



# **Cómo analizar y buscar contenidos de una manera adecuada**

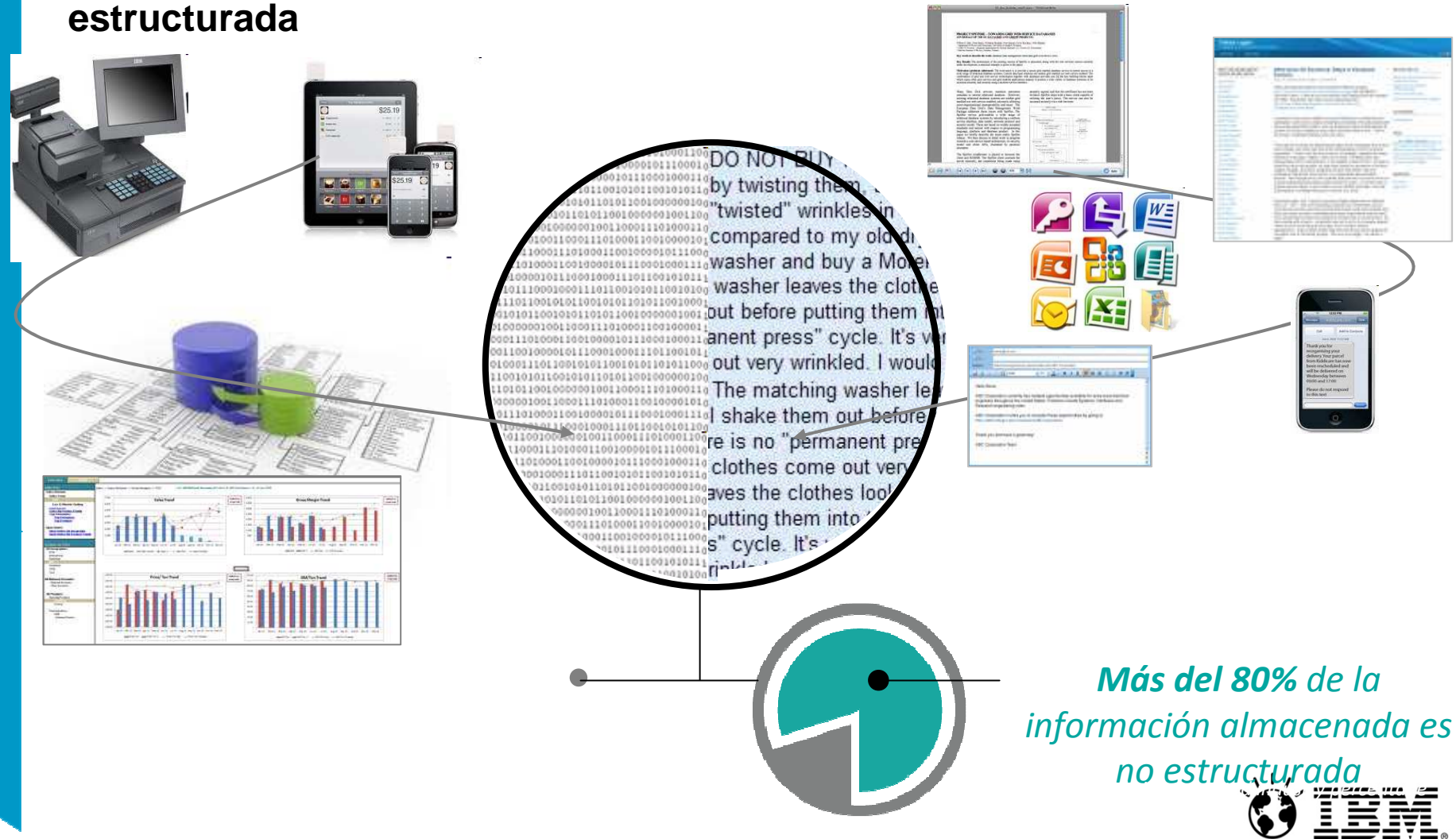
**Francisco Izquierdo**

**Especialista Técnico en Soluciones de Gestión Documental**

**#START013, 6 Noviembre 2012**

# El Imperativo del Análisis Textual

El concepto de Análisis de Información está cambiando desde datos transaccionales y estructurados hacia información interactiva y no estructurada



# Definiciones

## ¿Qué es el Análisis de Texto?

*Text Analytics* (NLP) describe un conjunto de técnicas estadísticas lingüísticas, y de aprendizaje automático que permiten que el texto a analizar y la extracción de información sea clave para la integración de negocios.

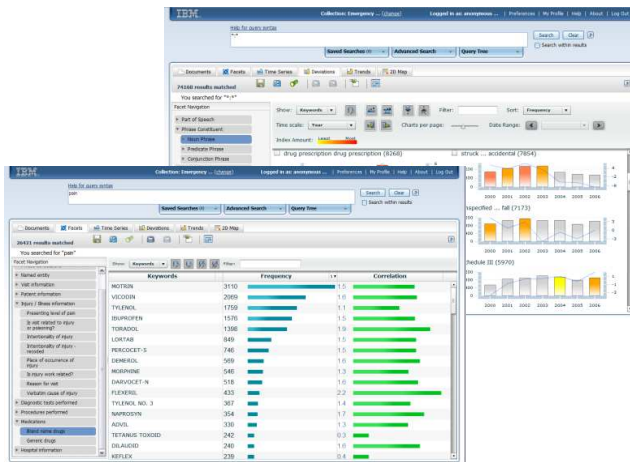
PC 143 (Hunter)  
 15 June 2006 23:47  
 Suspect identified himself as John Setsuko. Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving White Ford Mondeo, W563 WDL. Address given as 22 East Dene Ridge, Copdock, Ipswich. Searched at scene and found in possession of 1oz Cannabis Resin and lockable pocket knife.



Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Color	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession

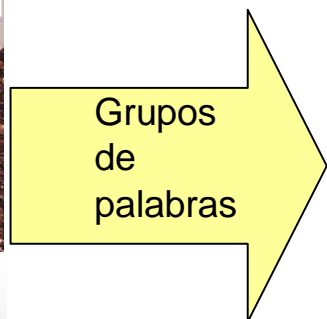
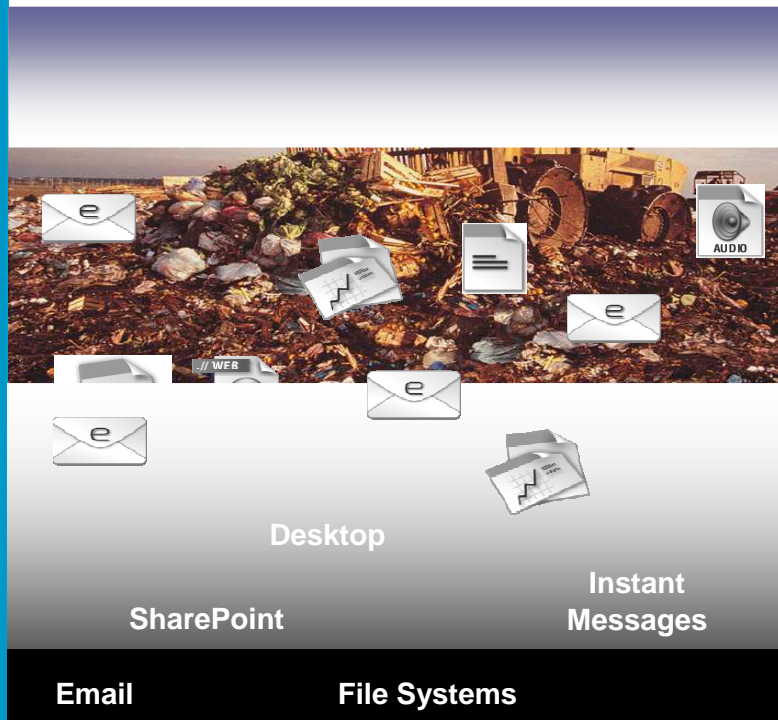
## Qué es Content Analytics?

*Content Analytics* (Text Analytics + Mining) se refiere al proceso de análisis de texto además de la capacidad para identificar visualmente y explorar las tendencias, patrones, y los hechos estadísticamente relevantes encontrados en distintos tipos de difusión de contenidos a través de fuentes de contenido internas y externas

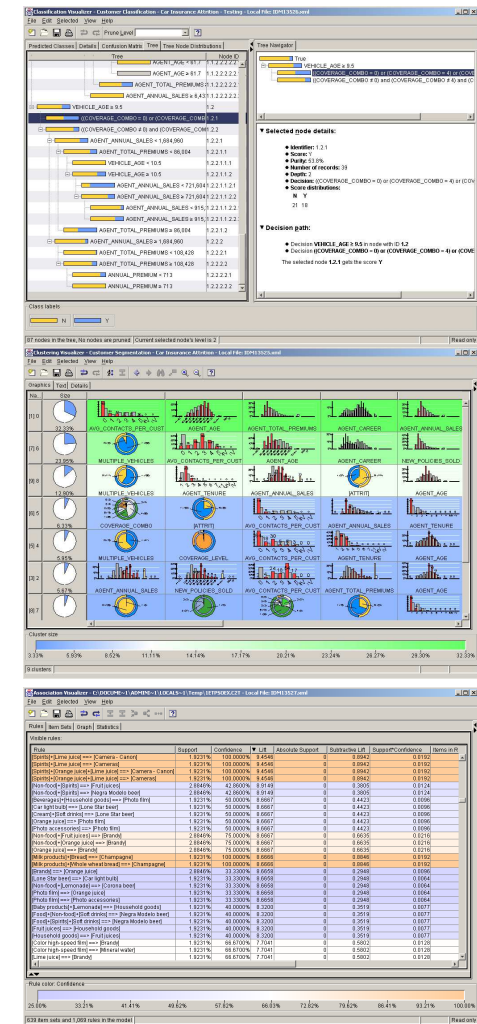


# ¿Qué es la Minería de Textos?

- La Minería de Textos es el proceso de extracción de **información utilizable a partir de datos textuales no estructurados**, mediante la identificación de conceptos, opiniones y tendencias.



01100101  
 11101011  
 01101110  
 00010100



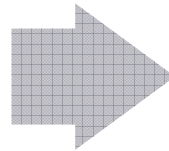
# Utilización para Análisis de Informes Policiales

## PC 143 (Hunter)

15 June 2006 23:47

Suspect **identified himself as** John Setsuko. **Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving** White Ford Mondeo, W563 WDL. Address **given as** 22 East Dene Ridge, Copdock, Ipswich.

**Searched at scene and found in possession of** 1oz Cannabis Resin **and** lockable pocket knife.



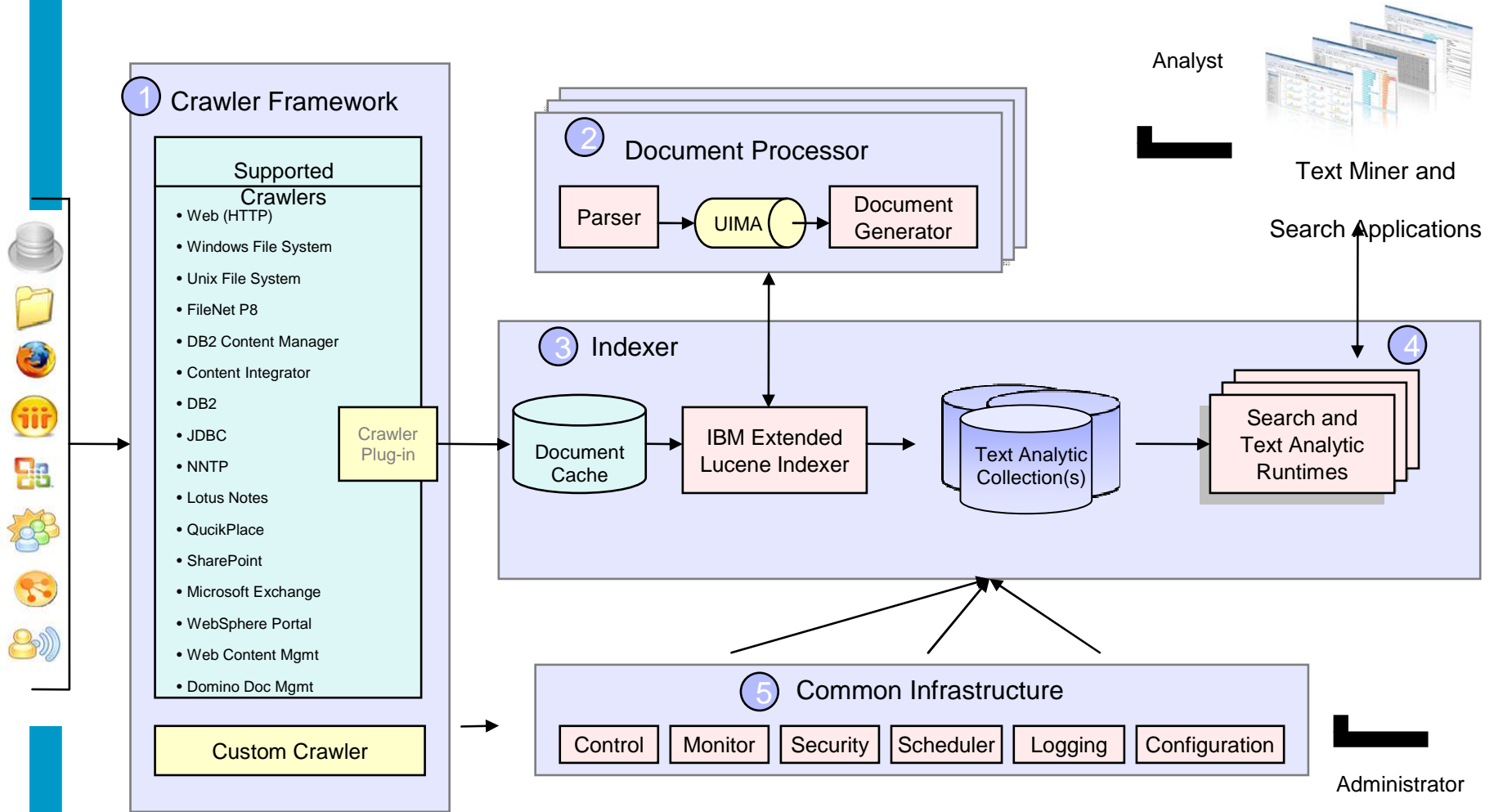
Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Colour	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession

Con este análisis ahora ya es posible::

- **Comprobar errores e inconsistencias en las bases de datos existentes**
- **Proporcionar información útil a los diferentes departamentos**
- **Mejorar considerablemente las capacidades de búsqueda documental**
- **Realizar tareas de resolución de identidades y relaciones**

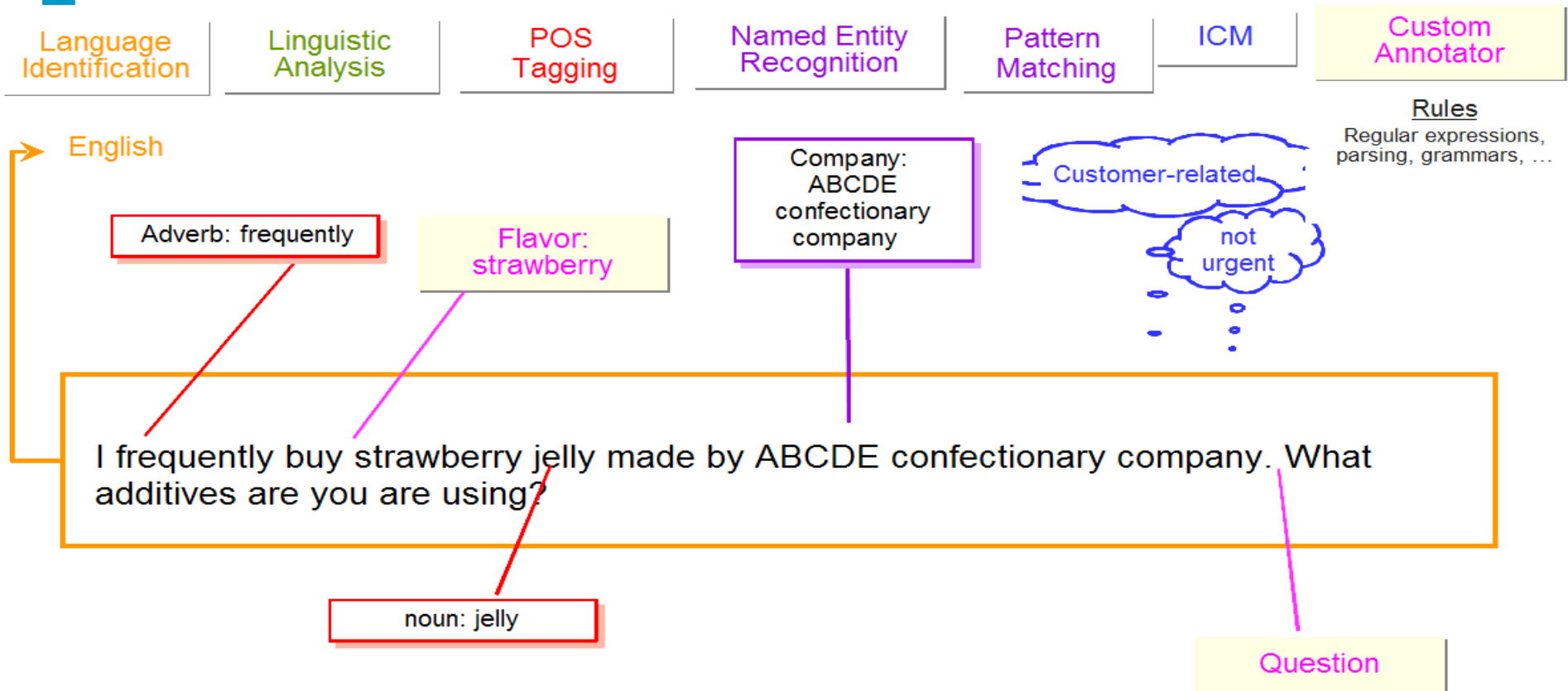


# Arquitectura de ICA



# ¿Qué son los Anotadores ICA?

**Anotador:** Se trata de un componente software que realiza tareas de análisis lingüístico, y como resultado obtiene y registra **anotaciones**





# ¿Y qué es lo que hacen?

Language Identification

Segmentation

Normalization

Classification

Disambiguation\*

Relationship Extraction\*

Rules

Regular expressions, parsing, grammars, ...

Fuzzy matching

Spelling correction, approx. lookup, hyphenation, ...

IBM Employee

I94989

Request for Proposal

IBM Product

end-of-business

solution

present

OmniFind Enterprise Edition

IBM Product

Marie,

I have an *OmniFind* RFP to complete for *Acme Corp.* by EOB Friday. We are presenting OEE, UIMA, *WebSphere Portal* and *LanguageWare* as the key components of the solution. To deliver this soln. we need LW support for creation of custom annotators, especially for *named entity recognition*. *DJ* and *Alex* mentioned that you were adding parsing to LW and were merging *James Luke's Swallow* with *LW workbench*. When will this be ready for customers? Also, could you provide *software services* as part of this RFP as the customer will need support in creating custom annotators?

Thomas ( [thomas.hampp@ie.ibm.com](mailto:thomas.hampp@ie.ibm.com) )

LanguageWare

Swallow and LanguageWare merging.

Acme Corp. request OmniFind RFP

Customer-related

Services Request

Urgent

English



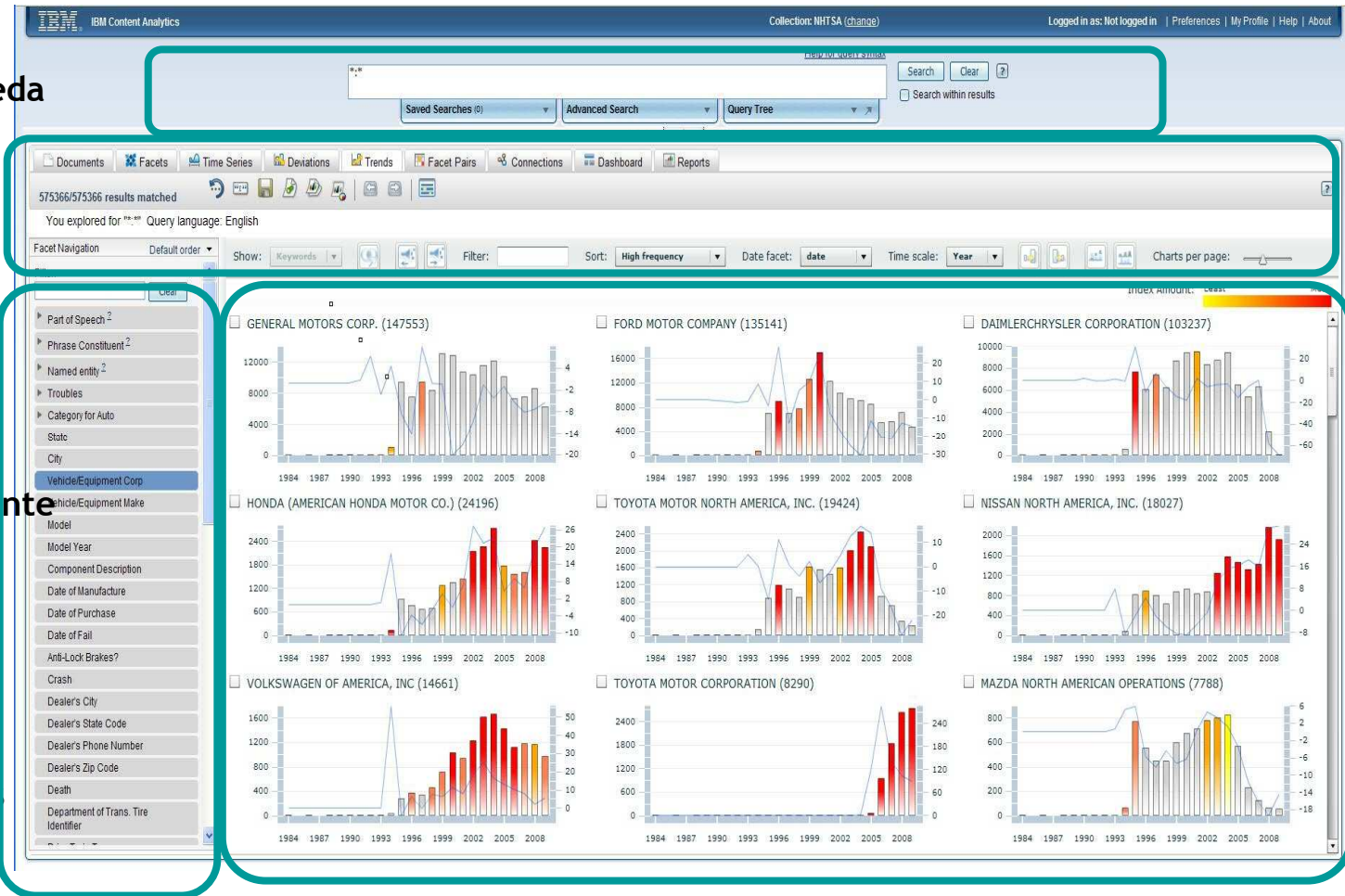


# Vista interactiva del Análisis de los Contenidos

Query, búsqueda

Puntos de vista, filtros y humbrales

Automáticamente se extraen y analizan los conceptos, las entidades, relaciones, metadatos y Clasificaciones



Visualización con capacidades de Drill Down





# Vista de Documentos

**IBM Cognos Content Analytics** Collection: FD... (change) Logged in as: Not ... | Preferences | My Profile | Help | About

Hide query input area... Help for query syntax

\*:\*

Search Clear ?

Search within results

Saved Searches (0) Advanced Search Query Tree

Documents Facets Time Series Deviations Trends Facet Pairs

Results 1-10 of 3000 (363562/363562 results matched) View by file Results per type: ALL - page: 10 - 1 2 3 4 5 6 7 8 9 10

You explored for "\*:\*" Query language: English

Facet Navigation	Default order	Source	Date	Title	Thumbnail
Filter:		Windows file system	3/6/08	<a href="#">MAUDE-1008531.xml</a>	
Part of Speech		1008531 967463 9612771-2008-00011 Manufacturer report 1 1 2008-03-06 Y N 2008-03-05 N Unknown Initial submission N representation 2008-02-22 Injury 977627 B 1 PLUS MODULAR STEM HIP STEM PLUS ORTHOPEDICS AG NA 75013441 0007. NA No JDI DA Device not returned to manufacturer UNK Hospitalization Required Intervention			
Phrase Constituent		Windows file system	1/30/08	<a href="#">MAUDE-989758.xml</a>	
Named entity		989758 948922 9616099-2008-00245 Manufacturer report 1 1 2008-01-30 N Y 2008-01-03 2008-01-03 Y Unknown Initial su N Foreign Health Professional Company representation 2008-01-03 Malfunction 991687 B 1 CYPHER SIROLIMUS-ELUTING COR STENT DRUG-ELUTING STENT (NIQ)CORDIS DE MEXICO NA CRB13275 13302064 PART#: NA Yes 2008-01-24 NIQ DA No 46 DURING AN INTERVENTIONAL PROCEDURE A CRACK WAS NOTED AT THE HUB OF THE DELIVERY SYSTEM BEFORE INFLATI INNER AND OUTER PACKAGING APPEARED NORMAL.THE PRODUCT HAS BEEN RECEIVED AND AN ANALYSIS IS PENDING. AC HAS BEEN REQUESTED AND WILL BE SUBMITTED WITHIN 30 DAYS OF RECEIPT. CYPHER SELECT PRODUCT IS NOT DISTRIB THE US; HOWEVER, IT IS SIMILAR TO US DISTRIBUTED CYPHER PRODUCT.NA Out-of-box failure Crack			
Report Information		Windows file system	1/31/08	<a href="#">MAUDE-989759.xml</a>	
Device Information		989759 925860 1823260-2008-01078 Manufacturer report 1 1 2008-01-31 N Y 2008-01-31 2007-12-03 N Unknown Initial su OTHER N Consumer 2008-01-16 Malfunction 991652 B 1 SOFTCLIX LANCET DEVICE LANCET DEVICE - FMK ROCHE DIAGNOS NI 11623656001 No FMK DA No 4655688 D CUSTOMER REPORTS, LANCET PROTRUDES PAST THE CAP OF THE SOFTCLIX D UNK IF BEFORE OR AFTER FIRING. ACCIDENTAL NEEDLE STICK OCCURRED, NO MEDICAL TREATMENT REQUIRED. NO ADVER EVENT REPORTED. REQUESTED RETURN OF SUSPECT DEVICE AND REPLACEMENT WAS SENT.UNK OINTMENT Retraction pr			
Manufacturer Information		Windows file system	3/6/08	<a href="#">MAUDE-1008532.xml</a>	
Patient Information		1008532 967464 9612771-2008-00012 Manufacturer report 1 1 2008-03-06 Y N 2008-03-06 2007-12-18 N Unknown Initial s N Company representation 2008-02-22 Injury 977624 V 1 PLUS FEMORAL COMPONENT PLUS ORTHOPEDICS AG NA 7500549 NA No JDL DA Device not returned to manufacturer UNK Hospitalization Required Intervention			
Hazards					

Search type: Keyword search

Facet Path:

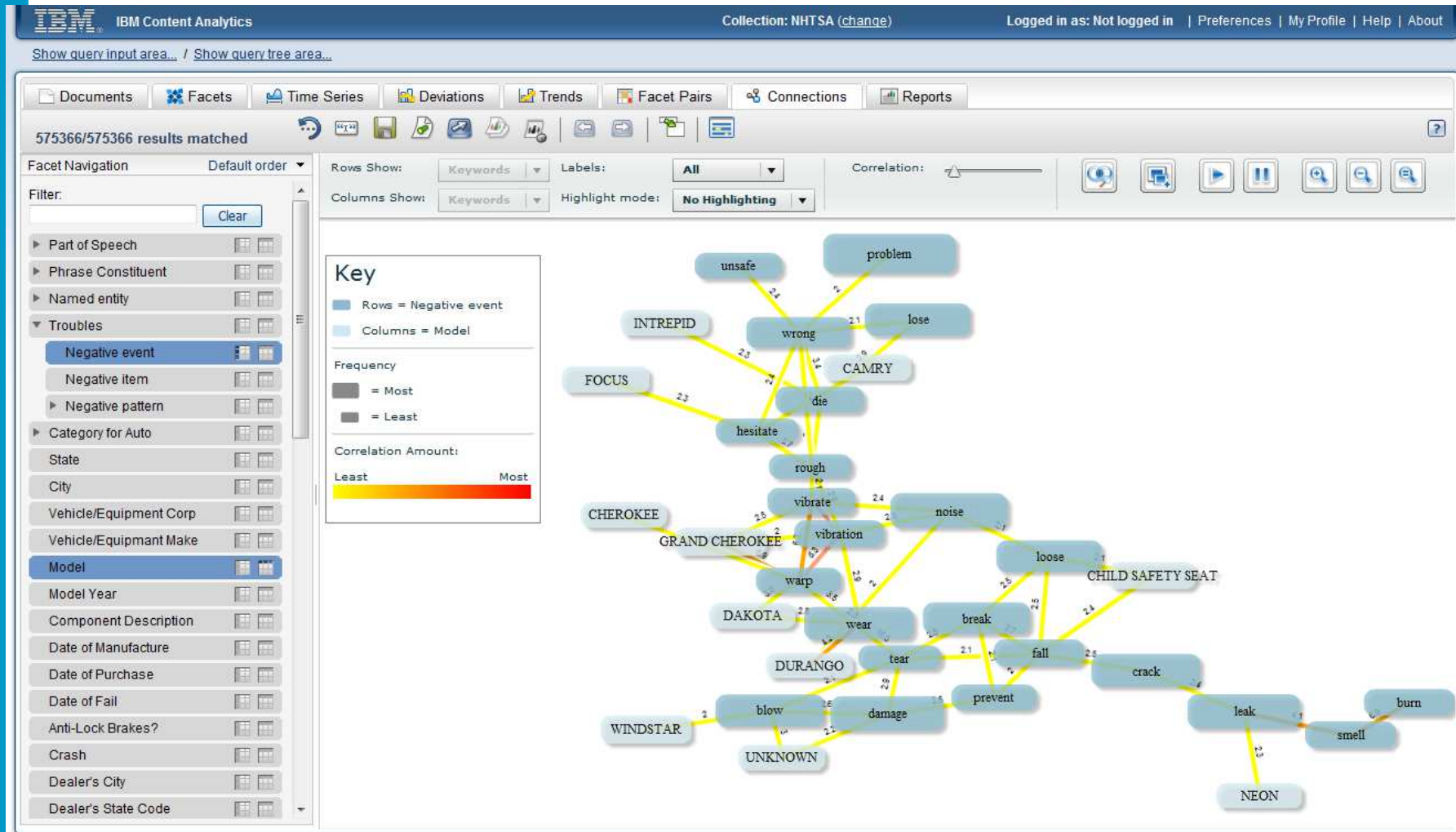
Keyword:

New search  Add to search

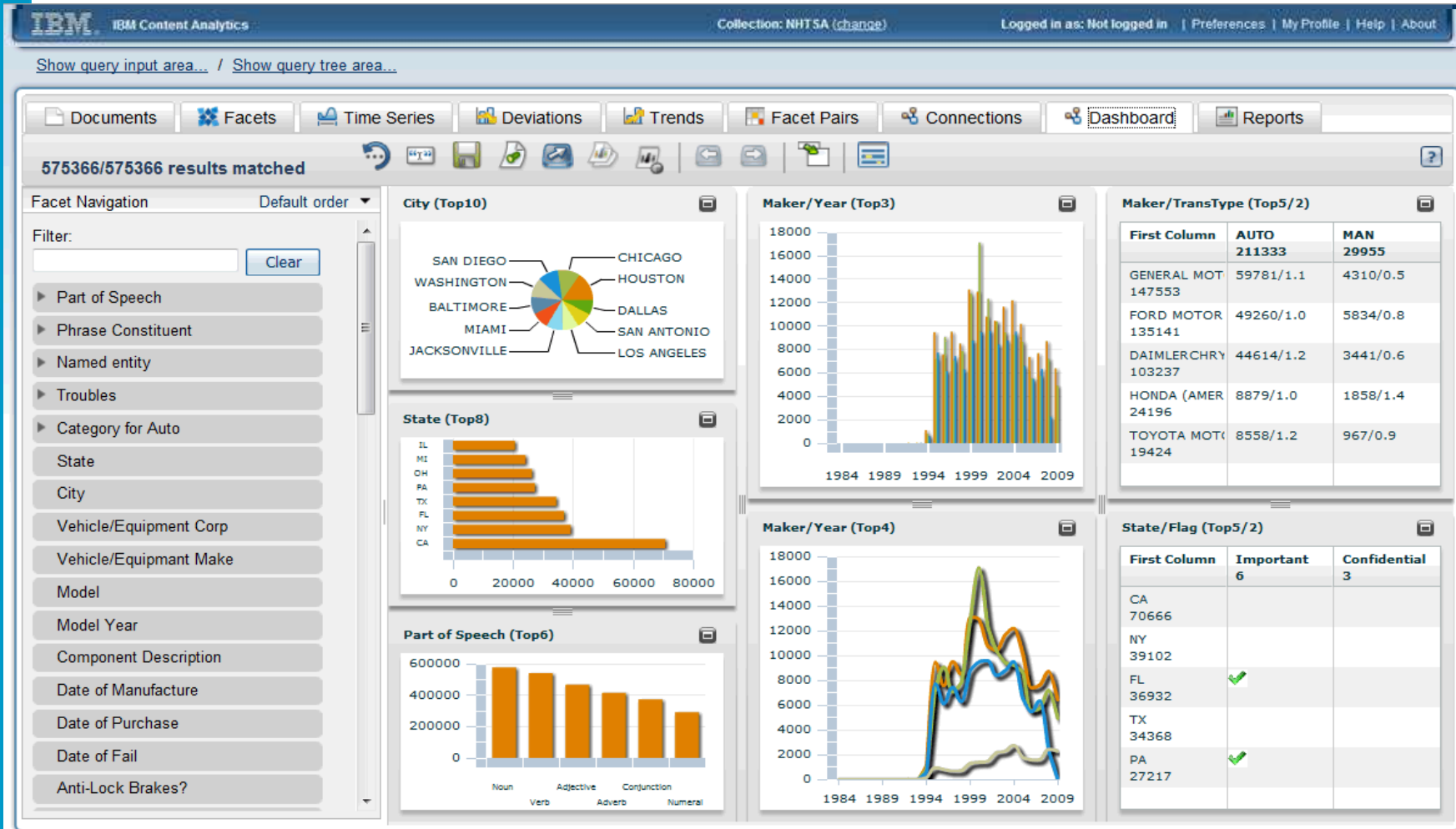
Search



# Conexiones: Vista con términos muy correlacionados entre sí

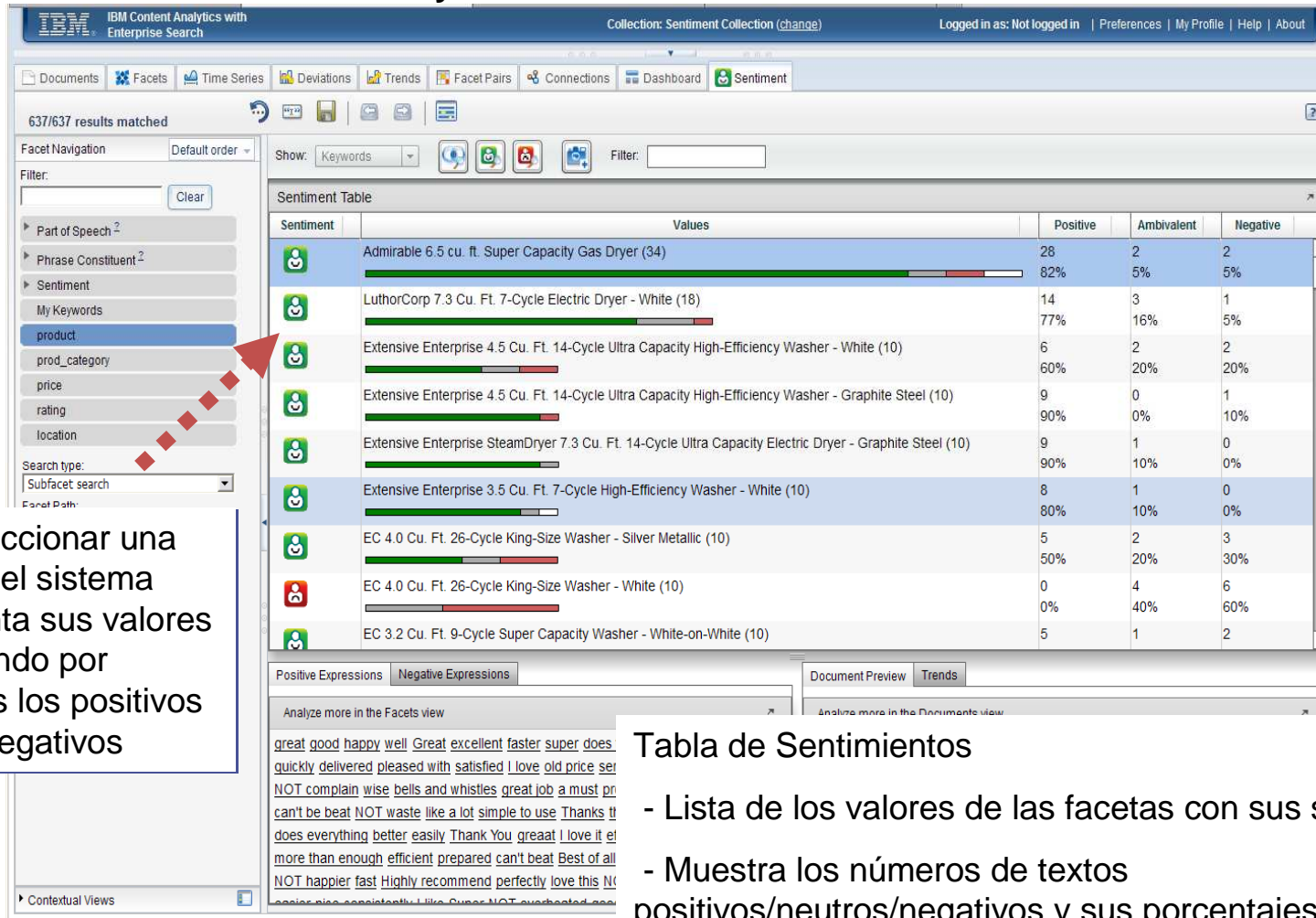


# Vista con paneles para Resúmenes Ejecutivos



# Análisis de Sentimientos

## Una vista nueva en Content Analytics



The screenshot displays the IBM Content Analytics interface. On the left, a 'Facet Navigation' panel shows a tree view with 'product.' selected. A red dashed arrow points from this selection to the 'Sentiment Table' on the right. The table lists various product models with their respective counts and sentiment percentages. Below the table, there are sections for 'Positive Expressions' and 'Negative Expressions' with sample text snippets.

Sentiment	Values	Positive	Ambivalent	Negative
Admirable 6.5 cu. ft. Super Capacity Gas Dryer (34)		28	2	2
LuthorCorp 7.3 Cu. Ft. 7-Cycle Electric Dryer - White (18)		14	3	1
Extensive Enterprise 4.5 Cu. Ft. 14-Cycle Ultra Capacity High-Efficiency Washer - White (10)		6	2	2
Extensive Enterprise 4.5 Cu. Ft. 14-Cycle Ultra Capacity High-Efficiency Washer - Graphite Steel (10)		9	0	1
Extensive Enterprise SteamDryer 7.3 Cu. Ft. 14-Cycle Ultra Capacity Electric Dryer - Graphite Steel (10)		9	1	0
Extensive Enterprise 3.5 Cu. Ft. 7-Cycle High-Efficiency Washer - White (10)		8	1	0
EC 4.0 Cu. Ft. 26-Cycle King-Size Washer - Silver Metallic (10)		5	2	3
EC 4.0 Cu. Ft. 26-Cycle King-Size Washer - White (10)		0	4	6
EC 3.2 Cu. Ft. 9-Cycle Super Capacity Washer - White-on-White (10)		5	1	2

**Al seleccionar una faceta el sistema presenta sus valores marcando por colores los positivos y los negativos**

**Tabla de Sentimientos**

- Lista de los valores de las facetas con sus sentimientos
- Muestra los números de textos positivos/neutros/negativos y sus porcentajes

# Pasos para desarrollar un anotador con ICA Studio

## 1 Desarrollar el anotador con ICA Studio

Construir un diccionario del dominio.

Desarrollar reglas para detectar los términos de importancia, las entidades y las relaciones

Crear un Anotador UIMA y probarlo sobre documentos de una colección

## 2 Exportar el anotador con ICA Studio en un PEAR file compatible con UIMA

Type	Offset	Context
@com.ibm.es.oze.lrw.sample.Product	4	The LanguageWare Resource Workbench
@com.ibm.es.oze.lrw.sample.Product	47	Resource Workbench What is LanguageWare
@com.ibm.es.oze.lrw.sample.Product	66	What is LanguageWare? IBM LanguageWare
@com.ibm.es.oze.lrw.sample.Product	351	t for more than 20 languages. LanguageWare is
@com.ibm.es.oze.lrw.sample.Product	666	tion has never been so great. LanguageWare is
@com.ibm.es.oze.lrw.sample.Product	853	ines and their resources. LanguageWare con
@com.ibm.es.oze.lrw.sample.Product	1378	rules and ontologies. The LanguageWare lib

## 3 Desplegar el Anotador en ICA

Importar el anotador UIMA desde la herramienta de administración de ICA y asociarlo a una colección.



# IBM Text Analytics Catalog

Nuevo

- Un catálogo con más de 300 anotadores para descargar de forma gratuita
- Amplio rango de dominios (Healthcare, Financiero, etc)
- Accesibles libremente desde la Web
- Arquitectura UIMA
- Compatible con ICA 2.2-3.0, LRW 7.2, Studio 3.0

Table of Contents

Category	Description
<a href="#">Academia</a>	Text analytics related to academics (universities, degrees, etc...)
<a href="#">Anatomy</a>	Text analytics for the human body
<a href="#">Animals</a>	Text analytics for the animal kingdom



# Global Insurance & Financial Service Organization

## Slashing risk exposure with analytics

### The need

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

### The solution

- 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

### The 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







# Financial Institution

## Exposing potential fraud with content analytics



### The need

A European financial Institution wanted to investigate fraudulent behavior by exploring internet sites for actions that might pose a threat to its members

### The solution

The company contracted with IBM to utilize IBM Content Analytics software to analyze a selected set of websites, investigate their findings and report findings back

### The benefits

Rapidly determine the types intrusion correlating bank terms with news about a known hacker using the out of the box extraction capabilities, prevention scenarios and frequently vulnerable operation systems.





# Government Intelligence Agency

## Intelligently Identifying Global Threats



### The need

- To improve organization's capability to understand terrorist elements and associated persons
- To quickly fuse information from disparate sources for agents and other staff
- To analyze and understand relationships between terrorist and associated activities

### The solution

IBM Content Analytics with Enterprise Search solution implemented to:

- Extract key facts and entities from unstructured information
- Create a social network hub that combines unstructured and unstructured content giving staff better visibility into terrorist activity and patterns

### The benefits

- Helps identify terror suspects and their associations & relationships
- Provides information to agents in hours, not days or weeks
- Uncovers previously unknown relationships between, suspects, persons of interest and activities





# Latin American Federal Police

## Improving Intelligence Operations



### The need

- To improve organization's capability to target and arrest criminals
- To fuse information from disparate sources to form a single view of a citizen
- To provide access to an overwhelming amount of content through a search interface

### The solution

- Extract facts, entities and concepts from unstructured sources like crime and investigation reports.
- Insights are surfaced through a search and correlation application for the purposes of Identity Resolution, single view of the citizen, and viewing of person records.
- 500M unstructured records across multiple systems, are collected, stored and analyzed
- Enables analytics-driven searches of both structured and unstructured information
- Implemented in a pattern Identity Insight and i2 that is repeated in other policing, intelligence, security and taxing agencies

### The benefits

- Help identify persons of interest and their relationships
- Help identify locations of past and planned criminal activity
- Increase human safety in a particularly difficult region





# A Healthcare Organization

## Proactive Patient Care



### The need

Medicare and Medicaid will begin charging penalties for what they see as excessive hospital readmissions. For many hospitals facing readmission rates for heart conditions as high as 25 percent, those Medicare and Medicaid readmissions alone will total more than USD1 million in fines. One hospital system in the United States realized that a key to reducing readmissions was to ensure that patients follow up on tests and treatments after discharge, staying healthier and reducing the likelihood of further adverse health events such as infection and relapse.

### The solution

- Hospital staff can now use the solution to analyze unstructured text for key discharge terminology, convert that text into structured data, and generate alerts and report for patients' primary care doctors and other caregivers
- Clearer data and better communication between health professionals helps ensure that patients keep their follow-up appointments and complete their post-discharge treatment
- Not only can patients stay healthier, but the hospital can also save millions of dollars on costly hospital readmissions.

### The benefits

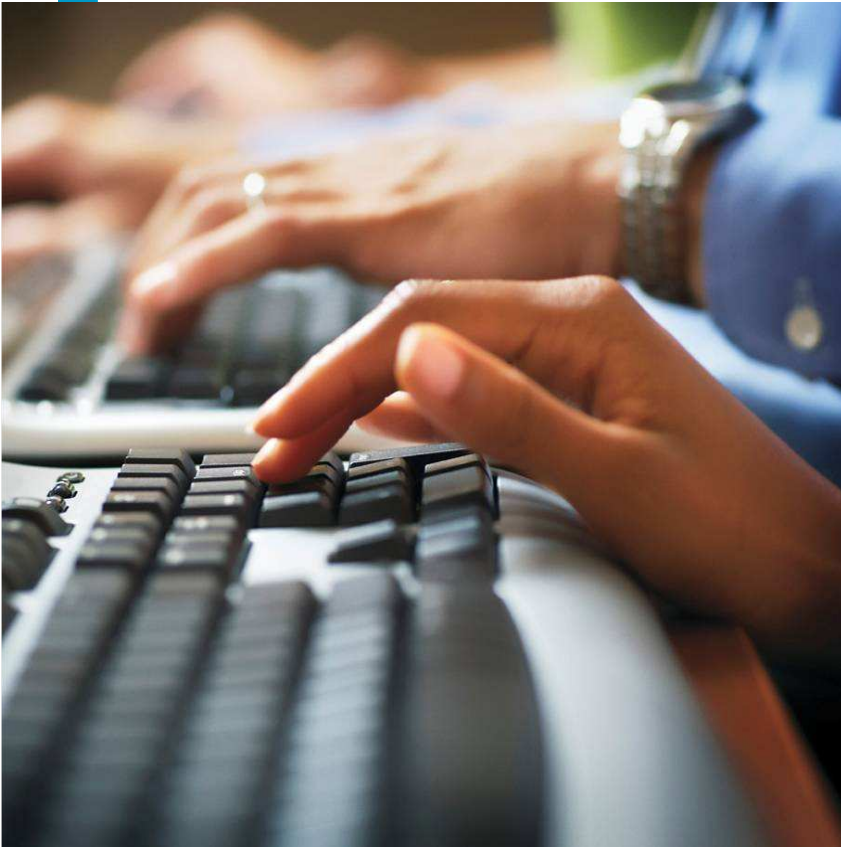
- Expected to prevent approximately USD1.1 million in Medicare and Medicaid penalty fines in areas of treatment with high readmission rates
- Expected to significantly reduce the overall number of hospital readmissions
- Expected to help improve recovery speeds by helping more patients follow up on treatments after discharge
- Expected to improve communications between staff, patients and follow-up caregivers by converting free text into searchable, reportable structured data





# A Legislative Document Company

Delivery fast, accurate updates to legislative material



## The Need

This company needed to improve the accuracy and speed of its regulation update service in order to stay ahead of new competitors and technology improvements. They needed to:

- Manage regulation data in a single platform
- Automate the legislation revision process to realize more accurate and faster revision and shorten the lead times for product delivery to local governments
- Quickly respond to customer needs which change dramatically over time

## The Solution

The company transformed its regulation management and updating system with natural language analysis technology. The new system:

- Automatically recognizes modifications to legislation and systematically updates the content
- Replaces a time-consuming, error-prone human process

## The Benefit

The client's new regulation and legislation management and update system brings a leading-edge system to the market by:

- Automating an error prone manual system
- Removing manual reading and understanding of new or updated legislation
- Improving legislation update speed by 50%
- Reducing search time for legislative information





# Consumer Products Research

Cross enterprise, secure search



*“It’s critical to us to be able to provide secure content search across the wide variety of data sources within our enterprise to enable our knowledge workers”*

— Customer

## The need

US consumer and industrial products company has a wide variety of data sources and internal environments that they need their users in different roles to be able to find information quickly and securely

## The solution

In 4 months, this research organization fully implemented enterprise search and classification, from IBM, to enable secure semantic search and classification of research assets for their internal portal that services all Institutes, as well as their public facing website

## The benefits

Using more than just keyword search, this organization was able to deploy content classification and contextual search to allow users to find accurate results based on the concepts they are looking for not just specific query terms alone





# ¡GRACIAS !



**Para Contactarnos:**

**Francisco Izquierdo**

**Entreprise Content Manager Especialista Técnico**

**[fizquierdo@es.ibm.com](mailto:fizquierdo@es.ibm.com)**

**Móvil: 670.62.64.98**

