Auto-categorization for Superior Accuracy

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

KAPS Group
Knowledge Architecture Professional Services

Agenda

- Introduction – Categorization Basics
- Creating the Right Foundation
- Categorization Rule Development
- Enhancing Accuracy
- Issues, Tips, and Techniques
- Conclusion

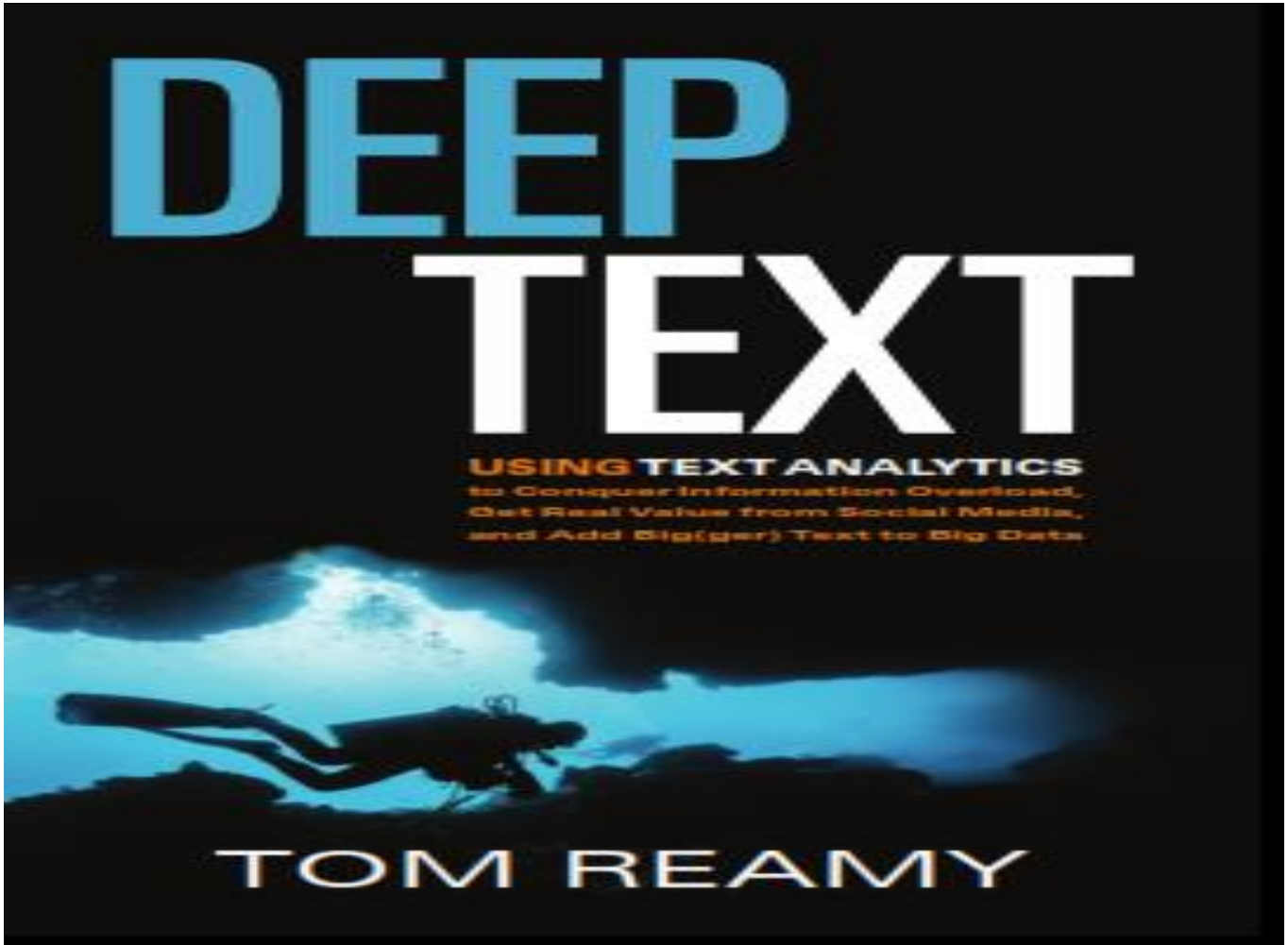
Introduction: KAPS Group

- Network of Consultants and Partners - 2002
- Text analytics consulting: Strategy, Development-taxonomy, text analytics foundation & applications
- Mini-Projects – get started or take to next level
  - Strategy, Mini-POC - Categorization
- Partners – Expert AI, Megaputer, SAS, Smart Logic, Lexalytics, BA Insight, BiText, Synaptica
- Clients: Genentech, Novartis, Northwestern Mutual Life, Financial Times, Hyatt, Home Depot, Harvard, British Parliament, Battelle, Amdocs, FDA, GAO, World Bank, IMF, IFC, Dept. of Transportation, RWJF, Intel, Kellogg Foundation, Foundry (IDG), etc.
- Presentations, Articles, White Papers – www.kapsgroup.com
- Program Chair – Text Analytics Forum

KAPS Group

A treasure trove of technical detail, likely to become a definitive source on text analytics – *Kirkus Reviews*

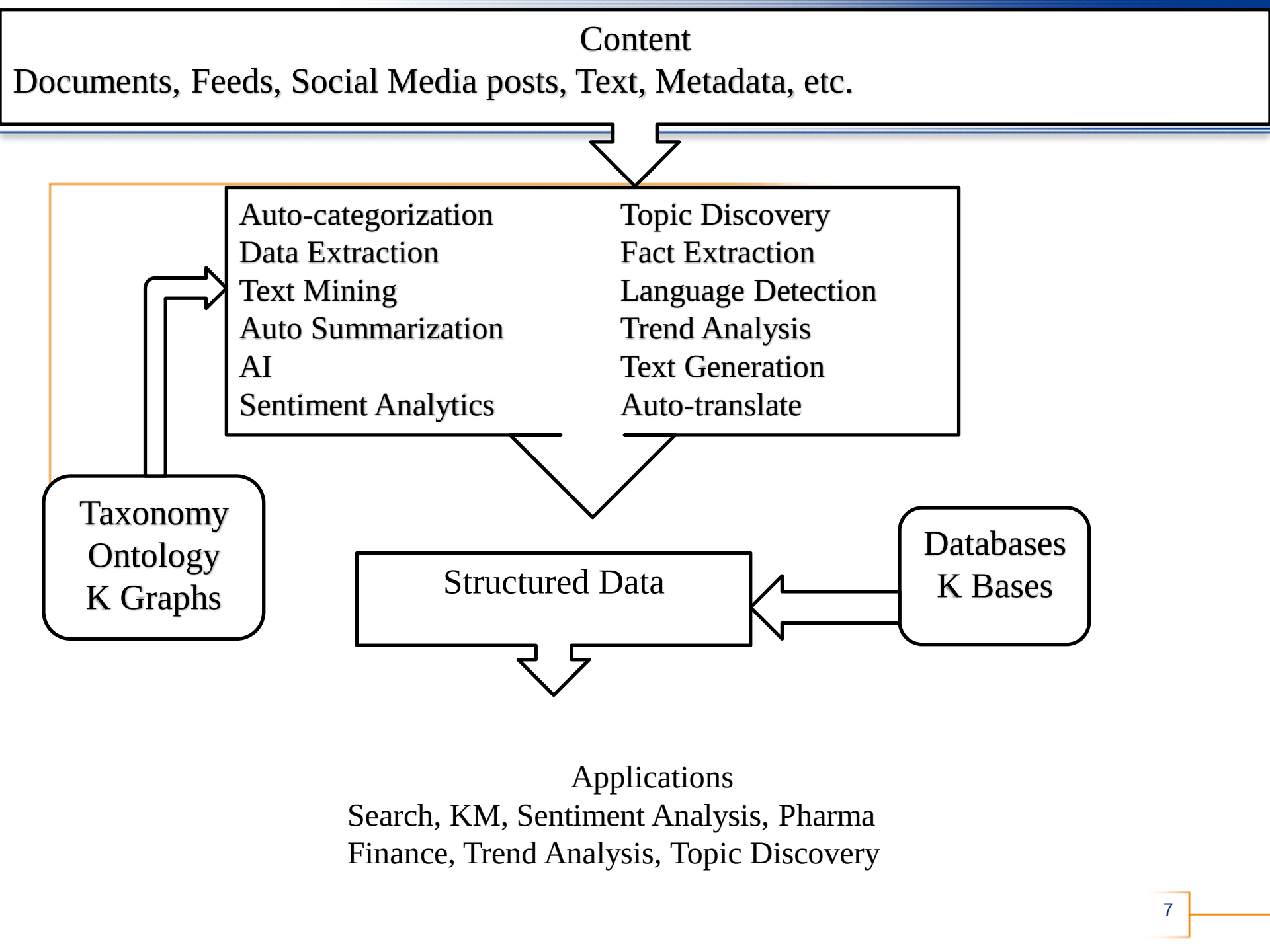
Book sign-M-6-6:30, T-5:30-6, W-3:14-4



Categorization Basics

Text Analytics Development: Categorization Basics

- Representation of domain knowledge – taxonomy, ontology
- Categorization – Most basic to human cognition
  - Most difficult to do with software
  - Explicit subject, tacit knowledge, sentiment, expertise
- Beyond Categorization – making everything else smarter
  - Disambiguation – within categorization and entity extraction
- No single correct categorization
  - Women, Fire, and Dangerous Things
- Building blocks
  - Taxonomy, Content, Supplementary Resources


Content
Documents, Feeds, Social Media posts, Text, Metadata, etc.
Auto-categorization
Data Extraction
Text Mining
Auto Summarization
AI
Sentiment Analytics
Topic Discovery
Fact Extraction
Language Detection
Trend Analysis
Text Generation
Auto-translate
Taxonomy
Ontology
K Graphs
Structured Data
Databases
K Bases
Applications
Search, KM, Sentiment Analysis, Pharma
Finance, Trend Analysis, Topic Discovery

Auto-Cat for Superior Accuracy
Categorization Techniques – Two Basic Approaches

- Machine Learning – Bayesian, Vector space, CNN, RNN
  - Create a statistical/neural net signature and compare new content
  - Results are poor, difficult to improve, needs large numbers of representative documents
- Categorization language - AND, OR, NOT
  - Advanced – DIST(#), ORDDIST#, PARAGRAPH, SENTENCE
  - Good results, flexible and power – DIST, etc.
  - Need to learn a categorization language

Text Analytics Workshop
Machine/Deep Learning and Rules

- Current trend – how to combine
- Claim – ML is faster to develop – only if unsupervised – typically bad results
- Selecting documents takes time and effort – and difficult to do well
- Rules (and Taxonomy) can provide structure and better training sets
- ML can provide terms for rules
- Categorization rule consists of logic and sets of evidence terms

KAPS Group
Knowledge Architecture Professional Services

Creating the Right Foundation

Auto-Cat for Superior Accuracy

The start and foundation: Knowledge Audit

- Knowledge Map - Understand what you have, what you are, what you want
- Contextual interviews, content analysis, surveys, focus groups, ethnographic studies, Text Mining
- Category modeling – Monkey, Panda, Banana
- 4 Dimensions – Content, People, Technology, Activities
- Strategic Vision and Change Management
  - Format – reports, enterprise ontology
  - Political/ People and technology requirements

Auto-Cat for Superior Accuracy
Creating the Right Foundation

- Right Software
  - Taxonomy Management, Content Management, editing rules, testing, refinement
- Good Taxonomy – orthogonal, Embodied Taxonomy – DOT
- Content Analysis -TM, content types
- Good content – all categorization starts with example documents
- User research
- Process – what is possible in that organization

KAPS Group
Knowledge Architecture Professional Services

Categorization Rule Development

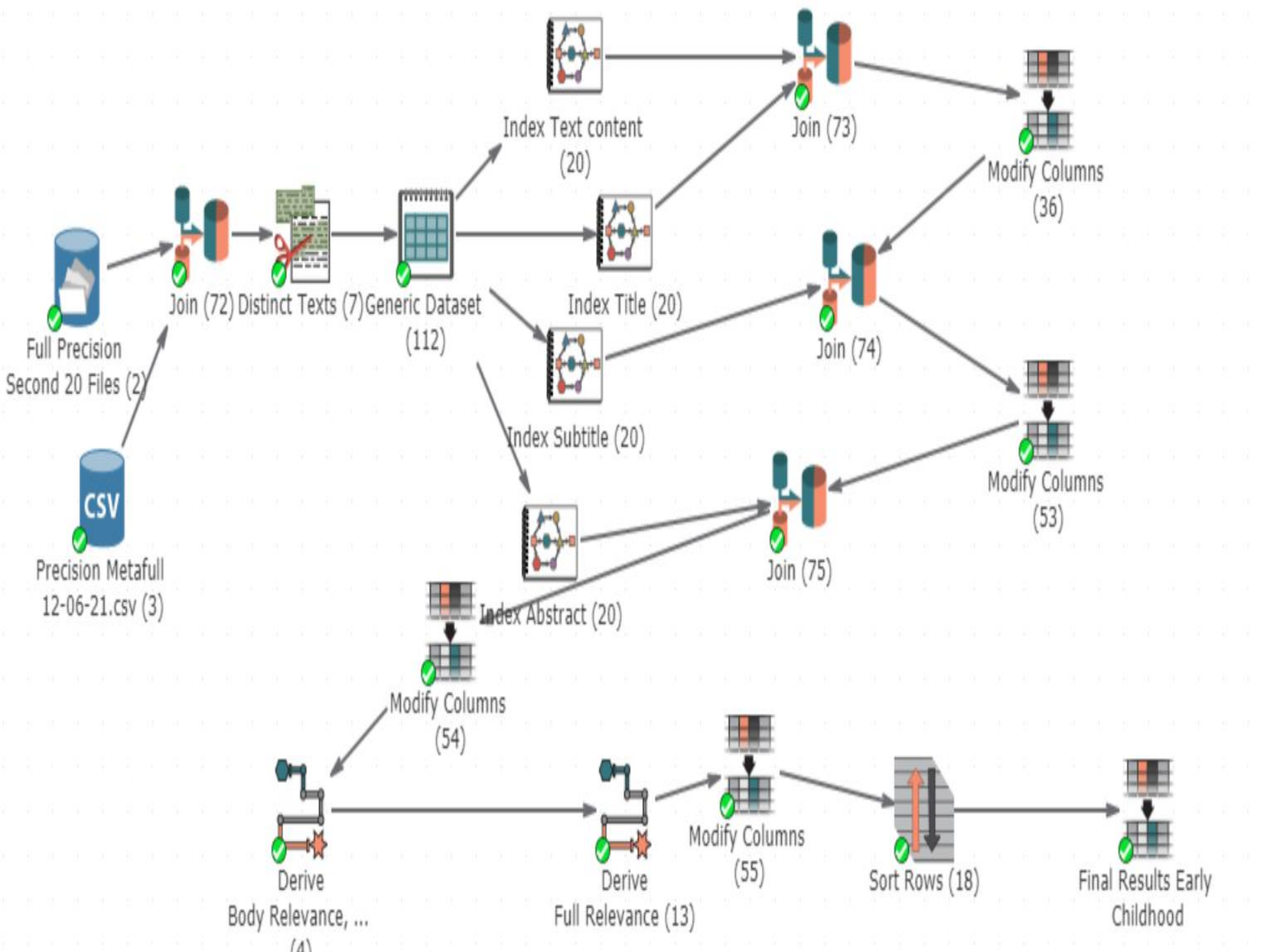
Auto-Cat for Superior Accuracy
Two Projects

- Foundation – Faceted Taxonomies
  - Topic - 12 broad Categories
  - Large documents
  - Complex Rules & Terms – 400+
  - Re-tag 4+ tags – Primary Tag only
- Publisher – 1,000 node taxonomy, 3-5 levels deep
  - Variety of types
    - Specific – entity type – Cameras
    - Broad concepts – Analytics
  - Mostly short documents – 3-5 pages, some longer – 10-20 pages
  - Re-tag all – current 1 – 60 tags – Primary & Seconary Tag
  - Develop a hybrid tagging application

KAPS Group

Auto-Cat for Superior Accuracy
Two Projects – Content Structure

- Foundation – Metadata and Internal Text
  - Title
  - Subtitle
  - Abstract
  - Text Indicators
- Publisher - Metadata and Internal Text
  - Title – highest
  - Description - highest
  - Sub-headline – Higher
  - H2 rules – High
  - First Three Sentences - High
  - Body – Normal


Index Text content
(20)
Join (73)
Modify Columns
(36)
Full Precision
Second 20 Files (2)
Join (72)
Distinct Texts (7)
Generic Dataset
(112)
Index Title (20)
Join (74)
Index Subtitle (20)
Modify Columns
(53)
CSV
Precision Metafull
12-06-21.csv (3)
Join (75)
Index Abstract (20)
Modify Columns
(54)
Derive
Body Relevance, ...
Derive
Full Relevance (13)
Modify Columns
(55)
Sort Rows (18)
Final Results Early
Childhood

Findings from the Evaluation of the RWJF **Community Coalition Leadership Program**

Asset ID	TaxLevelName1	Full Relevance	Text content	Title	Subtitle
420008	Health Leadership	113.29	FINAL REPORT Transforming Coaliti	Transforming Coalition Leadership	Findings from the Evaluation of the RWJF Com
417611	Health Leadership	58.80	Gardiner School Leadership Curr	Gardiner School Leadership Curriculum	
232132	Health Leadership	57.89	The UHI Lessons Learned project is	Enlisting Leaders in Community Change	The UHI Fellows Program
459627	Health Leadership	50.40	2020 Call for Applications Applicatio	Health Policy Research Scholars	
175721	Health Leadership	44.01	BUILDING CAPACITY FOR EVALUAT	Building Capacity for Evaluation	Self Evaluation Training
297657	Health Leadership	43.66	Created by MONITOR INSTITUT	Future Impact of Neuroscience and Behavi	
432875	Health Leadership	42.46	NLAPH training supports cross-secto	Cross-Sector Leadership	
460666	Health Leadership	39.70	Resourcing Networks for Equitable S	Resourcing Networks for Equitable Systen	
415879	Health Leadership	38.04	Validating a conceptual model for an	Validating a Conceptual Model for an Inter	
323955	Health Leadership	37.49	Changing the Face of Public Service I	Changing the Face of Public Service Leade	
432879	Health Leadership	34.98	Impact of the National Leadership Ac	Impact of the National Leadersip Academy	
454048	Health Leadership	34.84	The Harold Amos Medical Faculty I	The Harold Amos Medical Faculty Develop	
423204	Health Leadership	34.08	1 RWJF Strategic Communicatio	RWJF Strategic Communications Traning I	Year End Coaching Summary
463579	Health Leadership	31.64	2021 Call for Applications Applicatio	Health Policy Research Scholars	
459628	Health Leadership	29.04	2020 Call for Applications Applicatio	Interdisciplinary Research Leaders	
464252	Health Leadership	28.50	ORIGINAL RESEARCH Training	Training "Pivots" from the Pandemic	Lessons Learned Transitioning from In-Person
464464	Health Leadership	28.13	© 2021 ROBERT WOOD JOHNSON	Interdisciplinary Research Leaders	
417923	Health Leadership	27.66	2015 Call for Proposals Proposal Dea	Harold Amos Medical Faculty Developmen	
464522	Health Leadership	27.59	CREATING CULTURE THROUGH HE	Network Strategies & Cross-Collaboration	
423322	Health Leadership	26.57	RWJF and YUSA Senior Leadership M	RWJF and YUSA Senior Leadership Meetin	
459893	Health Leadership	25.98	2020 Call for Applications Applicatio	Harold Amos Medical Faculty Developmen	
463951	Health Leadership	25.42	2021 Call for Applications Applicatio	Harold Amos Medical Faculty Developmen	
463951	Health Leadership	25.42	2021 Call for Applications Applicatio	Harold Amos Medical Faculty Developmen	
463951	Health Leadership	25.42	2021 Call for Applications Applicatio	Harold Amos Medical Faculty Developmen	
463951	Health Leadership	25.42	2021 Call for Applications Applicatio	Harold Amos Medical Faculty Developmen	

Record 1 of 41

Data Statistics Distinct

Project

idg-autocategortization-2 C:\Users\tomr\StudioProjects\idg-autocategortization
.idea
ann
documents
gen
package
reports
rules
Analytics Rules
Analytics Terms
Artificial Intelligence Rules
Artificial Intelligence Rules.cr
GPUs Rules.cr
Machine Learning Rules.cr
Robotics Rules.cr
Artificial Intelligence Terms
Business Operations Rules
Business Operations Terms
Careers Rules
Careers Terms
Cloud Computing Rules
Cloud Computing Terms
Computers and Peripherals Rules
Computers and Peripherals Terms
Consumer Electronics Rules
Consumer Electronics Terms

main.jr | Artificial Intelligence Rules.cr | main.cr

```
1  // Title Rules
2  SCOPE SENTENCE IN SEGMENT (DOCTITLE)
3  {
4      DOMAIN("Artificial Intelligence":HIGHEST)
5      {
6          KEYWORD( EXPAND "Artificial Intelligence Terms\Artificial Intelligence Terms.c
7         AND NOT
8          KEYWORD( EXPAND "Artificial Intelligence Terms\Artificial Intelligence Terms-N
9      }
10
11     DOMAIN ("Artificial Intelligence":HIGHEST)
12     {
13         KEYWORD( EXPAND "Artificial Intelligence Terms\Artificial Intelligence Terms
14         <-7:7>
15         KEYWORD( EXPAND "Artificial Intelligence Terms\Artificial Intelligence Terms
16         AND NOT
17     KEYWORD( EXPAND "Artificial Intelligence Terms\Artificial Intelligence Terms-Neg.c
18     }
19 }
20
21 // Summary Rules - <desc></desc>
22 SCOPE SENTENCE IN SEGMENT (DOCSUMMARY)
23 {
```

Documents Taxonomy Properties Profiling

Filter by result: Filter by file name:

Validati...	File	Size	Durati...	Succe...	Categories	Extractions	Categorization					
							TP	FP	FN	Precision	Recall	F-Measure
	B2B-200-random/APIs_221967_GraphQL_-T.txt	5,396	00:00.4	✓	15	0	1	2	3	33.00%	25.00%	29.00%
	B2B-200-random/Amazon-Web-Services_189679_Top...	9,252	00:00.8	✓	39	0	1	1	6	50.00%	14.00%	22.00%
	B2B-200-random/Analytics_191176_How-Cargil.txt	8,638	00:00.5	✓	25	0	0	3	6	0.00%	0.00%	0.00%
	B2B-200-random/Analytics_191785_Real-time-.txt	8,271	00:00.6	✓	13	0	0	1	2	0.00%	0.00%	0.00%
	B2B-200-random/Analytics_193889_Episode-6_.txt	1,795	00:00.1	✓	6	0	0	0	6	0.00%	0.00%	0.00%
	B2B-200-random/Android-Security_189646_Securing-...	9,808	00:00.8	✓	44	0	0	4	3	0.00%	0.00%	0.00%
	B2B-200-random/Application-Management_250395_...	3,393	00:00.2	✓	4	0	0	0	3	0.00%	0.00%	0.00%
	B2B-200-random/Application-Management_350220_v...	4,252	00:00.3	✓	14	0	0	0	2	0.00%	0.00%	0.00%
	B2B-200-random/Application-Performance-Managem...	5,998	00:00.4	✓	22	0	0	1	1	0.00%	0.00%	0.00%
	B2B-200-random/Big-Data_188847_Real-time-.txt	8,563	00:00.6	✓	29	0	0	1	2	0.00%	0.00%	0.00%
	B2B-200-random/Budgeting_218420_Design-thi.txt	8,132	00:00.6	✓	21	0	0	1	25	0.00%	0.00%	0.00%
	B2B-200-random/Business-Process-Management_401...	11,075	00:01.0	✓	40	0	0	3	1	0.00%	0.00%	0.00%

Files: **200**

Close

JR main.jr | CL Analytics Terms.cl | CR Analytics Rules.cr | TXT Consumer-Electronics_579117_This-glori.txt | CR main.cr

```
<title>This glorious 65-inch 4K Roku TV is a jaw-dropping $400 today</title>
<subheadline>Watch the Super Bowl in style and save $200 for snacks.</subheadline>
<desc>Best Buy is selling a Westinghouse Roku smart TV for just $400.</desc>
<story_type>Deal</story_type>
<article_type>default</article_type>
The Super Bowl is coming, and if you plan to watch the big game you're going to need a bigger screen. Best Buy has you covered today with a Westinghouse 65-inch 4K Roku smart TV for a measly $400Remove non-product link'$200 off the MSRP and one of the best prices you'll ever see.
The TV itself features 4K resolution, as well as HDR 10. We wouldn't call that true HDR as this TV doesn't get quite bright enough to qualify, but you'll certainly notice the improved picture over similarly sized sets without it.'ith Roku on board, you have access to all your favorite streaming services including Apple TV, HBO, Hulu, Netflix, Starz, and more. This TV also works with Amazon Alexa and Google Assistant for voice control, as well as integrating with other smart home devices.
For ports, it has three HDMI one of which has HDCP 2.2, as well as one USB 2.0 for thumb drives and other storage devices. It's also packing one digital optical audio out and analog audio out. For connectivity, you get Wi-Fi and Ethernet, but no Bluetooth.
Add it all up and you've got one heck of a bargain for just $400. So go grab one and get read to watch the big game in style.
[<strong>Today's deal:</strong> Westinghouse 65-inch 4K Roku smart TV for $400 at Best BuyRemove non-product link]
```

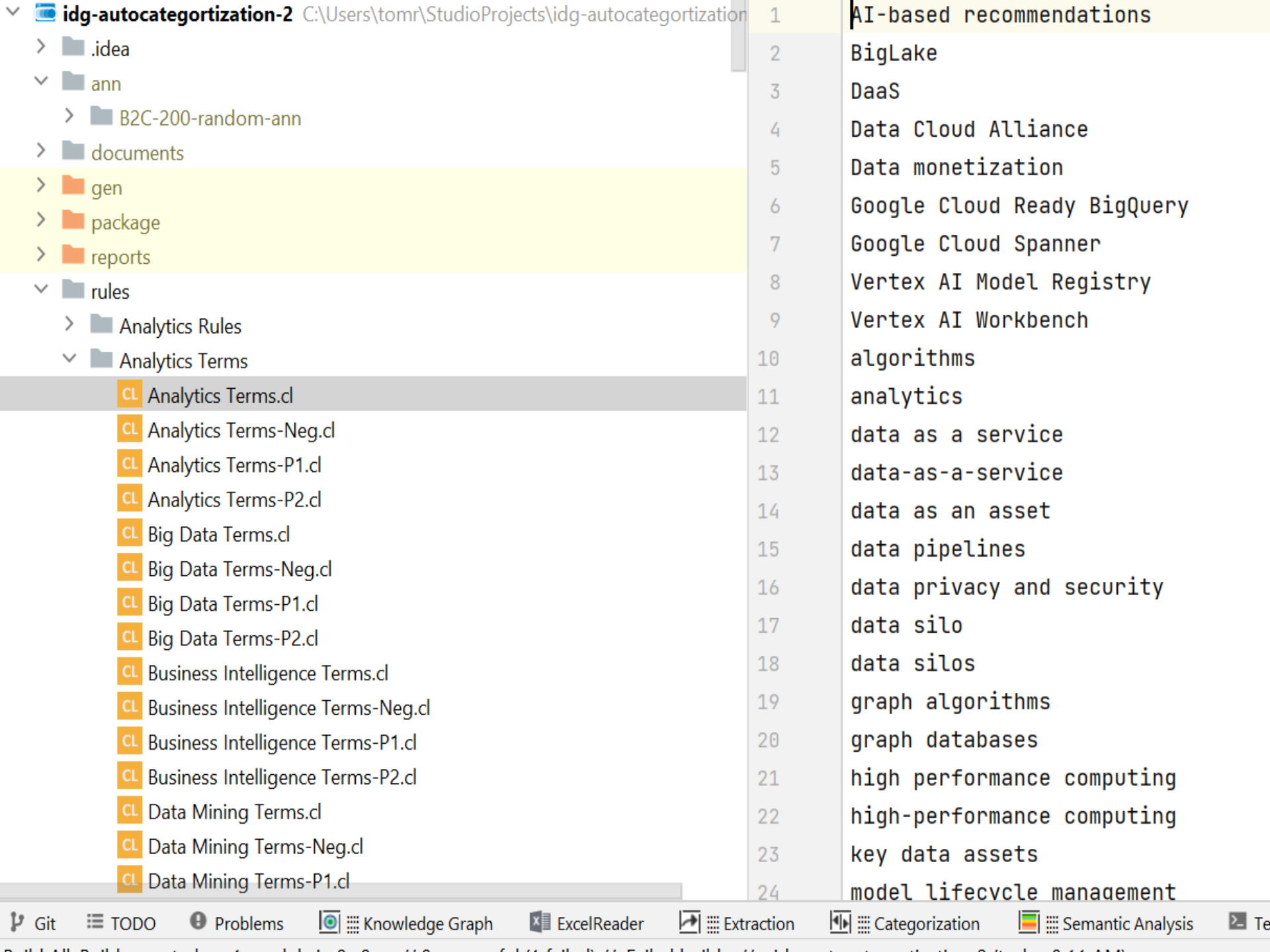
Extraction | Categorization | Semantic Analysis | Terminal | Statistics | Segments

Analysis Details

Description **B2C ran**

Auto-Cat for Superior Accuracy
What Makes a Good Term?

- Keywords – NO!!!
  - Mostly related terms, not terms that indicate what a document is about
  - Evidence terms
- 3 types of evidence terms
  - Single phrases that appear in target document and not others
  - 2 words/phrases that are near each other (7-10 words)
  - Negative terms – if found, discount - deal with overlapping taxonomy
- Good rules have two+ advantages over human tagging:
  - Consistency
  - Transparency


idg-autocategortization-2 C:\Users\tomr\StudioProjects\idg-autocategortization
.idea
ann
B2C-200-random-ann
documents
gen
package
reports
rules
Analytics Rules
Analytics Terms
Analytics Terms.cl
Analytics Terms-Neg.cl
Analytics Terms-P1.cl
Analytics Terms-P2.cl
Big Data Terms.cl
Big Data Terms-Neg.cl
Big Data Terms-P1.cl
Big Data Terms-P2.cl
Business Intelligence Terms.cl
Business Intelligence Terms-Neg.cl
Business Intelligence Terms-P1.cl
Business Intelligence Terms-P2.cl
Data Mining Terms.cl
Data Mining Terms-Neg.cl
Data Mining Terms-P1.cl
1 AI-based recommendations
2 BigLake
3 DaaS
4 Data Cloud Alliance
5 Data monetization
6 Google Cloud Ready BigQuery
7 Google Cloud Spanner
8 Vertex AI Model Registry
9 Vertex AI Workbench
10 algorithms
11 analytics
12 data as a service
13 data-as-a-service
14 data as an asset
15 data pipelines
16 data privacy and security
17 data silo
18 data silos
19 graph algorithms
20 graph databases
21 high performance computing
22 high-performance computing
23 key data assets
24 model lifecycle management
Git
TODO
Problems
Knowledge Graph
ExcelReader
Extraction
Categorization
Semantic Analysis

Enhancing Accuracy

Auto-Cat for Superior Accuracy
Measuring Accuracy

- What is Accuracy?
  - Statistical measure – never 100%
  - F-measure – combination of recall and precision
- Ask 3 humans and get 2 answers
- Two Approaches:
  - Ask human – agree with auto-tag
    - Maximum – 75%
    - Overcome with 3 humans, 2 = success
  - Ask human to tag and compare with auto-tag
    - Maximum – 90%

Auto-Cat for Superior Accuracy
Content Structure Rules

- Documents are not unstructured – variety of structures
  - Sections – Specific - “Abstract” to Function “Evidence”
- Content Structure Model – taxonomy of content types
  - Defined by Sections – text indicators or metadata
- Summary – human judgement on what the document is about
- Issues
  - Multiple text indicators
  - Variability of writing
  - Some documents had no sections – use clusters of co-occurring terms

KAPS Group
Knowledge Architecture Professional Services

Providing actuarial analyses and modeling of health reform ideas to stabilize individual insurance markets and continuing RWJF's actuarial challenge

Fund Description: To continue dissemination and analytical activities associated with the results of the 2017 RWJF Actuarial Challenge, in which teams of actuaries proposed solutions to stabilize the individual insurance market.

SUMMARY

This project will continue the work of the RWJF Actuarial Challenge. The Actuarial Challenge took place in early 2017. The final results included policy suggestions for stabilizing the individual market, including elements such as reinsurance, auto enrollment, and other market reforms. Milliman organized the challenge and simulated the winning proposals, providing estimates of how they would impact enrollment and public and private spending. As the prospect for bipartisan health reform increases, there is an increased demand for disseminating these results and for potentially engaging in some additional actuarial modelling. The challenge process and results are reviewed as technically credible and politically nonpartisan. As the effort to repeal and replace has abated, there may be an opportunity to bring some bipartisan suggestions for reform forward. Several organizations and RWJF are planning meetings and presentations for the next several months, with the goal of sharing the challenge results with policymakers and other stakeholders. At some point, this will most likely result in engaging Milliman in simulating some refined version of some elements of the winning proposals. It may ultimately be recommended to stage a second round of the challenge. The deliverables will include meetings, presentations, discussions with stakeholders, and, potentially, additional simulations. The policy environment and demand for these products will help determine the size and scope of this project.

COMMONWEALTH OF VIRGINIA
DEPARTMENT OF TRANSPORTATION

WORK ORDER

Contract ID. No.: P00091296B00 FHWA No.: BH-BR03(259); BH-BR03(261) Work Order No.: 2

State Project No.: BRDG-041-718, B660; BRDG-041-719, B661 Category: MISC

Original Contract Value $ 646,308.25 Total of Other Work Orders $ 0

NOTE: If additional space is needed, use an additional sheet(s) and label as Supplemental Attachment #.

I. LOCATION AND DESCRIPTION OF PROPOSED WORK:

Time Extension

Dec. 22, 2010 to March 13, 2011 Suspension of work.
March 14, 2009 to April 15, 2011 Extension of 33days

50 days total time extension

One month additional Maintenance of Traffic

II. RESPONSIBLE CHARGE ENGINEER'S EXPLANATION OF NECESSITY FOR PROPOSED WORK:

This Work Order is needed to extend the contract time to allow the contractor to place the Asphalt Concrete TY. SM9.5A during warmer weather. Asphalt producers have shut down and will not be open until warmer weather returns. All remaining work to be completed at current contract prices.

"Burleigh Construction Company Inc. and VDOT agree that this Work Order fully resolves and settles all claims, demands or damages of any kind relating to or arising out of the work set forth in this Work Order, including but not limited to delay, impact and acceleration."

The additional Maintenance of Traffic cost are to cover the cost of rented traffic control equipment during the time when additional work was taking place.

III. FUNDING SOURCE/CHARGE Federal 80% / State 20%

IV. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT PRIOR TO APPRVOAL OF THIS WORK ORDER IS Dec. 21, 2010

V. THE FIXED DATE TIME LIMIT FOR THIS CONTRACT UPON APPROVAL OF THIS WORK ORDER IS Apr. 15, 2011

Auto-Cat for Superior Accuracy
Structure Rules Basic Logic

- Count terms that are in the list and in the first 100 words unless there are negative terms within 7 words
- Count terms that are in the list and that are within 500 words after a Document Summary Indicator unless there are negative terms within 7 words
  - Document Summary Indicators – 29 terms “Executive Summary”, “Issue Brief”, “Abstract”
- Terms in the list can be phrases or sets of terms within 7 words of each other
- Negative terms are ones that often show up but should belong to another category – they vary by category
  - Child & Family Well-being – “Coverage”, “Obesity”, “Nurses”

KAPS Group

Auto-Cat for Superior Accuracy
Overview Average Scores

	Recall	Precision	Precision Top 10
With Sections	95%	92%	99%
Full Text	71%	41%	81%
Difference	24%	51%	18%

Issues, Tips, Techniques

Auto-Cat for Superior Accuracy
Issues, Tips, Techniques

- Rule Templates
  - Separate logic and terms, not organic growth of rules and terms mixed together
- Importance of Scope Notes – even with auto-cat
  - Important during rule development
  - Important if hybrid-auto
- Category name in file name
- General rule – the more specific the category, the fewer the terms
- Empirical – try different weights, different significance thresholds

KAPS Group
Knowledge Architecture Professional Services

Conclusion

Auto-Cat for Superior Accuracy Conclusion

- Importance of content structure model
  - Bag of words = Bag of Sh..T
- Rounds of refinement
  - Step 1 – build recall – more terms
  - Step 2 – build precision – fewer terms
- Need a minimum of three rounds of refinement
  - Round 1 – 90%
  - New content accuracy drops, build to 90%
  - New content accuracy drops to 80%
- Plan on ongoing refinement and governance