp & Enterprise Search – DC Nov 14-17