



Industry Use Cases and the Underlying Content Analytics Technology used in Big Data and Predictive Analytics

Brian Swanson, Vice President Cognitive Services

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE MAY 2015		2. REPORT TYPE		3. DATES COVERED 00-00-2015 to 00-00-2015	
4. TITLE AND SUBTITLE Industry Use Cases and the Underlying Content Analytics Technology used in Big Data and Predictive Analytics				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DataSkill, Inc, 5675 Ruffin Road, Suite 100, San Diego, CA, 92123				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Presented at the 12th Annual Acquisition Research Symposium held May 13-14, 2015 in Monterey, CA.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 17	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Industry Domains



Customer Insight

- Customer experience
- Customer satisfaction and survey analysis
- Product and service quality
- Churn prediction
- Marketing campaign development and execution
- Product enhancements



Crime Analytics

- Community policing
- Investigation analytics
- Incident management
- Antiterrorism initiatives
- Antiterrorism initiatives
- Cyber crime investigation



Healthcare

- Diagnostic assistance
- Clinical treatment
- Critical care intervention
- Research for improved disease management
- Fraud detection and prevention
- Voice of the patient
- Claims management
- Prevention of readmissions
- Patient discharge and follow-up care



Insurance

- Risk assessment
- Fraud detection
- Policy and underwriting analysis
- Claims analysis, payment validation and loss review
- Reserve trending and optimization



Finance

- Anti-money laundering
- Internet banking fraud
- Operational efficiency
- Risk management and compliance

Insurance & Financial Services



Use Case

- Reduce loss ratio on claims
- Attack fraud
- Maintain optimal level of reserves

Approach

- Automate the search of 15 different data sources going back 15 years for greater insight into claim losses and insured policy lifecycle changes
- Enable knowledge-driven searches of both structured and unstructured information
- Provide one version of the truth by validating policy data across applications and databases
- Rapidly build additional internal/external data sources as needed

Benefits

- Improve risk assessment models by uncovering unexpected patterns and associations among existing data sources
- Set adequate reserves with a better understanding of the factors contributing to claims losses
- Pinpoint fraud with data mining to identify triggers that may signal bogus claims
- Save millions of dollars in staff time and get results more quickly by automating the risk assessment process

Manufacturing



The Use Case

- Quickly identify defects that can lead to recalls and negatively impact business
- Analyze defect information in a cost-effective way
- Utilize that data as feedback for the planning and development of new products
- Enhance quality, image and competitiveness, and improve customer satisfaction

The Approach

- Analyze structured information (automaker, model, year)
- Analyze unstructured information (descriptions of problems, opinions about the automaker)
- Drill down into data along several dimensions of frequency, time, deviation, trends, and more
- Provide reports that allow the user to visualize the results clearly and easily

The Benefits

- Reduce by at least 1% the cost required for handling recalls, which are estimated to cost automakers up to tens or even hundreds of billions of dollars a year
- Improve customer satisfaction and competitiveness by enabling the automakers to produce higher quality cars based on market demand as expressed in the NHTSA data
- Notify the automaker if data that match user-specified search criteria are reported to NHTSA

Education



Use Case

- Increase job placement rates for university graduates
- Gain unprecedented insight into hiring trends to align university curriculum with employers' needs
- Enhance quality, image and competitiveness, and improve customer satisfaction

Approach

- Crawl through thousands of online job postings, analyzing the unstructured data to provide an unprecedented perspective on the job market
- Aggregate the view of employers' requirements across the industry
- Monitor emerging employment trends including high-demand degrees and skills, essential concepts and methodologies, and required programming languages and product knowledge

Benefits

- Gained the ability to respond quickly and cost-effectively to changing industry needs, launching a new course in 2.5 months instead of 12 months, a 76 percent improvement
- Increased demand for new courses in business information systems to 300 percent the current capacity, demonstrating the marketplace need and the university's competitiveness
- Improved the employability of students by matching coursework to high-demand skills in the job market

Telecommunications



Use Case

- Improve customer satisfaction, secure & maintain market share
- Understand the “voice of their customer” and prevent contract cancellation
- Identify new opportunities and quickly establish new services
- Rapidly respond to incidents

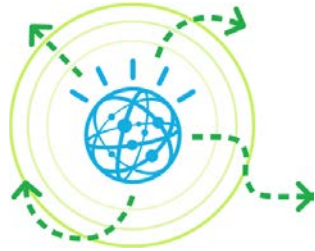
Approach

- Analyze call center notes, surveys, and customer emails
- Quickly detect likely candidates for customer churn
- Identify customer issues and suggests FAQ candidates for posting to a self-service Web site
- Mine for trends, patterns and unusual product and services associations with customer experiences

Benefits

- Improve accuracy to detect likely churn candidates by 50%
- Improve rates for model and service upgrades to loyal customers
- Improve self-service FAQ system
- Monitor voice of customer for new offerings and services

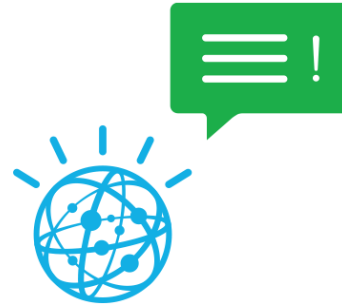
Technologies



Search

Securely connect to, search and explore all of your organization's data, regardless of format or where it is stored and managed.

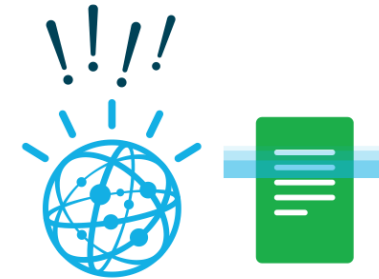
- ✓ Provision key business functions with 360-degree view of information
- ✓ Gain rapid ROI from better use and re-use of available information



Content Analytics

Mine your unstructured data to reveal trends, patterns and insights from unstructured content for high-value projects such as:

- ✓ Anticipating and identifying product defects
 - ✓ Reducing customer churn
 - ✓ Improving customer and patient care
- ... and more ...



Cognitive Services

Integrate cognitive services to enhance, scale and augment human expertise.

Embed cognitive capabilities such as:

- Question answering
 - User modeling
 - Machine translation
 - Concept expansion
- ... and more ...

Applying the Technology



Search and analytics tools provide quantitative answers e.g. the WHO, WHAT, WHERE and WHEN

Content Analytics and Cognitive services provide qualitative answers e.g. the *HOW & WHY*

The Challenge of Scale

How do you reduce big data to 'human size'?



Cognitive Services



Content Analytics




Search



Big Data

Content Analytics Technology

Text	According to finance report, IBM Corp. 's EPS increased by 10.1%.	
Identify Language	English	
Segment Sentence		
Identify Token	According to finance report IBM Corp. 's EPS increased by 10.1%	Indexing
Normalize Character Case	according	
Lemmatize Token	corporation increase	
Assign Part of Speech Tag	adjective preposition noun(singular) noun(singular) noun(singular) preposition noun(singular) noun(proper) possessive verb(past tense) numeral	Built -in Facet
Identify Domain Specific Term	IBM Corp. EPS	Custom Facet
Extract Domain Specific Phrase	IBM Corp. 's EPS 10.1% Positive (finance – increase) ↑	

Content Analytics Challenges

Words have multiple Part-of-Speech tag candidates commonly:

- “according”: adjective / verb (present participle)
- “finance”: noun (singular) / verb (base form / present tense)
- “report”: noun (singular) / verb (base form / present tense)
- “s”: possessive / has / is / was
- “increased”: verb (past tense / past tense participle)

Upper case character doesn't always indicate sentence beginning. It is also used for:

- abbreviation
- proper noun (e.g. place, organization, people name)
- normal noun in several languages (e.g. German)
- title (e.g. chapter, news article, book)
- enumeration (e.g. A. B. C.)

Latin alphabet doesn't always indicate English text. It is commonly used for other languages too (e.g. French, Spanish, etc.)

Period doesn't always indicate sentence ending. It is also used for:

- abbreviation
- decimal point
- 1000 separator in several languages (e.g. German)
- enumeration (e.g. A.B.C.)

According to finance report, IBM Corp.'s EPS increased by 10.1%.

Need to identify phrasal expressions by scanning minimum number of tokens

Need to store millions of words in small memory
Need to achieve high throughput for looking up

Token boundary doesn't always have white space character
Several east Asian languages doesn't use any indicators for token boundaries. It is determined by context. (e.g. Japanese, Chinese, Korean, Thai)

“EPS” doesn't always mean “Earnings Per Share”. It has different meaning in different domain.
e.g. Wikipedia lists 35 different meanings for “EPS”:
- “External Power Supply”
- “European Protected Species”
- “Electro-Plasma System” :-)

Company name is a domain specific term.
For finance domain, it needs to recognize all companies names listed on NYSE at least. Though it is not enough at all for analyzing finance report from other countries outside U.S.

Content Analytics Example

Content Analytics with Natural Language Processing describes a set of linguistic, statistical, and machine learning techniques that allow text to be analysed and key information extraction for business integration

Scalable Approach to Understanding and Extracting Language

1. Language Detection
2. Parts of Speech
3. Phrase Constituents (Concepts and Context)
4. Higher Level Extractions (NER, Sentiment, Custom)

EC 4.0 Cu. Ft.
26-Cycle King-Size Washer –
White. I hate this machine. Have had 3 calls on machine. You can't wash **large items**, Wont' clean in the middle. **Leaves dry spots** through the clothes, I can only do **1/2 basket** of clothes. Will **not clean** or **mix bleach** in with the water.....



Product	EC
Category	washer
Size	4.0 Cu. Ft
Model	26-Cycle King Size
Color	white
Issue	large items
Issue	leaves dry spots
Issue	1/2 basket
Issue	not clean
Issue	mix bleach

Data Mining Unstructured Data

The image displays six screenshots of the IBM Watson Content Analytics interface, each with a blue callout box highlighting a specific feature:

- Document Analysis:** Shows a document titled "944379.xml" with a list of extracted entities such as "Part of Speech", "Phrase Constituent", "My Keywords", "vehicle information", "incident info", "Rumor", "Weather", "Condition", and "Flags".
- Facets:** Shows a list of facets for a document, including "Part of Speech", "Noun Phrase", "Predicate Phrase", "Verb - Noun", "Conjunction Phrase", "My Keywords", "vehicle information", "manufacture", "make", "model", "year", "anti-lock brakes", "cruise control", "date purchased", "dealer information", "drive type", and "Data Facet Navigation".
- Dashboard:** Shows a dashboard with various charts and graphs, including a line chart and a bar chart, with a legend for "Part of Speech".
- Time Series:** Shows a time series chart with a date facet, time scale, and maximum time units per chart.
- Sentiment:** Shows a sentiment analysis chart with a legend for "Sentiment" (Positive, Ambivalent, Negative) and a bar chart showing the distribution of sentiment values.
- Connections:** Shows a network graph with nodes representing terms and edges representing relationships, including terms like "be ... issue", "be ... problem", "be ... anything", "be ... it", "be ... common", "be ... safety", "park ... lot", "park ... brake", "SILVERADO", and "MAL".
- Facet Pairs:** Shows a table of facet pairs with columns for "Row", "Column", and "Value". The table includes terms like "take ... vehicle", "drive ... home", "slide ... door", "be ... noise", "find ... anything", "lower ... ball", "SIENNA", "TOWN AND COUNTRY", "F-150", "COBALT", "GRAND AM", "BLAZER", and "ALTIMA".
- Deviations / Trends:** Shows a chart of deviations and trends for various manufacturers, including "GENERAL MOTORS CORP (11407)", "FORD MOTOR COMPANY (89617)", "DAIMLERCHRYSLER CORPORATION (8099)", and "TOYOTA MOTOR CORPORATION (8754)".

Cognitive Services



Question Answer

Direct responses to users inquiries fueled by primary document sources



Machine Translation

Globalize on the fly. Translate text from one language to another.



User Modeling

Personality profiling to help engage users on their own terms.



Relationship Extraction

Intelligently finds relationships between sentences components (nouns, verbs, subjects, objects, etc.)



Message Resonance

Communicate with people with a style and words that suits them



Visualization Rendering

Graphical representations of data analysis for easier understanding



Concept Expansion

Maps euphemisms or colloquial terms to more commonly understood phrases



Language Identification

Identifies the language in which text is written

Informed Decision Making: Search vs. Expert Q&A

Decision Maker

Has Question

Distills to 2-3 Keywords

Reads Documents, Finds Answers

Search Engine

Finds Documents containing Keywords

Expert

Understands Question

Produces Possible Answers & Evidence

Analyzes Evidence, Computes Confidence

Delivers Response, Evidence & Confidence

Decision Maker

Asks NL Question

Considers Answer & Evidence

Cognitive Q & A Technology

