Text Analytics Evaluation Case Study - Amdocs

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

Text Analytics World
October 20 New York

KAPS Group
Knowledge Architecture Professional Services

KAPS Group
Knowledge Architecture Professional Services

Agenda

- Introduction – Text Analytics Basics
- Evaluation Process & Methodology
  - Two Stages – Initial Filters & POC
- Initial Evaluation Results
- Proof of Concept
  - Methodology
  - Results
- Final Recommendation
- Conclusions

Introduction to Text Analytics
Text Analytics Features

- Noun Phrase Extraction
  - Catalogs with variants, rule based dynamic
  - Multiple types, custom classes – entities, concepts, events
  - Feeds facets
- Summarization
  - Customizable rules, map to different content
- Fact Extraction
  - Relationships of entities – people-organizations-activities
  - Ontologies – triples, RDF, etc.
- Sentiment Analysis
  - Rules – Objects and phrases

Introduction to Text Analytics
Text Analytics Features

- 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 – NEAR (#), PARAGRAPH, SENTENCE
- This is the most difficult to develop
- Build on a Taxonomy
- Combine with Extraction
  - If any of list of entities and other words

Evaluation Process & Methodology

- Start with Self Knowledge
  - Think Big, Start Small, Scale Fast
- Eliminate the unfit
  - Filter One- Ask Experts - reputation, research – Gartner, etc.
    - Market strength of vendor, platforms, etc.
    - Feature scorecard – minimum, must have, filter to top 3
  - Filter Two – Technology Filter – match to your overall scope and capabilities – Filter not a focus
  - Filter Three – In-Depth Demo – 3-6 vendors
- Deep POC (2) – advanced, integration, semantics
- Focus on working relationship with vendor.

Evaluation Process & Methodology
Amdocs Requirements / Initial Filters

- Platform – range of capabilities
  - Categorization, Sentiment analysis, etc.
- Technical
  - API's, Java based, Linux run time
  - Scalability – millions of documents a day
  - Import-Export – XML, RDF
- Total Cost of Ownership
- Vendor Relationship - OEM
- Usability, Multiple Language Support

Vendors of Taxonomy/ Text Analytics Software

- Attensity
- Business Objects – Inxight
- Clarabridge
- ClearForest
- Concept Searching
- Data Harmony / Access Innovations
- **Expert Systems**
- GATE (Open Source)
- **IBM Infosphere**
- Lexalytics
- Multi-Tes
- Nstein
- **SAS - Teragram**
- SchemaLogic
- **Smart Logic**
- Synaptica

Initial Evaluation
4 Demos

- Smartlogic
  - Taxonomy Management, good interface
  - 20 types of entities, API's, XML-Http
  - Full Platform – no Sentiment Analysis
- Expert Systems
  - Different Approach – Semantic Network – 400,000 words / 3,500 rules, 65 types of relationships
  - Strong out of the box – 80%, no training sets
  - Language concerns – no Spanish, high cost to develop new ones
  - Customization – add terms and relationships, develop rules – uncertain how much effort, use their professional linguists

Initial Evaluation
4 Demos

- SAS-Teragram
  - Full Platform – categorization, entity, sentiment – integrated
  - API's, XML, Java – ease of integration
  - Strong history of company, range of experience
- IBM – Classification, Concept Analytics – Two products
  - Classification Module – statistical emphasis
    - Once trained, it could "learn" new words
    - Rapid development / depends on training sets
  - Content Analytics, Languageware Workbench
    - Full Platform

Initial Evaluation – Findings

- SAS & IBM – Full Platform, OEM Experience, multilingual
  - Proven ability to scale, customizable components, mature tool sets
- SAS was the strongest offering
  - Capabilities, experience, integrated tool sets
- IBM good second choice
  - Capabilities, experience - multiple products – strength and weakness
- Single Vendor POC - Demonstrate it can be done
  - Ability to dive more deeply into capabilities, issues
  - Stronger foundation for future development, Learn the software better
  - Danger of missing better choice
- Two Vendor POC
  - Balance of depth and full testing

Phase II - Proof Of Concept - POC

- Measurable Quality of results is the essential factor
- 4 weeks POC – bake off / or short pilot
- Real life scenarios, categorization with your content
- 2 rounds of development, 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
- Need to balance uniformity of results with vendor unique capabilities – have to determine at POC time
- Taxonomy Developers – expert consultants plus internal taxonomists

POC Design: Evaluation Criteria & Issues

- Basic Test Design – categorize test set
  - Score – by file name, human testers
- Categorization & Sentiment – Accuracy 80-90%
  - Effort Level per accuracy level
- Quantify development time – main elements
- Comparison of two vendors – how score?
  - Combination of scores and report
- Quality of content & initial human categorization
  - Normalize among different test evaluators
- Quality of taxonomists – experience with text analytics software and/or experience with content and information needs and behaviors
- Quality of taxonomy – structure, overlapping categories

Text Analytics POC Outcomes
Categorization of CSR Notes

- Content –2,000 CSR notes categorized by humans
  - Variation among human categorization
- Recall (finding all the correct documents)
- Precision (not categorizing documents from other categories)
  - Precision is harder than recall
  - Two scores – raw and corrected – only raw for IBM precision
  - First score was very low, with an extra round got it up
- Uncategorized documents – 50,000 – look at top 10 in each category

Text Analytics POC Outcomes
Categorization Results

	SAS	IBM	
Recall-Motivation	92.6	90.7	
Recall-Actions	93.8	88.3	
Precision – Mot.	84.3		
Precision-Act	100		
Uncategorized	87.5		
Raw Precision	73	46	

Text Analytics POC Outcomes
Vendor Comparisons

- SAS has a much more complete set of operators – NOT, DIST, START
  - IBM team was able to develop work arounds for some – more development effort
  - Operators impact most other features – Sentiment analysis, Entity and Fact Extraction, Summarization, etc.
- SAS has relevancy – can be used for precision, applications
- Sentiment Analysis – SAS has workbench, IBM would require more development
  - SAS also has statistical modeling capabilities
- Development Environment & Methodology
  - IBM as toolkit provides more flexibility but it also increases development effort, enforces good method

Text Analytics POC Outcomes
Vendor Comparisons - Conclusions

- Both can do the job
  - Product vs. Tool Kit (SAS has toolkit capabilities also)
- IBM will require more development effort
  - Boolean Operators – NOT, DIST, START, etc.
    - In rules, entity and fact extraction
  - Sentiment Analysis – rules, statistical
  - Summarization
  - Rule building more programming than taxonomy
- IBM harder to learn – POC had 2X effort for IBM

Text Analytics Evaluation Conclusions

- Start with Self Knowledge – text analytics not an end in itself
- Initial Evaluation – filters, not scorecards
  - Weights change output – need self knowledge for good weights
- Proof of Concept – essential
  - OOB doesn't tell you how it will work in real world
  - Content and Scenarios is your real world
- Importance of operators, relevance for a platform
- Sentiment needs full platform
- Everyone has room for improvement

Questions?

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

KAPS Group
Knowledge Architecture Professional Services