Future Directions in Text Analytics

Tom Reamy
Chief Knowledge Architect
KAPS Group
http://www.kapsgroup.com
Program Chair – Text Analytics World
http://www.textanalyticsworld.com

KAPS Group
Knowledge Architecture Professional Services

Agenda

- Introduction
- What is Text Analytics?
- Conference Themes
  - Big Data and Text Analytics – 2 way street
  - Social Media: Beyond Negative and Positive
  - Enterprise Text Analytics
- The Present and Future of Text Analytics – Survey Results
- Strategic Vision
  - Text Analytics as Platform
  - Need to spread the word
- Conclusions

Introduction: Personal

- Deep Background: History of Ideas – dissertation – Models of Historical Knowledge
- Artificial Intelligence research at Stanford AI Lab
- Programming – designed two computer games, educational software
- Started an Education Software company, CTO
  - Height of California recession
- Information Architect – Chiron/Novartis, Schwab Intranet
  - Importance of metadata, taxonomy, search – Verity
- From technology to semantics, usability
- From library science to cognitive science
- 2002 – started consulting company

KAPS Group
Knowledge Architecture Professional Services

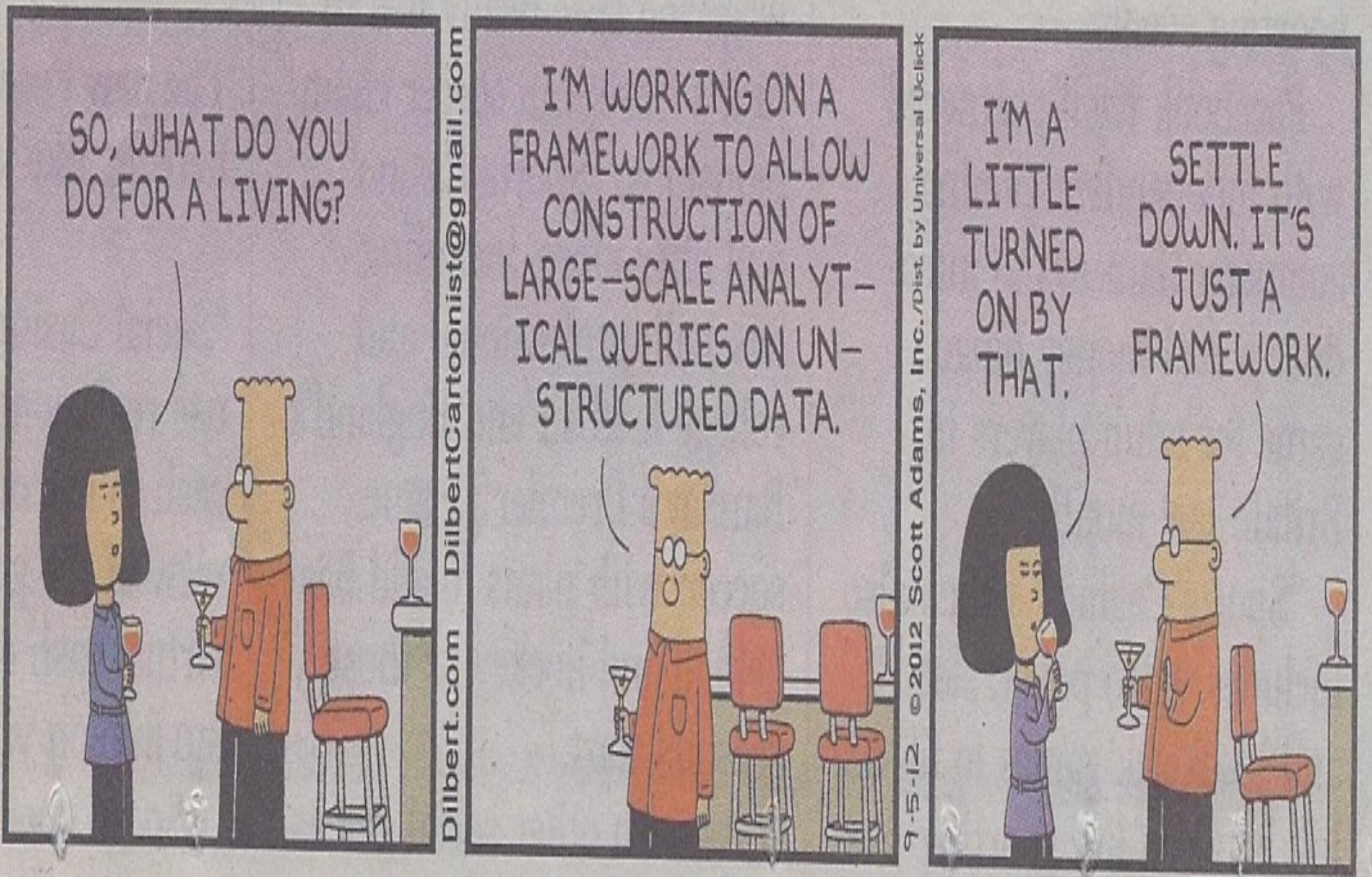
Introduction: KAPS Group

- Knowledge Architecture Professional Services – Network of Consultants
- Applied Theory – Faceted taxonomies, complexity theory, natural categories, emotion taxonomies
- Services:
  - Strategy – IM & KM - Text Analytics, Social Media, Integration
  - Taxonomy/Text Analytics development, consulting, customization
  - Text Analytics Fast Start – Audit, Evaluation, Pilot
  - Social Media: Text based applications – design & development
- Partners – Smart Logic, Expert Systems, SAS, SAP, IBM, FAST, Concept Searching, Attensity, Clarabridge, Lexalytics
- Clients:
  - Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard Business Library, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, etc.
- Presentations, Articles, White Papers – www.kapsgroup.com

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
- Focus on enterprise text analytics – shift to sentiment analysis - easier to do, obvious pay off (customers, not employees)
  - Backlash – Real business value?
- Enterprise search down, taxonomy up –need for metadata – not great results from either – 10 years of effort for what?
- Text Analytics to the rescue!
- Have we arrived? Gartner just released a report on –Text Analytics!

DILBERT By Scott Adams



Introduction: Future Directions
What is Text Analytics?

- Text Mining – NLP, statistical, predictive, machine learning
- 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
- 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

Future Directions of Text Analytics
Text and Data: Two Way Street

- Why are we talking about big data at a text conference?
  - Big Text is bigger than Big Data
  - Text Analytics and Big Data enrich each other
- Text Analytics – pre-processing for TM
  - 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

Future Directions for Text Analytics
Text and Data: Two Way Street

- New types of applications
  - New ways to make sense of data, enrich data
- Harvard – Analyzing Text as Data
  - Detecting deception, Frame Analysis
- Narrative Science – take data (baseball statistics, financial data) and turn into a story
- Political campaigns using Big Data, social media, and text analytics
- Watson for healthcare – help doctors keep up with massive information overload

Future Directions for Text Analytics
Social Media: Beyond Simple Sentiment

- Beyond Good and Evil (positive and negative)
  - Social Media is approaching next stage (growing up)
  - Where is the value? How get better results?
- Importance of Context – around positive and negative words
  - Rhetorical reversals – “I was expecting to love it”
  - Issues of sarcasm, (“Really Great Product”), slanguage
- Granularity of Application
  - Early Categorization – Politics or Sports
- Limited value of Positive and Negative
  - Degrees of intensity, complexity of emotions and documents
  - Addition of focus on behaviors – why someone calls a support center – and likely outcomes

Future Directions for Text Analytics
Social Media: Beyond Simple Sentiment

- Two basic approaches [Limited accuracy, depth]
  - Statistical Signature of Bag of Words
  - Dictionary of positive & negative words
- Essential – need full categorization and concept extraction to get full value from social media
- New Taxonomies – 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

Future Directions for 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
  - Detect relationships among speakers and changes over time
  - Strength of social ties, informal hierarchies
- Combination with other techniques
  - Expertise Analysis – plus Influencers
  - Quality of communication (strength of social ties, extent of private language, amount and nature of epistemic emotions – confusion+)
  - Experiments - Pronoun Analysis – personality types
  - Analysis of phrases, multiple contexts – conditionals, oblique

Future Directions for Text Analytics
Social Media: Beyond Simple Sentiment

- Expertise Analysis
  - Experts think & write differently – process, chunks
  - Categorization rules for documents, authors, communities
- Applications:
  - Business & Customer intelligence, Voice of the Customer
  - Deeper understanding of communities, customers – better models
  - Security, threat detection – behavior prediction, Are they experts?
  - Expertise location- Generate automatic expertise characterization
- Crowd Sourcing – technical support to Wiki's
- Political – conservative and liberal minds/texts
  - Disgust, shame, cooperation, openness

Future Directions for Text Analytics
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.
  - cci and **is upset that he has the asl charge** and **wants it off** **or** her is going to **cancel** his act
- More sophisticated analysis of text and context in text
- Combine text analytics with Predictive Analytics and traditional behavior monitoring for new applications

Future Directions: Enterprise Text Analytics

- Text Analytics is the Platform / Foundation for all kinds of unstructured information applications
- CM/ Search/ Search-based Applications Platform
  - Business Intelligence, Customer and Competitor Intelligence
  - eDiscovery, litigation support, compliance
  - Fraud detection
  - Recommendation engines
  - Reputation and opinion monitoring applications
- Internal and external publishing auto-tagged
- [Insert your favorite idea here]

Enterprise Text Analytics Information Platform

- 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
  - BI, CI limited sources //labor intensive// SBA need language
- 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, and much more

Enterprise Text Analytics
Information Platform: Tagging Documents

- How do you bridge the gap – taxonomy to documents?
- 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 business 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)!

Enterprise Text Analytics
Information Platform: Content Management

- Hybrid Model
  - 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
- Hybrid – Automatic is really a spectrum – depends on context
  - All require human effort – issue of where and how effective
- External Information - human effort is prior to tagging
  - Build on expertise – librarians on categorization, SME's on subject terms

Future Directions: Survey Results

- Who – mix, TA vendor/consultant, finance, education
- Size – Greater than $1B – 28.6%
  - $50M - $1B- 20.6%
  - Less than $50M – 27%
- Function
  - Executive – 44%, Manager – 25%, Staff – 21%
- TA Knowledge – Expert – 38%, General – 40%, Novice – 22%
- Use of TA
  - All / many – 17.5%, Big Data & Social media – 10% each
  - Just Getting Started – 28.6%, Not Yet started -11%
  - Enterprise Text Analytics – 17.5% - Surprise!!!

Future Directions: Survey Results

- Who owns Text Analytics?
  - IT- 8%
  - Marketing – 13% (highest novice – 38%)
  - R&D- 40%
  - No One – 16%

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%

Future Directions: Survey Results

- What factors are holding back adoption of TA?
  - Lack of clarity about value of TA – 23.4%
  - Lack of knowledge about TA – 17.0%
  - Lack of senior management buy-in - 8.5%
  - Don't believe TA has enough business value -6.4%
- Other factors
  - Financial Constraints – 14.9%
  - Other priorities more important – 12.8%
- Lack of articulated strategic vision – by vendors, consultants, advocates, etc.

Strategic Vision for Text Analytics 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

Strategic Vision for Text Analytics
Adding Intelligence – High Level

- Understand your customers
  - What they are talking about and how they feel about it
- Empower your employees
  - Not only more time, but they work smarter
- Understand your competitors
  - What they are working on, talking about
  - Combine unstructured content and rich data sources – more intelligent analysis
- Integration of all of the above – Platform
  - Integration at the semantic level

Strategic Vision for Text Analytics
Building the Platform - Strategic Vision

- Info Problems – what, how severe
- Formal Process - KA audit – content, users, technology, business and information behaviors, applications - Or informal for smaller organization,
- Contextual interviews, content analysis, surveys, focus groups, ethnographic studies, Text Mining
- Category modeling – Cognitive Science – how people think
  - Monkey, Panda, Banana
- Natural level categories mapped to communities, activities
    - Novice prefer higher levels
    - Balance of informative and distinctiveness
- Text Analytics Strategy/Model – What is text analytics?

New Directions in Text Analytics Conclusions

- Text Analytics as enriching foundation/ platform
  - Big data, social media, Search, SBA, and text
  - Smart Enterprise – Semantic Enterprise
- New models are opening up
  - Beyond sentiment – emotion & behavior, cognitive science
  - Enterprise Hybrid Model, Data and Text models
- Needs to develop better methods
  - Easier, smarter, more integrated
  - Integration of NLP and categorization, better visualization
- Needs a more articulated strategic vision
  - Same process to develop vision and platform
- Text Analytics World will cover the whole spectrum

New Directions in Text Analytics
Text Analytics World

- Thursday Keynote – Sue Feldman IDC
  - Search and Text Analytics
- Big Data, Predictive Analytics, and Text
  - New approaches and applications
- Enterprise Text Analytics (2 days)
  - Applications, Tools, Techniques, How-To
- Social Media, Voice of the Customer, and Text
  - Beyond simple sentiment, Twitter bits
- Great Sponsors- Expert Systems, Smart Logic
  - Visit and learn
- Panel – Future of Text Analytics – Discussion
- This is a great time to be getting into text analytics!

Questions?

Tom Reamy
tomr@kapsgroup.com
KAPS Group
http://www.kapsgroup.com

Upcoming: Text Analytics World – San Francisco, April 17-18

SAS A2012 – Las Vegas, Oct 8-9

Taxonomy Boot Camp – Washington DC, Oct 16-17

Gilbane – Boston, November 27-29

KAPS Group
Knowledge Architecture Professional Services