Giorgio Sicurella

ECM Architect

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

Content Analytics
l'anello mancante tra la ricerca full-text e la Business Intelligence

- 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:
  - semplificare la condivisione del patrimonio informativo aziendale;
  - facilitare e rendere efficace l'analisi di grandi quantità di documenti;
  - ricavare indicazioni e rilevare tendenze relative alla propria attività dall'esame dei contenuti "non strutturati" o "semi strutturati"

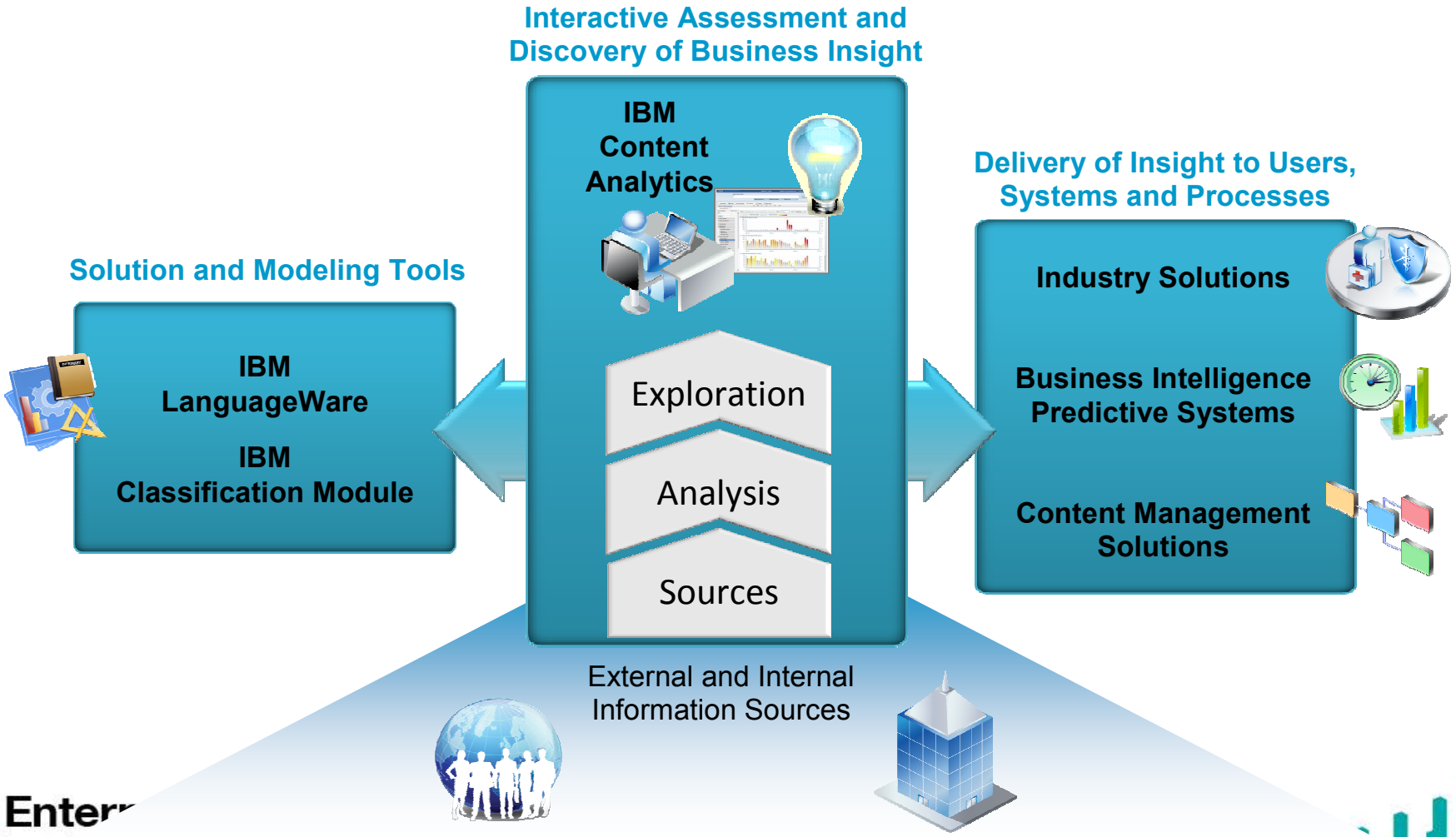
IBM Content Analytics
caratteristiche

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

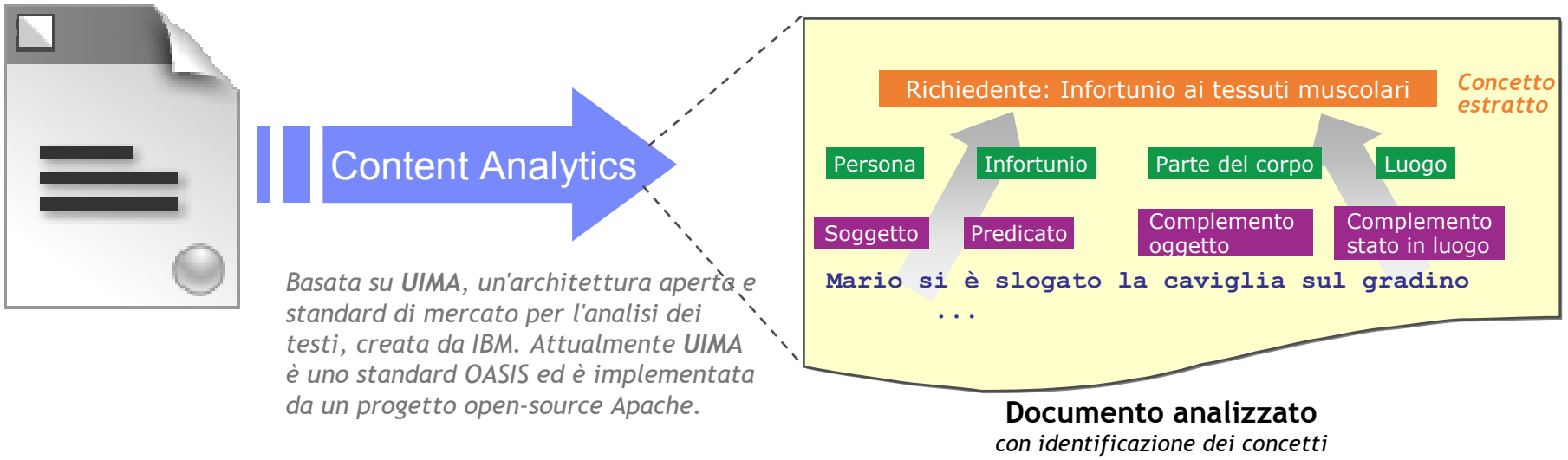
IBM Content Analytics
caratteristiche



IBM Enter
Content Management
Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics

La chiave di volta tecnologica: l'analisi dei contenuti

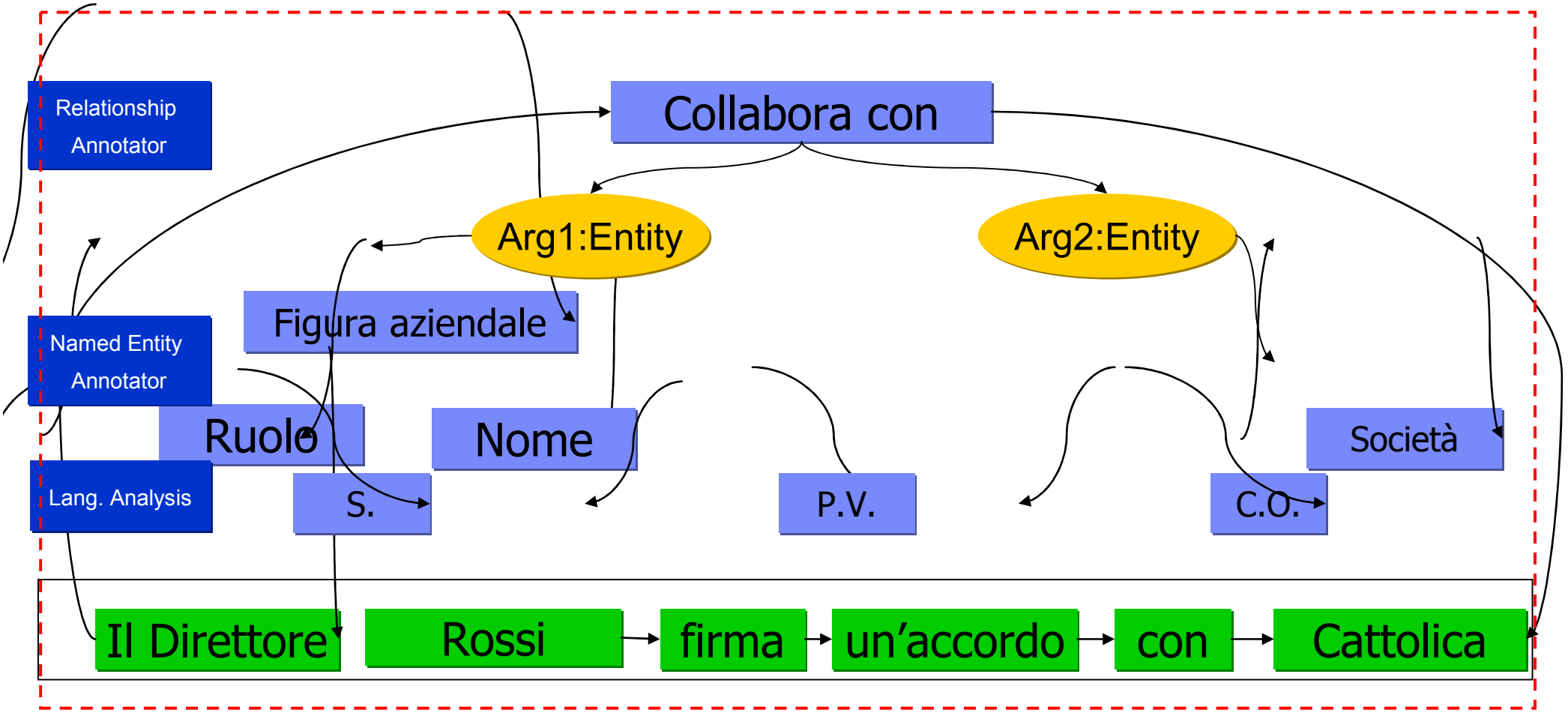


*Basata su **UIMA**, un'architettura aperta e standard di mercato per l'analisi dei testi, creata da IBM. Attualmente **UIMA** è uno standard OASIS ed è implementata da un progetto open-source Apache.*

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 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
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

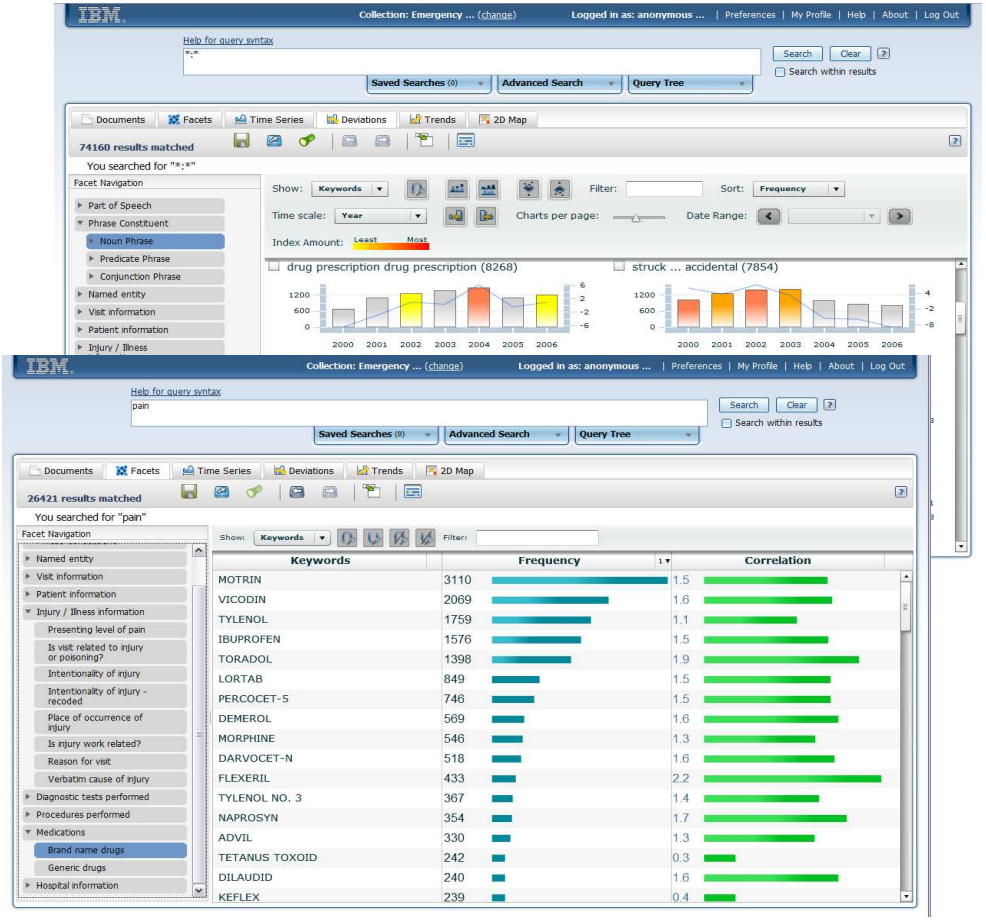
Raffina:

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

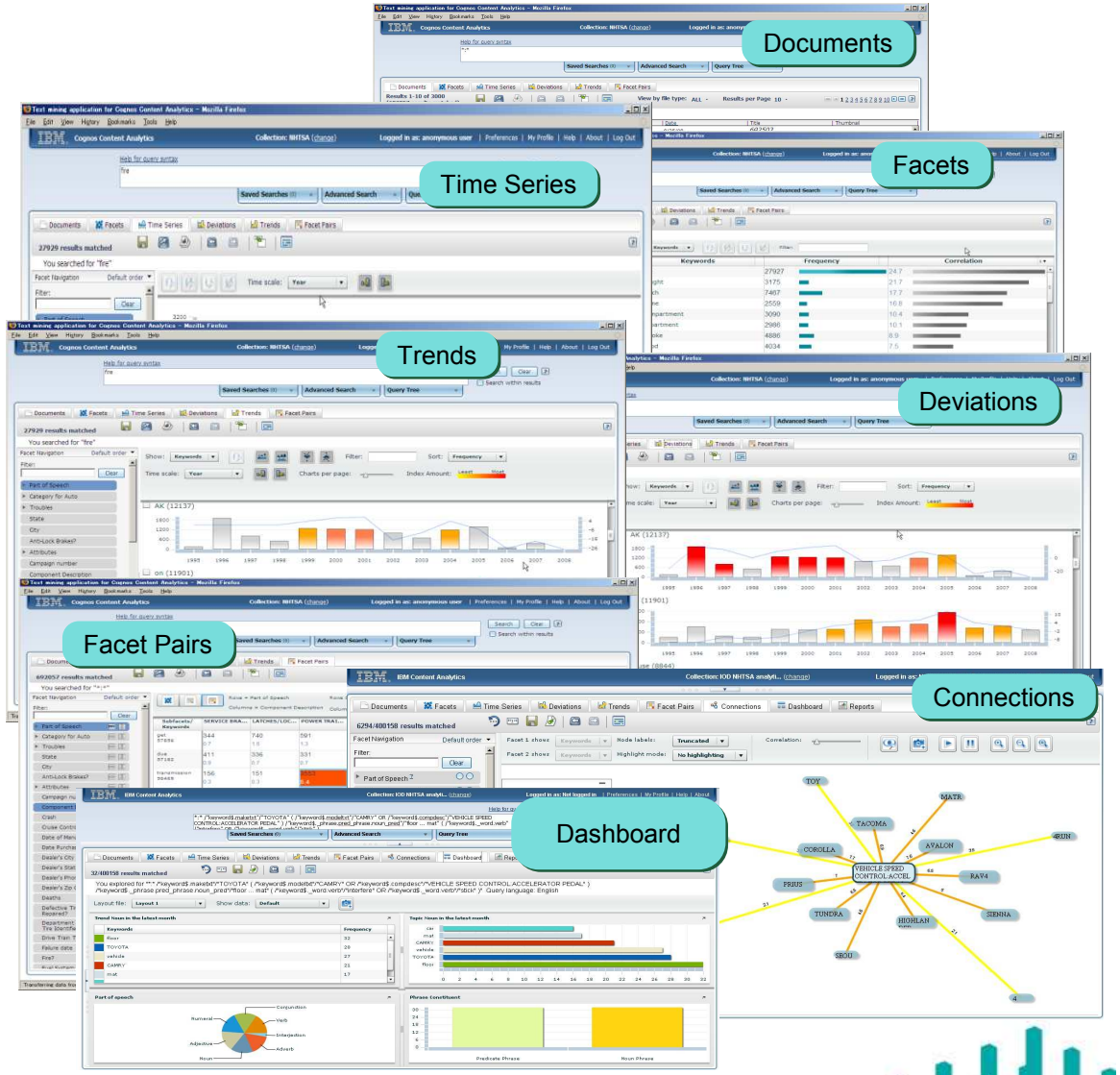


IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining

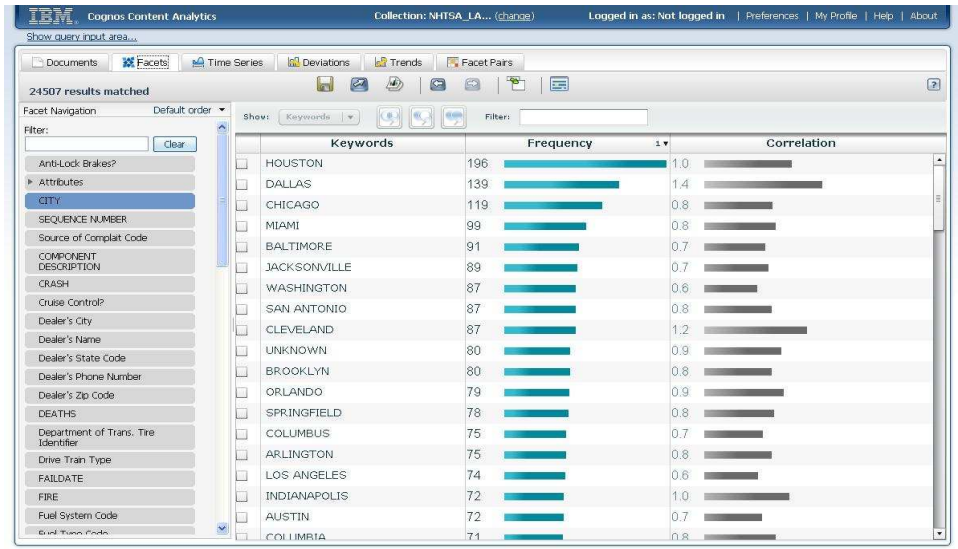
- **Estrazione interattiva delle informazioni**
  - Facilita il reperimento di documenti che hanno indici valorizzati in modo distinto rispetto ad un determinato aspetto dell'analisi
- **Diverse viste dei dati**
  - **Documenti** elenca i documenti selezionati da una query
  - **Facets** aggrega i contenuti in base alla "prospettiva" relativa alla facet
  - **Serie temporale** mostra come cambia nel tempo la frequenza di tutti i contenuti selezionati
  - **Deviazioni** mostra come cambia nel tempo la frequenza dei contenuti evidenziando le deviazioni dalla media
  - **Trend** come la vista deviazioni, ma evidenzia gli innalzamenti anomali della frequenza
  - **Facet Pairs** mostra la correlazione tra i valori di due facet diverse
  - **Connections** mostra il grafico delle relazioni delle keyword delle facet pairs selezionate



IBM Enterprise Content Management

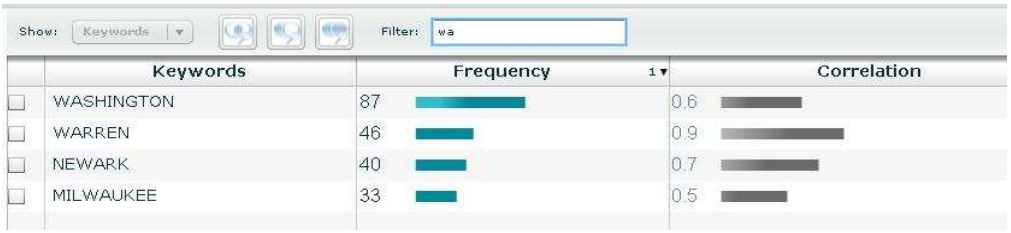
Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining: FACETs



- Fornisce una vista gerarchica dei risultati per parola chiave associata alle facet
- Mostra frequenza ed indici di correlazione tra facets
- Drill-down tramite l'aggiunta della parola chiave selezionata alla condizione di ricerca

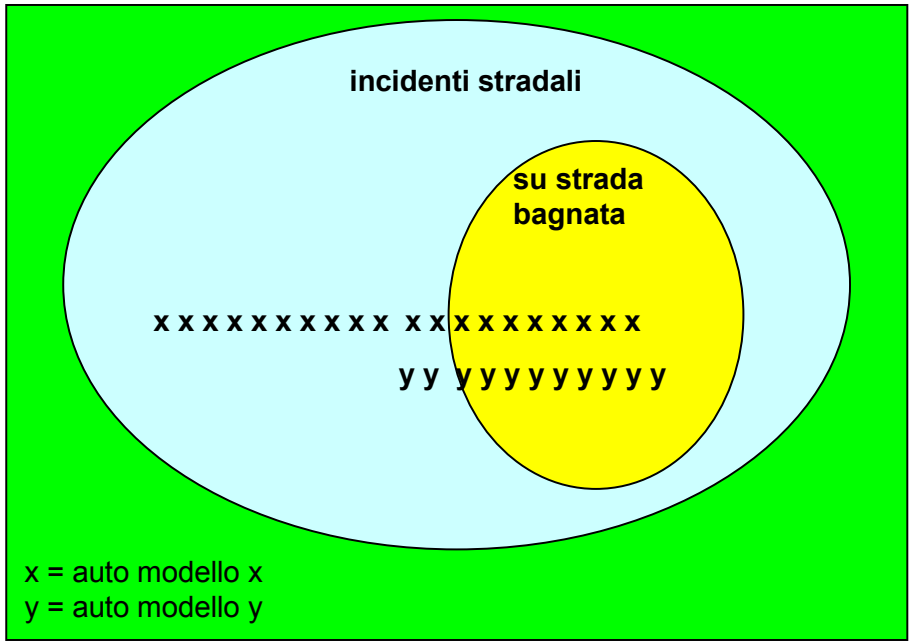
- Applicazione di filtri
- Ordinamento su più colonne



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
Funzionalità di Text Mining: ***Frequenza e Correlazione***



Frequenza auto x = 20
Frequenza auto y = 11

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

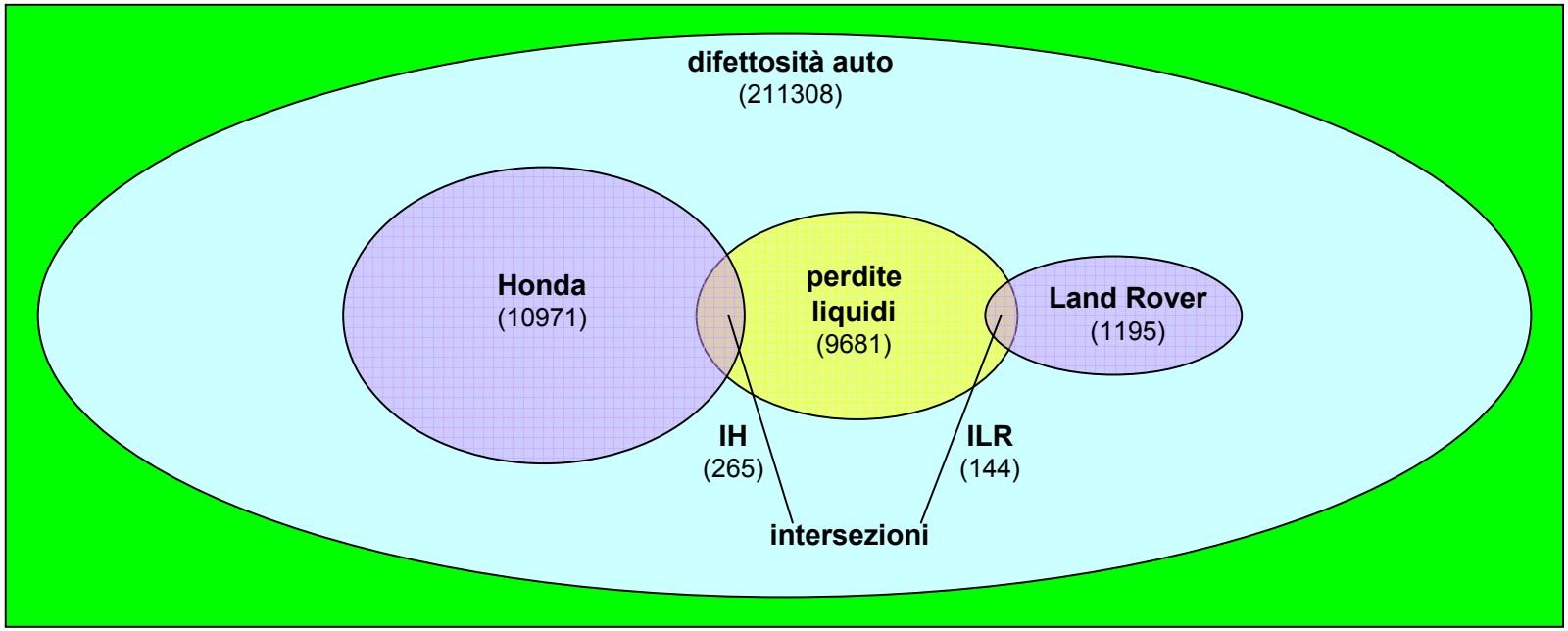
NEL CONTESTO
incidenti stradali su strada bagnata

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 o oggetti all'interno dell'insieme 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

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

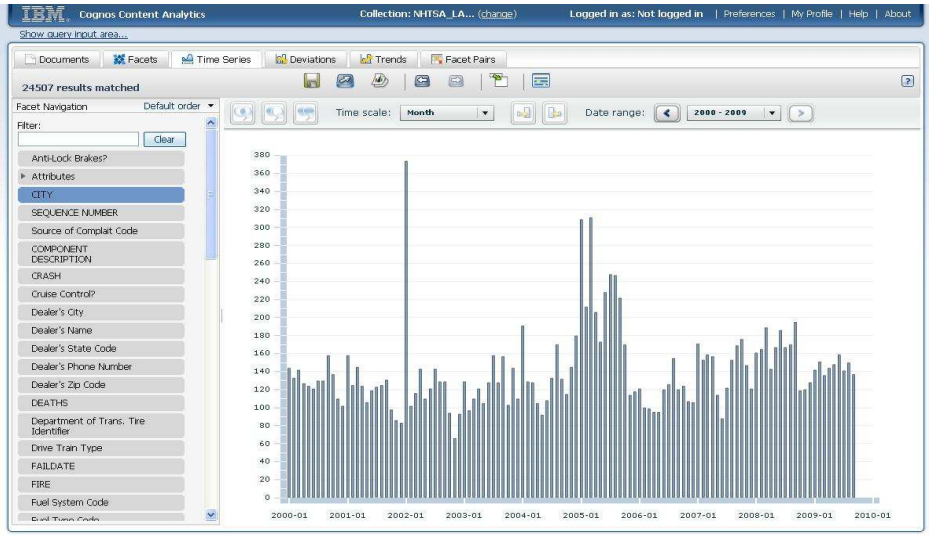
$$\text{correlazione Honda - perdite} = \frac{0,00125}{0,05190 * 0,04580} * 0,77 = 0,4$$

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

IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

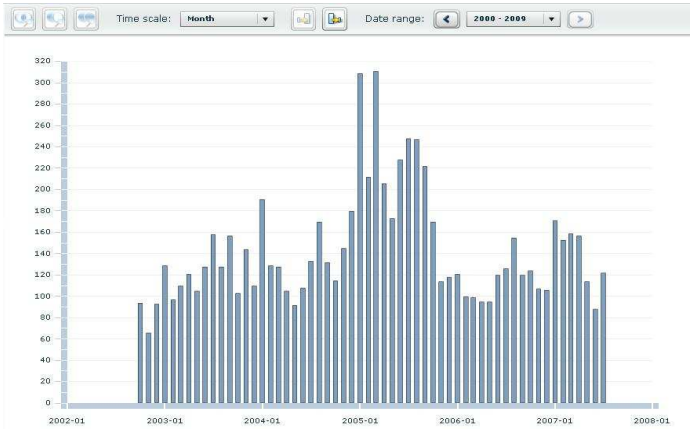
IBM Content Analytics
*Funzionalità di Text Mining: **Serie temporali***



- 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

Zoom in

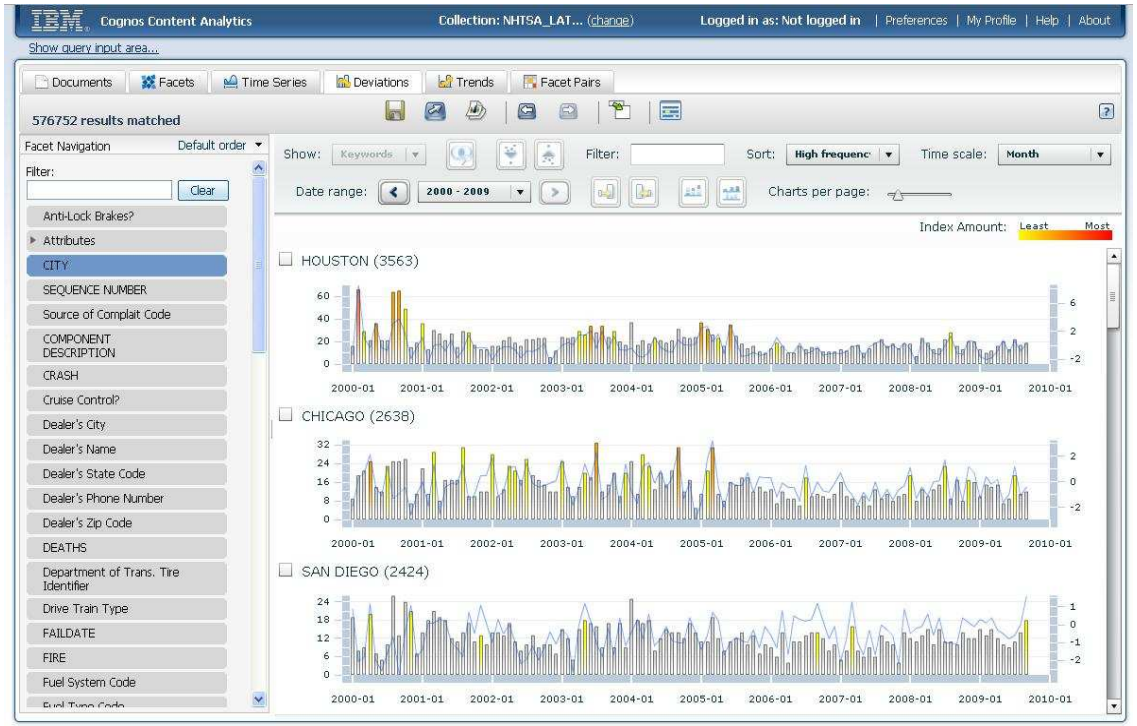
Zoom out



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 e gli score calcolati su barre del grafico separate
- Deviazioni
  - Mostra come la quantità di occorrenze di una parol chiave devia dalla media delle altre parole chiave
- Trend
  - Mostra come in ogni periodo di tempo la quantità d occorrenze di una parola chiave devia dalla media di tutti i documenti che soddisfano la riceca corrente

- Permette di raffinare la ricerca usando il valore della facet e la data
- Ordina, filtra, zoom-in, zoom-out

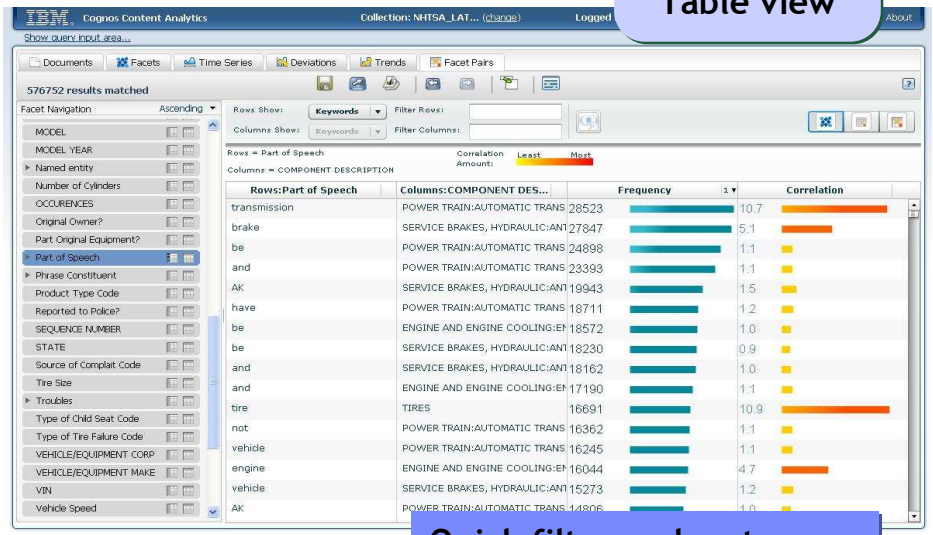
IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

IBM Content Analytics
*Funzionalità di Text Mining: **Facet Pairs***

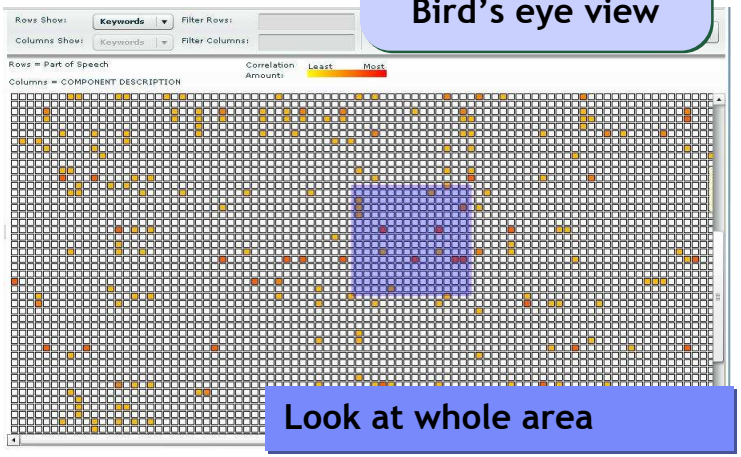
- Mostra la correlazione tra parole chiave appartenenti a due facet diverse
- 3 modalità di visualizzazione

Table view



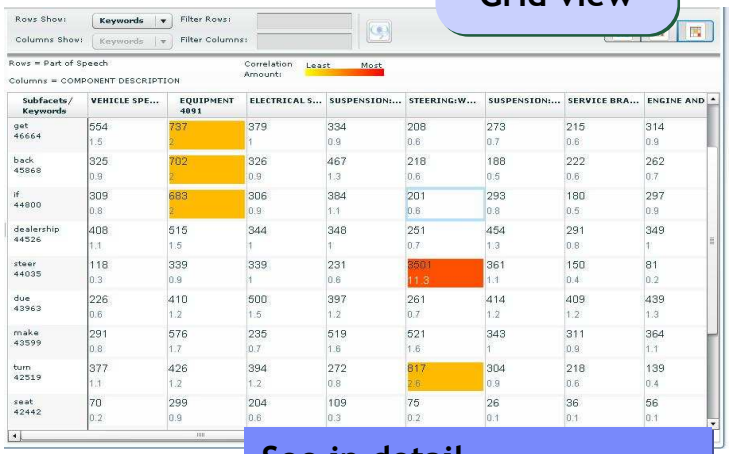
Quick filter and sort

Bird's eye view



Look at whole area

Grid view



See in detail

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 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

FileNet

documentum a division of EMC

Hummingbird

OPEN TEXT CORPORATION Livelink ENTERPRISE SUITE

IBM

CM & CMOD

Database

DB2

SYBASE

ORACLE

Informix

Microsoft

MySQL

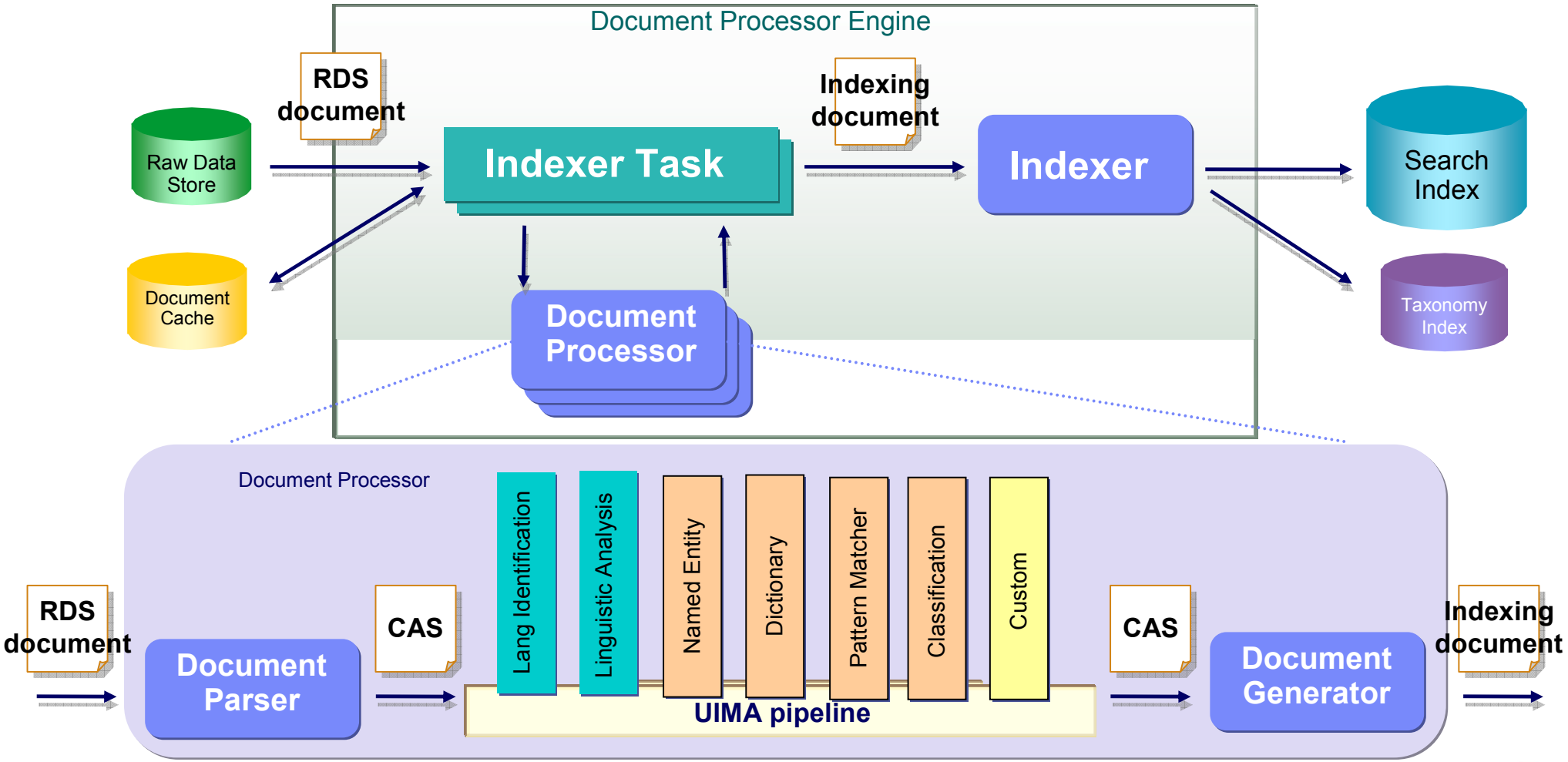
IBM Websphere MQ
with Event Publisher

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

IBM Enterprise Content Management
Contenuti al centro per decisioni più intelligenti.

Native security support

IBM Content Analytics
Struttura del Document Processor Engine



IBM Enterprise Content Management

Contenuti al centro per decisioni più intelligenti.

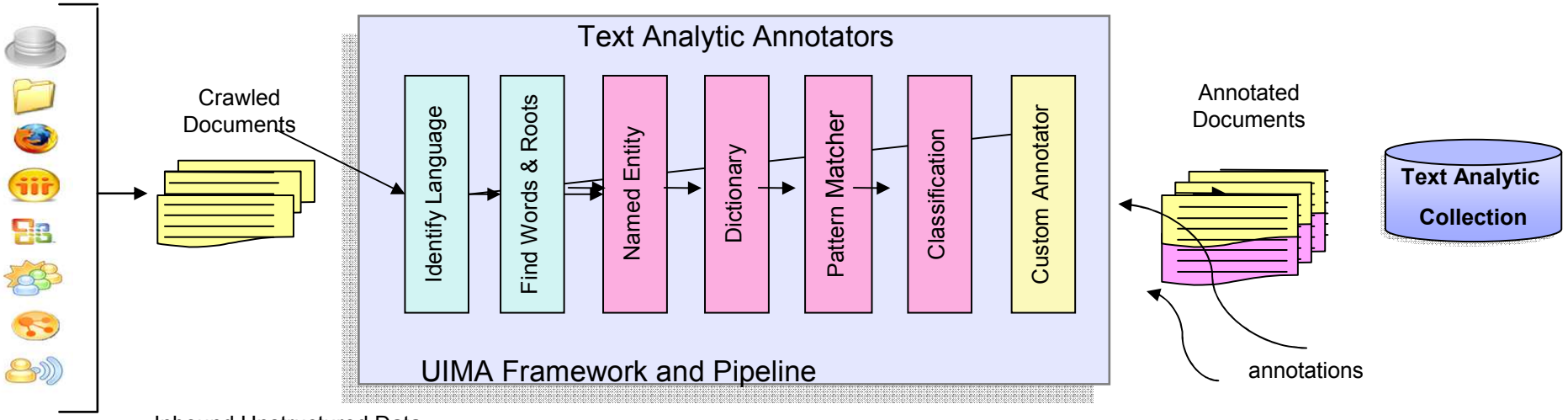
IBM Content Analytics
Unstructured Information Management Architecture

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



Inbound Unstructured Data

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.

Domande?