



Maurizio Gallotti

ECM Client Technical Professional

IBM Software | Industry Solutions | Enterprise Content Management

Content Analytics

Acquisire conoscenza dalle informazioni strutturate e non strutturate

**IBM Enterprise
Content Management**

Contenuti al centro per decisioni più intelligenti





IBM Content Analytics

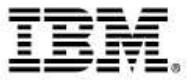
l'esplosione delle informazioni



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





- Nella ricerca full-text i criteri sono espressi mediante termini o espressioni regolari
- La Business Intelligence mostra dati e informazioni ricavate dai metadati disponibili dalle sorgenti dati strutturate
- **Content Analytics** consente di:

- estrarre, ordinare ed analizzare informazioni di valore contenute nei documenti

- semplificare la condivisione del patrimonio informativo aziendale

- facilitare e rendere efficace l'analisi di grandi quantità di documenti



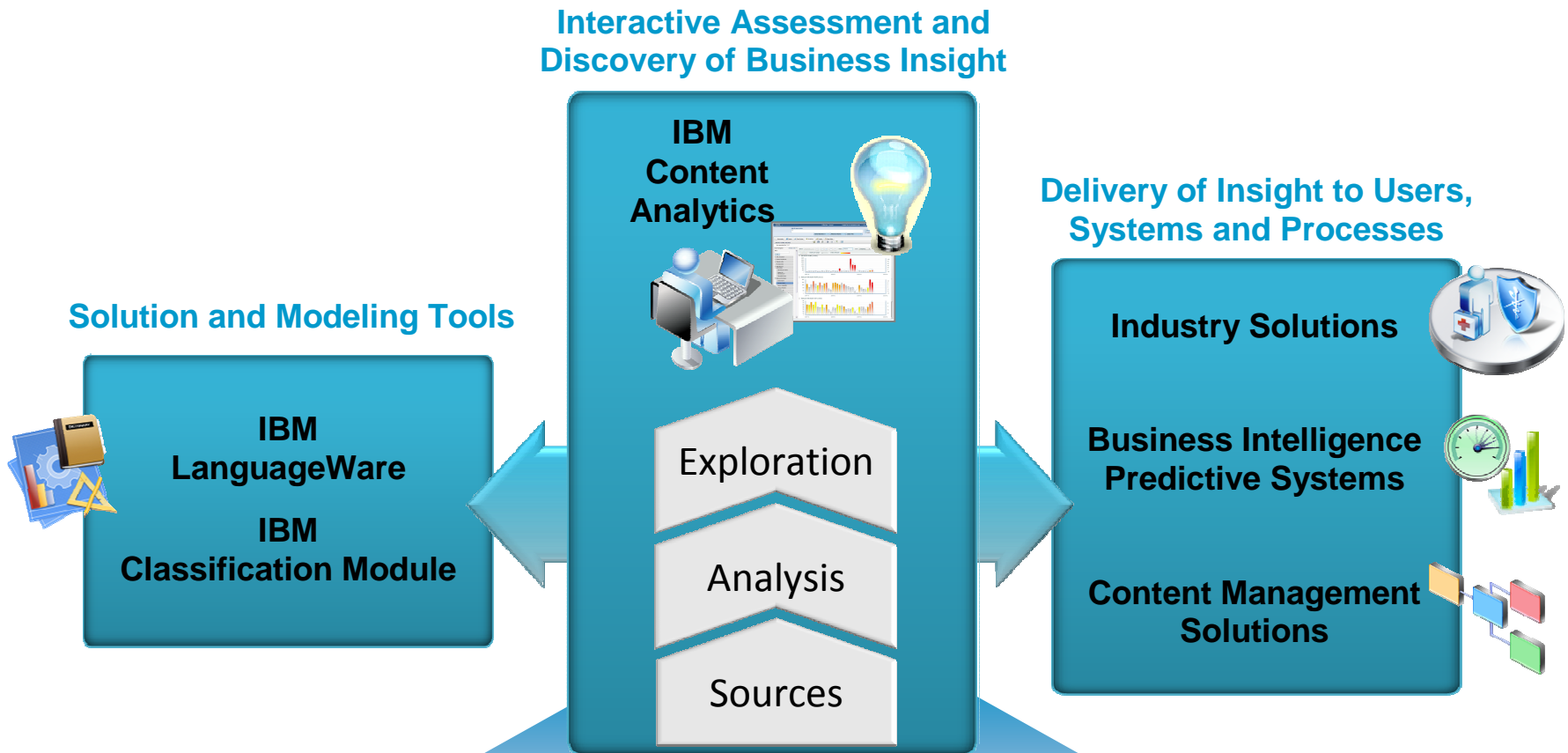
***Trova ed estrae
informazioni,
effettua sui dati non
strutturati le stesse
analisi che vengono
attualmente fatte sui
dati strutturati***

- Combina dati strutturati e non strutturati rendendo l'analisi agevole e fluida
- Permette di analizzare in modo semplice e veloce i contenuti esplorandoli da diverse prospettive
- Identifica ed evidenzia automaticamente qualsiasi relazione inusuale tra i contenuti
- Si integra facilmente con gli strumenti di business intelligence Cognos
- Accede a tutte le tipologie di documento e a numerosissime fonti documentali
- Offre la possibilità di espandere le capacità di analisi, compresa quella semantica, attraverso tecnologie aperte





IBM Content Analytics *caratteristiche*



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





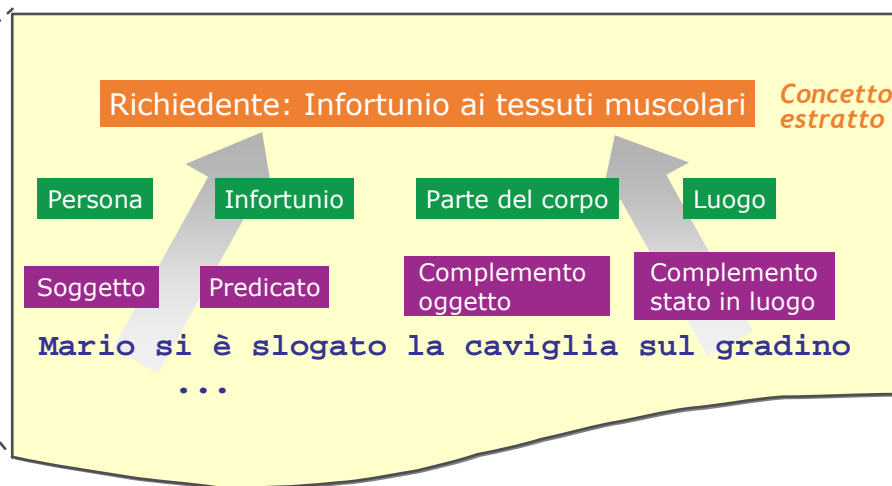
IBM Content Analytics

La chiave di volta tecnologica: l'analisi dei contenuti



Content Analytics

Basata su UIMA, un'architettura aperta e standard di mercato per l'analisi dei testi, creata da IBM. Attualmente UIMA è uno standard OASIS ed è implementata da un progetto open-source Apache.



Documento analizzato
con identificazione dei concetti

- Content Analytics comprende la struttura di una frase e crea degli indici che facilitano l'esplorazione delle "informazioni" contenute nei documenti
- Estrae:
 - **Entità**, identificando item individuali come nomi, compagnie, prodotti, date, persone, luoghi, prezzi, ..
 - **Fatti**, definiti come collezioni di entità che identificano eventi, ruoli, circostanze, azioni, incidenti,
 - **Concetti**, definiti come collezioni di fatti che identificano tendenze, comportamenti, (es. sentimenti, reputazioni, frodi, affinità..)

IBM Enterprise Content Management

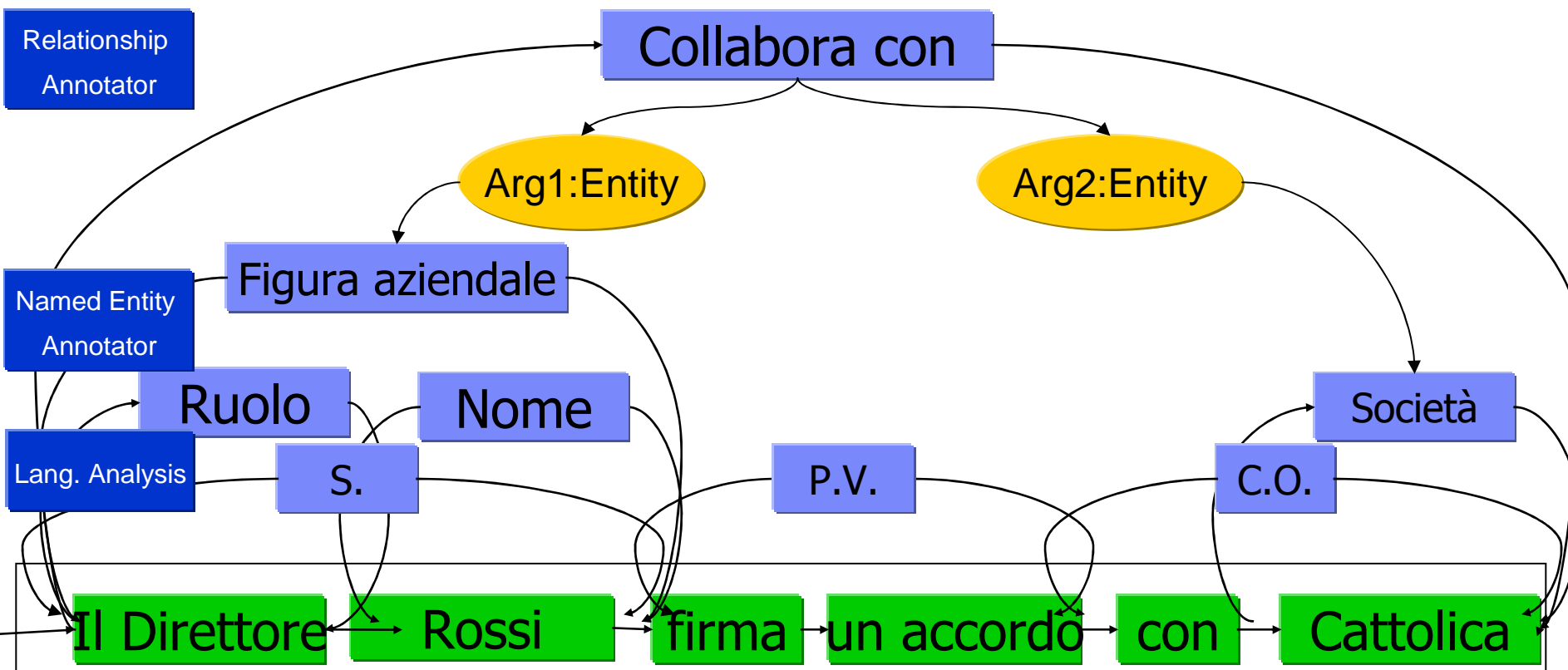
Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics

Analisi dei contenuti



<FiguraAziendale><Ruolo>Direttore</Ruolo><Persona>Rossi</Persona></FiguraAziendale> firma un accordo <Società>Cattolica</Società>

IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.





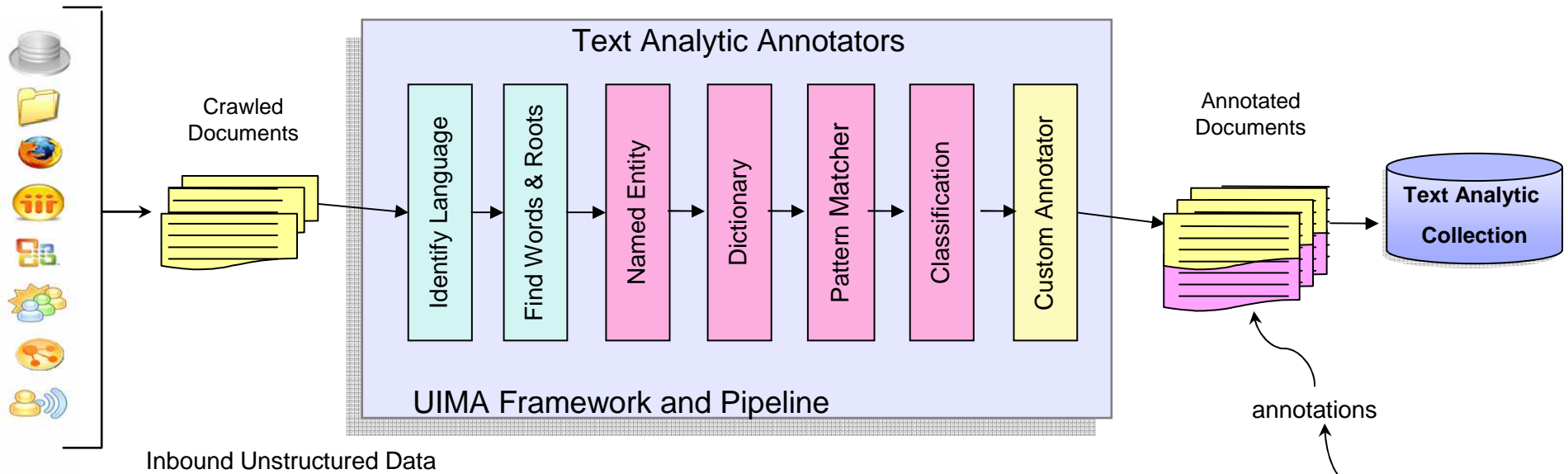
IBM Content Analytics Unstructured Information Management Architecture (UIMA)

UIMA Background

- Created by IBM in the late 1990s
- Accepted into the Apache Incubator in 2006
- Approved as OASIS Standard in 2009

ICA Implementation

- Multiple languages
- Multiple best of breed analytic technologies
- Open & customizable text analytics pipeline



Device Malfunction Description:

It was reported that during a gastric bypass roux-en-y procedure, on the 3rd firing with a blue load on the stomach there was bleeding. Not sure if staples were formed properly. They over sewed the staple line. There was no PT consequence.

Extracted / Derived Information

| | |
|---------------------|------------------------|
| Involved Body Part | stomach |
| Type of Injury | bleeding |
| Procedure Performed | gastric bypass surgery |

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics *UIMA and Watson*

IBM WATSON

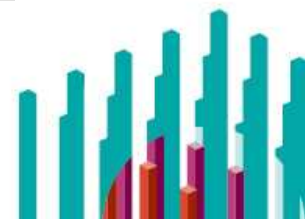
< What is Watson? > Why Jeopardy!? >

THIS NAME OF A BOY
WHO CAN FLY
CAN BE APPLIED
TO ANY EMOTIONALLY
UNDEVELOPED MAN

Who is Peter Pan ?

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





EVEN A BROKEN
ONE OF THESE
ON YOUR WALL
IS RIGHT
TWICE A DAY

ANCHE SE ROTTO
UNO DI QUESTI
SUL TUO MURO
DICE IL GIUSTO
DUE VOLTE AL GIORNO

Il supercomputer che partecipa al quiz tv e sbaraglia i concorrenti umani

A «Jeopardy», sulla Nbc, dalle risposte bisogna risalire alle domande. Battuti i due campioni dello scorso anno



la sfida a Jeopardy

MILANO - Qualcuno l'ha definito «il discendente più brillante di Hal 9000», il computer dell'astronave di 2001: *Odissea nello spazio*, perché come la macchina presente nel film di Stanley Kubrick ha una singolare «intelligenza artificiale». Si chiama Watson, è un supercomputer ideato dall'Ibm e in questi giorni è il protagonista assoluto di *Jeopardy*, uno dei giochi a quiz più famosi d'America, che va in onda sulla Nbc dal 1964. Nonostante gareggi con Brad Rutter e Ken Jennings, i due più forti campioni dell'anno passato, Watson sta dimostrando di essere il numero uno e nei primi due giorni di gara ha letteralmente sbaragliato gli avversari.

IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics *UIMA and Watson*



- Watson è disegnato secondo lo standard UIMA. Questa architettura software è lo standard per lo sviluppo di programmi che analizzano informazioni non strutturate come testo, audio ed immagini.

- Watson utilizza Apache UIMA per sfruttare il “natural language processing” in parallelo sui processori POWER7, consentendo di effettuare migliaia di computazioni analitiche simultaneamente sul cluster di server per rispondere ad ogni domanda nel più breve tempo possibile.

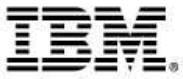
Link utili per approfondimenti :

- http://www-03.ibm.com/innovation/us/watson/?S_CMP=swnews211
- <http://www-03.ibm.com/innovation/us/watson/watson-for-a-smarter-planet/watso>

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics Funzionalità

Scopre:

Identifica ed etichetta automaticamente gli attributi e le entità fondamentali all'interno dei contenuti,

Analizza tutte le sorgenti di contenuti, estraendo le parole chiave ed elementi delle frasi

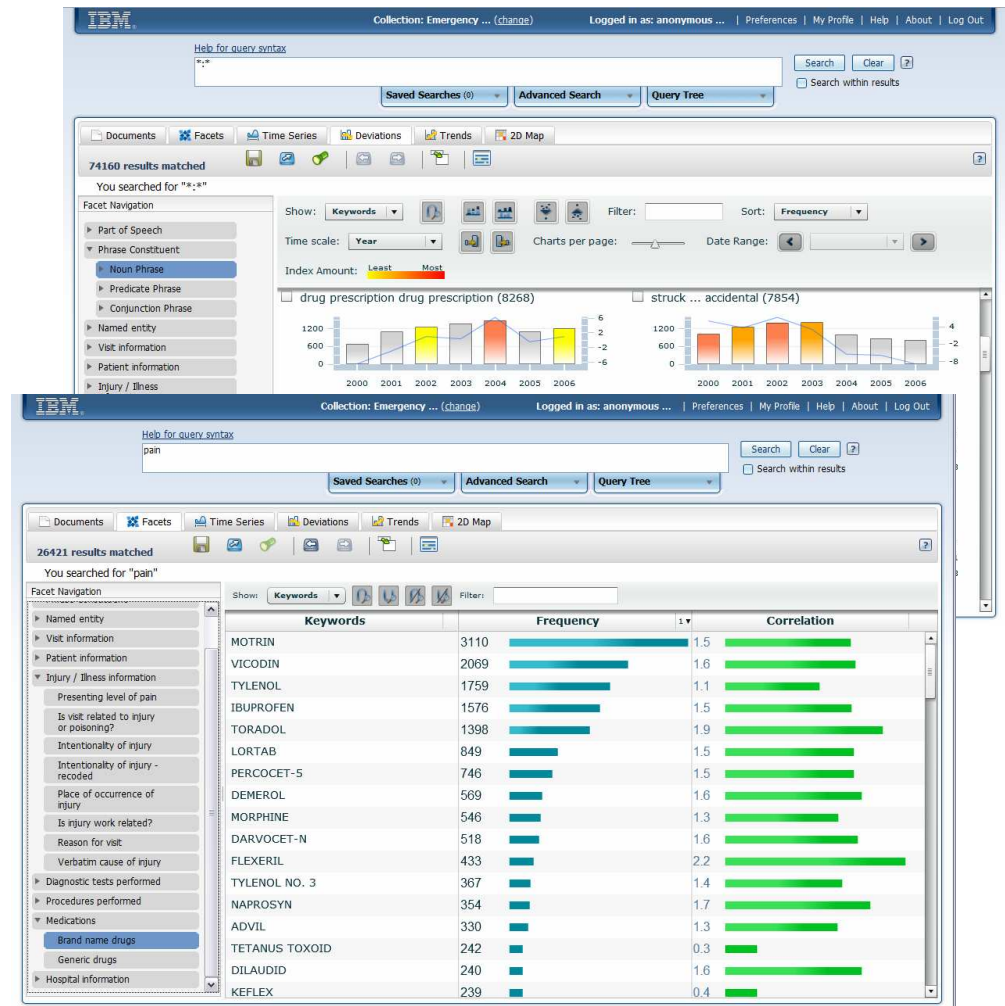
Raffina:

Consente la navigazione ed il drill-down sulla base degli attributi, delle entità e delle dimensioni estratte

Visualizza:

Utilizza una modalità avanzata di visualizzazione che consente il mining.

Evidenzia deviazioni ed anomalie in modo da consentire di prendere decisioni ed intraprendere azioni correttive



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics

Funzionalità di Text Mining

• Diverse viste dei dati

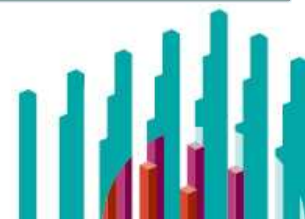
- Documenti
- Facets
- Serie temporale
- Deviazioni
- Trend
- Facet Pairs
- Connections
- Dashboard

The collage displays several screenshots of the IBM Content Analytics interface, each highlighting a different data view:

- Documents:** Shows a search results page with a list of documents and a search bar.
- Facets:** Displays a list of facets (keywords) with their frequency and correlation.
- Time Series:** Shows a bar chart representing data over time.
- Trends:** Shows a line graph representing trends over time.
- Deviations:** Shows a bar chart representing deviations from a baseline.
- Connections:** Shows a network graph with nodes and edges, representing relationships between entities.
- Dashboard:** Shows a comprehensive view with multiple charts, including a network graph, a bar chart, and a pie chart.

IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics

Funzionalità di Text Mining: FACETS

- Attraverso le facets è possibile esplorare i contenuti sia attraverso i dati strutturati disponibili che le informazioni estratte dai documenti
- Consente di visualizzare i contenuti aggregati per keyword
- Mostra frequenza ed indici di correlazione tra facets
- Drill-down tramite l'aggiunta della parola chiave selezionata alla condizione di ricerca

The screenshot displays the IBM Content Analytics interface. The top navigation bar includes tabs for Documents, Facets, Time Series, Deviations, Trends, Facet Pairs, Connections, Dashboard, and Reports. The main content area shows a search result of 142516/400158 results matched. The Facet Navigation panel on the left lists categories like Vehicle/Equipment Corp, Vehicle/Equipment Make, Model, Model Year, and Component Description. The main table displays Keywords, Frequency, and Correlation. The table shows two rows: FORD with a frequency of 20478 and a correlation of 0.9, and CHEVROLET with a frequency of 18795 and a correlation of 1.0. A second screenshot below shows a search result of 91641/400158 results matched. The Facet Navigation panel on the left lists categories like Verb - Noun, Conjunction Phrase, Named entity, Troubles, and Negative event. The main table displays Keywords, Frequency, and Correlation. The table shows three rows: be ... problem with a frequency of 5005 and a correlation of 1.0, be ... recall with a frequency of 3246 and a correlation of 1.2, and have ... problem with a frequency of 1807 and a correlation of 1.1.

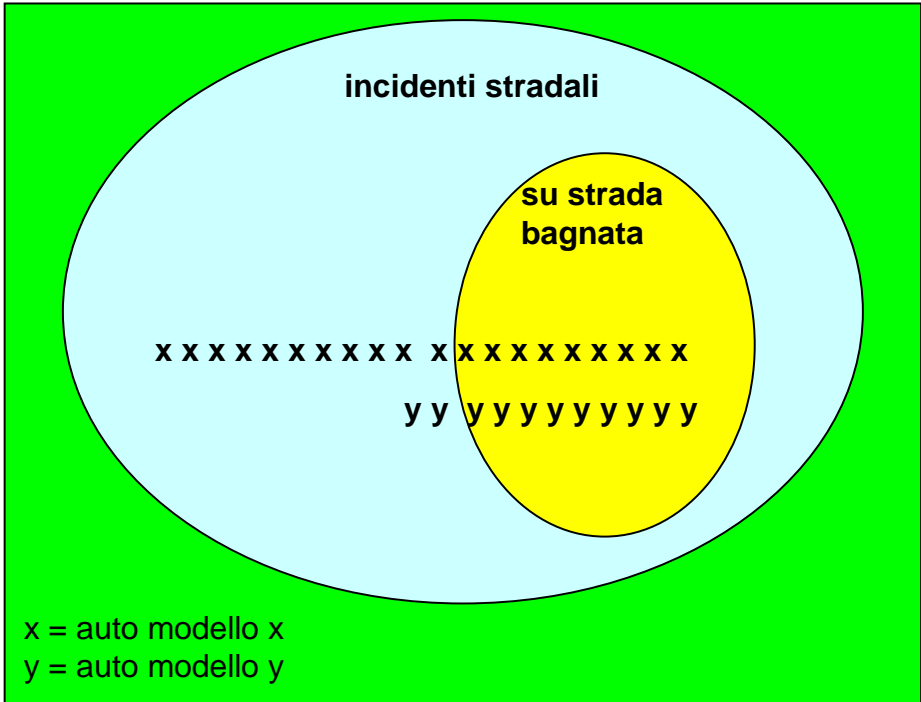
| Keywords | Frequency | Correlation |
|-----------|-----------|-------------|
| FORD | 20478 | 0.9 |
| CHEVROLET | 18795 | 1.0 |

| Keywords | Frequency | Correlation |
|------------------|-----------|-------------|
| be ... problem | 5005 | 1.0 |
| be ... recall | 3246 | 1.2 |
| have ... problem | 1807 | 1.1 |

IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.





Frequenza auto x = 20
 Frequenza auto y = 11

SU STRADA BAGNATA
 frequenza auto x = 10
 frequenza auto y = 9

NEL CONTESTO
 incidenti stradali su strada bagnata

VALORE CORRELAZIONE PIU' ELEVATO PER AUTO y

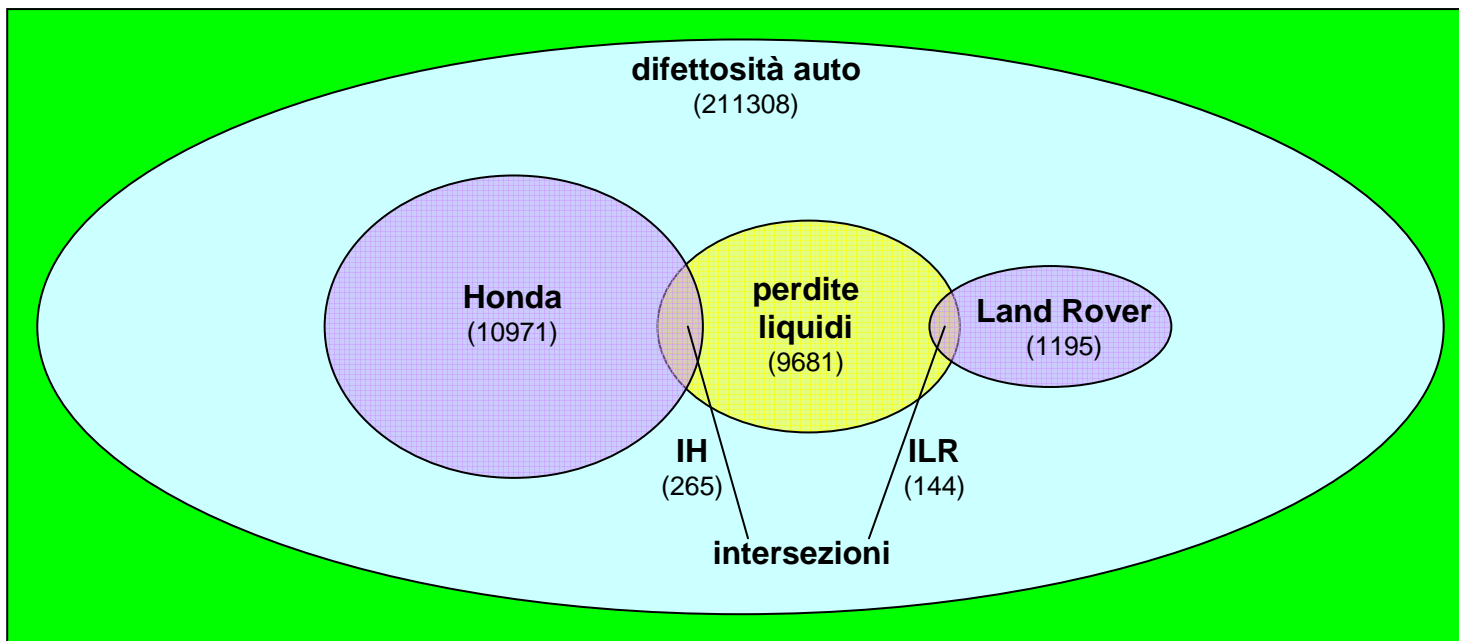
- La frequenza esprime il numero di documenti che contengono la keyword (evento, oggetto, ecc.)
- La correlazione indica la rilevanza reciproca di due sottoinsiemi all'interno dell'aggregato preso in considerazione.





IBM Content Analytics

Funzionalità di Text Mining: Frequenza e Correlazione



densità = occorrenze sottoinsieme / totale

| | occorrenze | densità |
|-----------------|------------|---------|
| Honda | 10971 | 0,05190 |
| Land Rover | 1195 | 0,00565 |
| perdite liquidi | 9681 | 0,04580 |
| IH | 265 | 0,00125 |
| ILR | 144 | 0,00068 |

correlazione = $\frac{\text{densità intersezione}}{\text{prodotto densità sottoinsiemi interessati}}$ * indice di correlazione

$$\text{correlazione Honda - perdite} = \frac{0,00125}{0,05190 * 0,04580} * 0,77 = 0,4$$

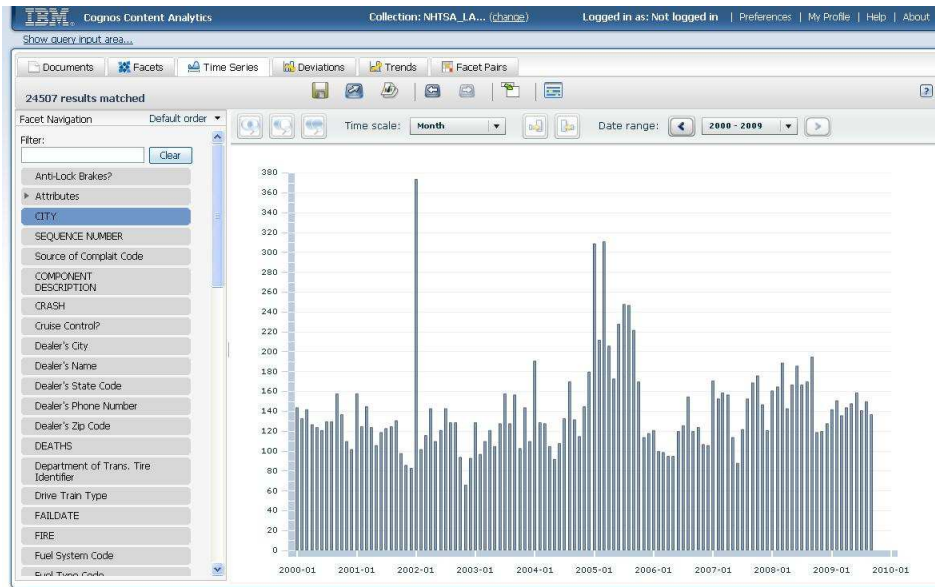
$$\text{correlazione Land Rover perdite} = \frac{0,00068}{0,00565 * 0,04580} * 0,80 = 2,1$$



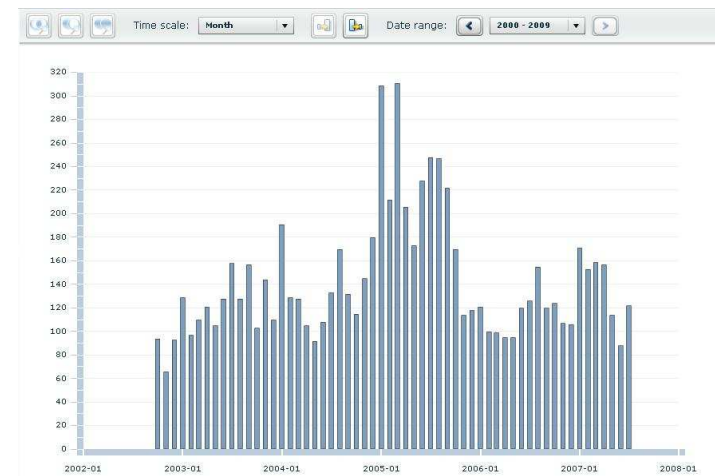
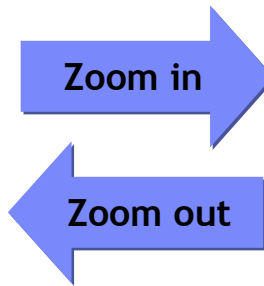
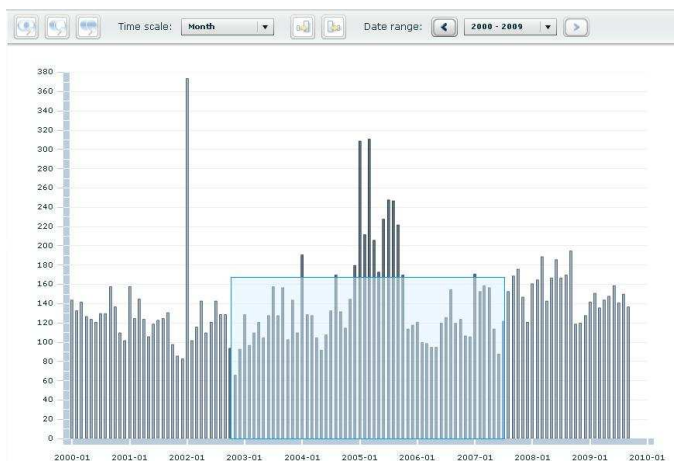


IBM Content Analytics

Funzionalità di Text Mining: Serie temporali



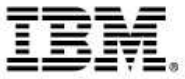
- Mostra quanto spesso i documenti soddisfano la condizione di ricerca in un determinato periodo di tempo
- Consente di raffinare la ricerca tramite la selezione di una o più vincoli di date



IBM Enterprise
Content Management

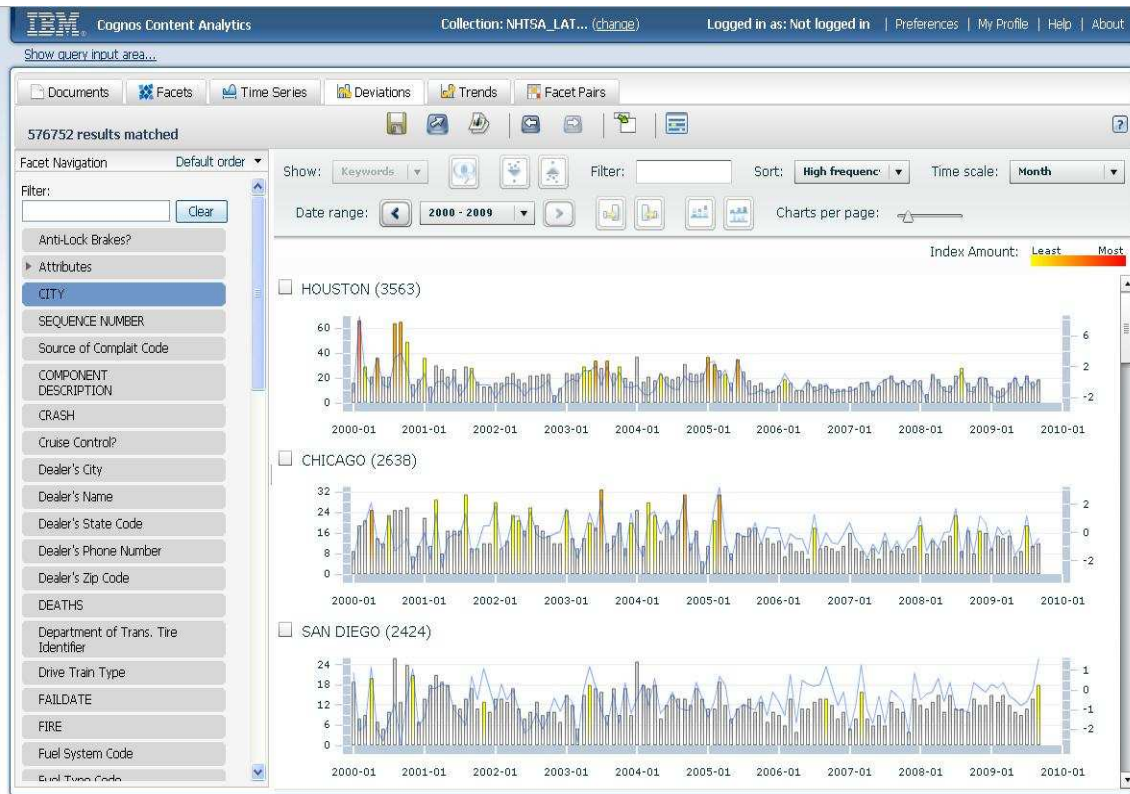
Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics

Funzionalità di Text Mining: Deviazioni e Trend



- Mostra la frequenza delle parole chiave ed i valori calcolati
- Deviazioni
 - Mostra incrementi inaspettati della frequenza di una parola chiave nel tempo
- Trend
 - Mostra incrementi inaspettati della frequenza di un insieme di parole chiave (facet) nel tempo
- Permette di raffinare la ricerca usando il valore della facet e la data
- Ordina, filtra, zoom-in, zoom-out

IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.





IBM Content Analytics

Funzionalità di Text Mining: Facet Pairs

- Mostra la correlazione tra parole chiave appartenenti a due facet diverse
- 3 modalità di visualizzazione

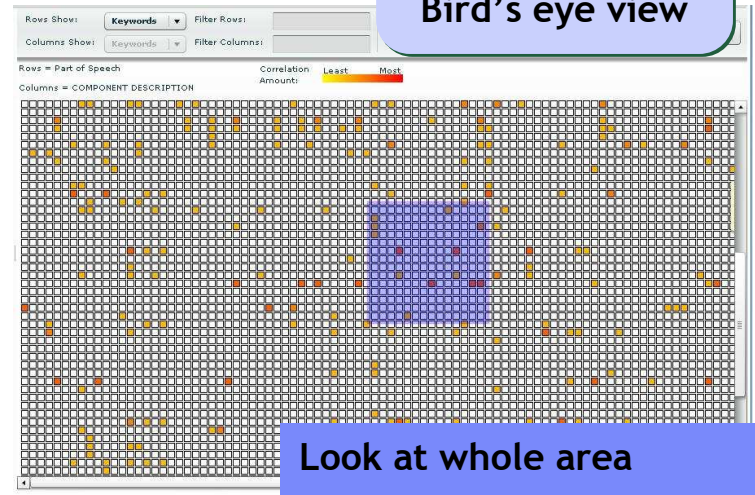


Table view

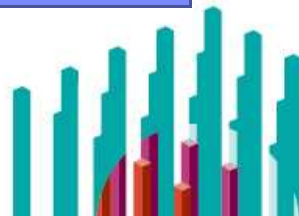
| Rows:Part of Speech | Columns:COMPONENT DES... | Frequency | Correlation |
|---------------------|-------------------------------|-----------|-------------|
| transmission | POWER TRAIN:AUTOMATIC TRANS | 28523 | 10.7 |
| brake | SERVICE BRAKES, HYDRAULIC:AN1 | 27847 | 5.1 |
| be | POWER TRAIN:AUTOMATIC TRANS | 24898 | 1.1 |
| and | POWER TRAIN:AUTOMATIC TRANS | 23393 | 1.1 |
| AK | SERVICE BRAKES, HYDRAULIC:AN1 | 19943 | 1.5 |
| have | POWER TRAIN:AUTOMATIC TRANS | 18711 | 1.2 |
| be | ENGINE AND ENGINE COOLING:EM | 18572 | 1.0 |
| be | SERVICE BRAKES, HYDRAULIC:AN1 | 18230 | 0.9 |
| and | SERVICE BRAKES, HYDRAULIC:AN1 | 18162 | 1.0 |
| and | ENGINE AND ENGINE COOLING:EM | 17190 | 1.1 |
| tire | TIRES | 16691 | 10.9 |
| not | POWER TRAIN:AUTOMATIC TRANS | 16362 | 1.1 |
| vehicle | POWER TRAIN:AUTOMATIC TRANS | 16245 | 1.1 |
| engine | ENGINE AND ENGINE COOLING:EM | 16044 | 4.7 |
| vehicle | SERVICE BRAKES, HYDRAULIC:AN1 | 15273 | 1.2 |
| AK | POWER TRAIN:AUTOMATIC TRANS | 14906 | 4.0 |

Quick filter and sort

Grid view

| Subfacets/ Keywords | VEHICLE SPE... | EQUIPMENT 4891 | ELECTRICAL S... | SUSPENSION:... | STEERING:W... | SUSPENSION:... | SERVICE BRA... | ENGINE AND |
|---------------------|----------------|----------------|-----------------|----------------|---------------|----------------|----------------|------------|
| get 46664 | 554 | 737 | 379 | 334 | 208 | 273 | 215 | 314 |
| back 45668 | 325 | 702 | 326 | 467 | 218 | 188 | 222 | 262 |
| if 44800 | 309 | 688 | 306 | 384 | 201 | 293 | 180 | 297 |
| dealership 44526 | 408 | 515 | 344 | 348 | 251 | 454 | 291 | 349 |
| steer 44035 | 118 | 339 | 339 | 231 | 8001 | 361 | 150 | 81 |
| due 43963 | 226 | 410 | 500 | 397 | 261 | 414 | 409 | 439 |
| make 43599 | 291 | 576 | 235 | 519 | 521 | 343 | 311 | 364 |
| turn 42519 | 377 | 426 | 394 | 272 | 817 | 304 | 218 | 139 |
| seat 42442 | 70 | 299 | 204 | 109 | 75 | 26 | 36 | 56 |

See in detail

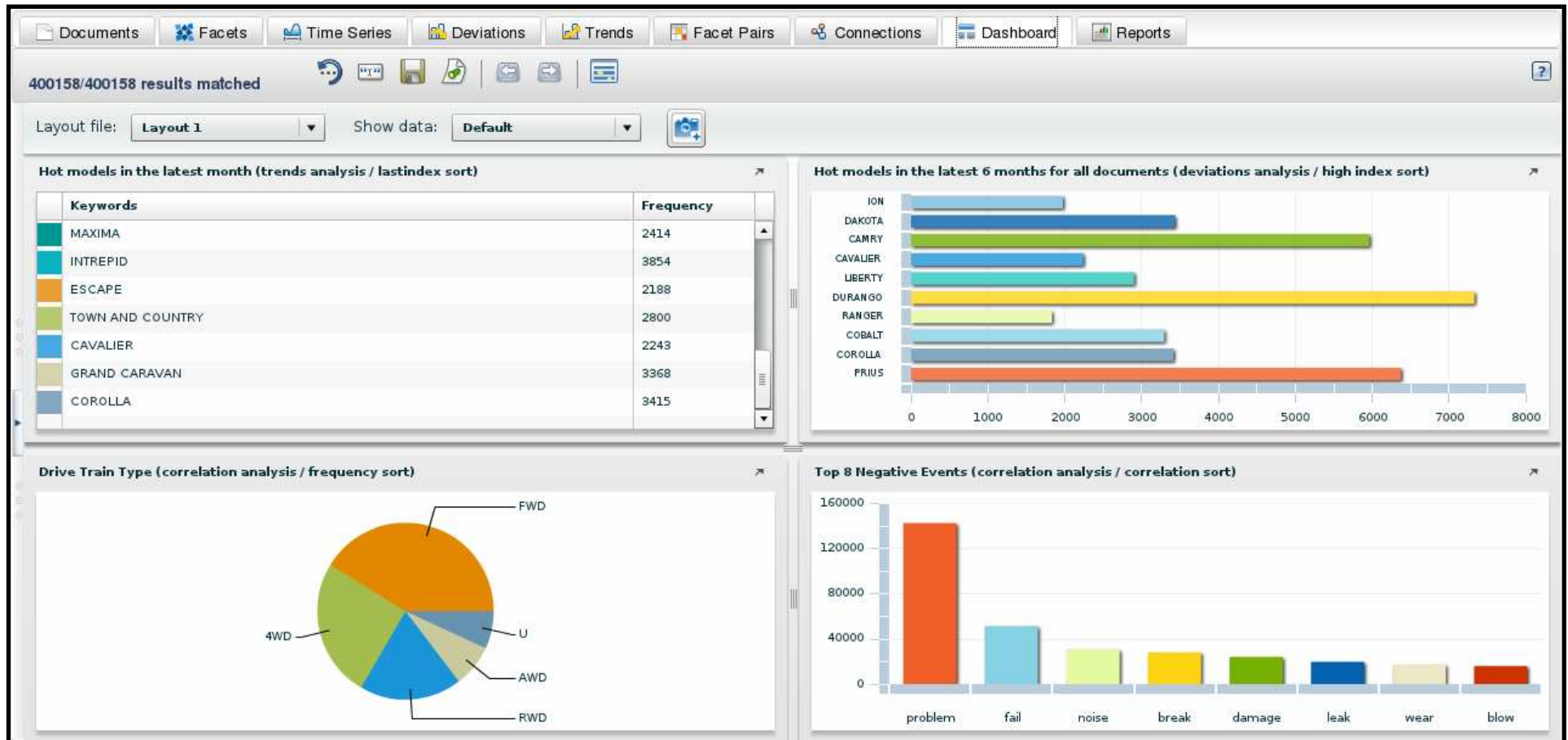




IBM Content Analytics

Funzionalità di Text Mining: Dashboard

La vista **Dashboard** mostra varie tavole predefinite in vista unica, per visualizzare e correlare vari aspetti dei dati analizzati



IBM Enterprise
Content Management

Contenuti al centro per decisioni più intelligenti.








IBM Content Analytics









Fonti Documentali analizzabili

Web

- HTTP
- HTTPS
- WebSphere Portal  Web pages
- WebSphere Portal  Document Manager
- IBM Workplace Web  Content Management
- Newsgroup (NNTP)

Collaboration

Lotus. **Microsoft**

- Lotus Notes databases 
- Domino.doc 
- QuickPlace 
- Quickr 
- Lotus WCM 
- Lotus Connections 
- MS Exchange public folders
- Microsoft SharePoint  Services (2003 & 2007)
- Windows file systems 
- UNIX file systems

ECM



Database

DB2.

SYBASE

ORACLE

Informix

Microsoft



IBM Websphere MQ
with Event Publisher

Mainframe:
VSAM, IMS, CA-
Datacom,
Software AG
Adabas

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.



Native security support





Domande ?

