

# IBM Content Analytics Proof Of Technology

11. Mai 2011, Wien



**Why IBM Content Analytics:**

Analyze dynamically, then decommission unnecessary content and preserve and exploit what matters.

# Disclaimer

©Copyright IBM Corporation 2011. All rights reserved. U.S. Government Users Restricted Rights -Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

## **THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY.**

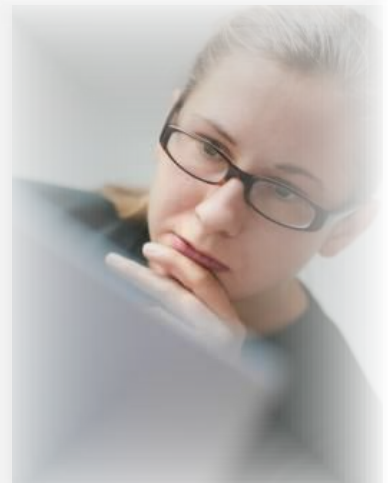
WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS AND/OR SOFTWARE.

IBM, the IBM logo, ibm.com, Cognos, FileNet, OmniFind and all IBM FileNet products are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)

Other company, product, or service names may be trademarks or service marks of others.

## Wissen ist Macht – Was wissen wir von unserem Unternehmen?

- Wissen Sie, wie viele Dokumente es in Ihrem Unternehmen gibt ?
- Wissen Sie, was Ihre Mitarbeiter „unstrukturiert dokumentieren“ ?
- Wissen Sie, wie viele Dokumente es über Ihr Unternehmen im Web gibt ?
- Wissen Sie, wie Ihre Kunden Ihr Unternehmen sehen ?
- Wissen Sie, was Ihre Kunden über Ihr Unternehmen im Web schreiben ?
- Sind Sie 100% sicher, dass Ihnen keine Information entgeht ?
- Sind Sie 100% sicher, dass Sie Ihnen neue Erkenntnisse nicht Helfen würden ?

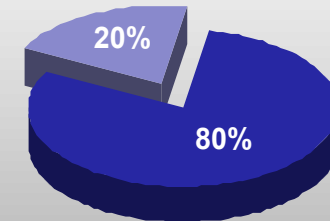


## Trendthema unstrukturierte Daten

- **247 Milliarden** Emails werden pro Tag versandt (Radicati Group)
- **47 Millionen** neue Web Seiten wurden 2009 im Internet erstellt (Netcraft)
- **25 Milliarden** Beiträge werden jeden Monat bei Facebook ausgetauscht (Facebook)
- **mind. 6 Millionen \$** verliert ein Unternehmen mit 1000 „Wissensarbeitern“ im Jahr durch die erfolglose Suche nach Informationen (IDC)
- **Business Intelligence** Initiativen beziehen sich idR auf strukturierte Daten...
- **...und lassen 80%** der verfügbaren aber unstrukturierten Informationen unberücksichtigt

### Unternehmensdaten

- Unstrukturierte Daten
- Strukturierte Daten



### Wachstum



- Unstrukturierte Inhalte nehmen jährlich um 65-200% zu



Strukturierte Daten enthalten meist: *wer, was, wieviel* und *wann*



Unstrukturierte Daten/ Content enthalten das **Warum** und **Wie**

# WATSON verwendet Text Mining – Was bedeutet das ?

“Der Kunde ist nicht zufrieden mit Mobil Telefon, Anruf vom 18.11.2010 – Kunde möchte zu Yellow inc wechseln”



## Wort Extraktion

Kd  
Kunde  
Yellow  
inc  
zufrieden  
nicht  
wechseln  
Mobil  
Telefon

## Begriffs- Extraktion

Kunde  
Yellow inc  
Mobil Telefon  
Nicht zufrieden

## Named Entities Recognition (NER)

**Kunde** -> CRM-Begriff  
Kd?  
**Yellow inc** -> Telco Company (nicht die Farbe)  
**Mobil Telefon** -> Telco-Begriff  
nicht zufrieden  
**18. Nov** > Date



70'er

80'er

90'er

## WATSON verwendet Text Mining – Was bedeutet das ?

“Der Kunde ist nicht zufrieden mit Mobil Telefon, Anruf vom 18.11.2010 – Kunde möchte zu Yellow inc wechseln”



### Sentiment-Analyse

Kunde (Kd) → Mobil Telefon →  
unzufrieden (negativ); Wechseln zu  
(negative Vorhersage)  
→ yellow inc (Wettbewerber)

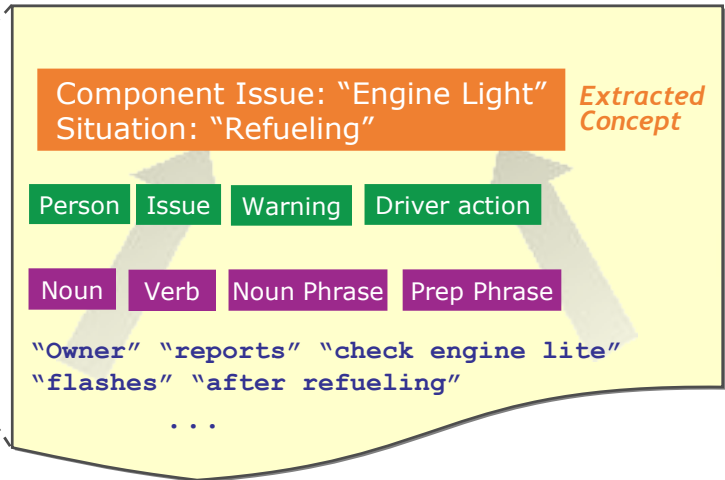
### Kombiniert mit strukturierten Daten

Schnelle Entscheidung  
Wechselwilliger Kunde  
hoher Kundenwert ?  
  
→ individuelles  
Angebot



Heute

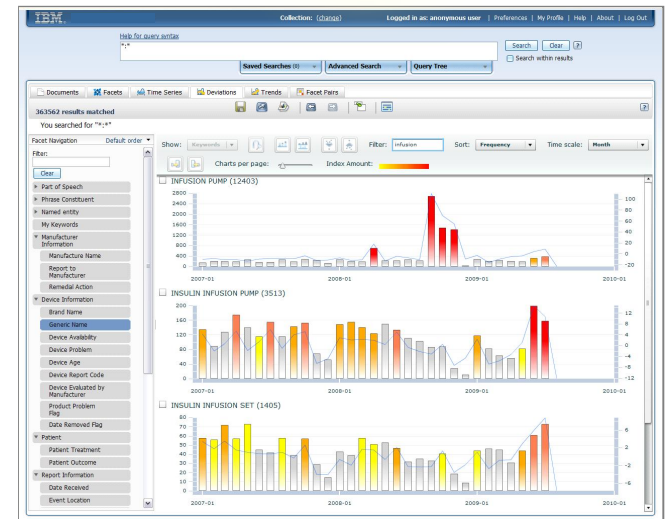
# IBM Content Analytics (ICA)



Analyse von  
 Content und Daten



Das Werkzeug zur dynamischen Analyse



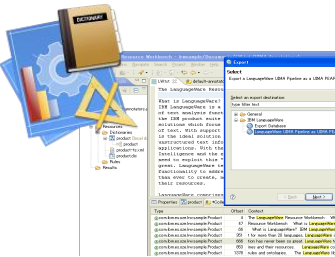
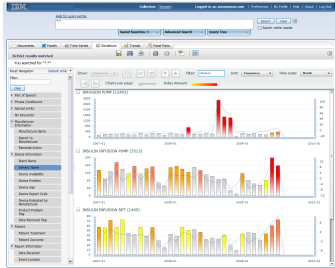
## Merkmale von IBM Content Analytics

- Einfache Bedienung
- Out-of-the-box Erkenntnisse ohne Fachkenntnisse
- Integrierte Volltextsuche zur Kombination mit Analyse
  - für Ad-hoc Auswertungen ohne Neukonfiguration des Systems
  - Suchbegriff wird als Dokumentenfilter verwendet
- Dynamische Analyse für Entscheidungsprozesse
  - Untersuchen / Analyse von unstrukturiertem Content und strukturierten Daten
  - Automatische Erkennung von Trends und Korrelationen
- UIMA Analyseframework – OpenSource Standard zur Datenanalyse
- LWR – IBM Language Ware – direkte Integration
- Sprachunterstützung für die gängigen Sprachen
- Analyseergebnisse können für andere Anwendungen und Prozesse bereitgestellt werden (Exportfunktion für DB und Warehouse)

# Eine zuverlässige Content Analyse Umgebung erlaubt ...

## Sofortiges Ausnutzen von Fähigkeiten im Standard

- Unterstützt über 30 Datenquellen und über 150 Dateiformate für die Analyse.
- Beinhaltet wertvolle Annotatoren, um automatisch Konzepte und Entitäten - ohne zusätzliche Anpassungen - zu erkennen.
- Sechs verschiedene benutzerfreundliche grafische Ansichten, um intuitiv neue Einblicke zu bekommen.
- Dynamisches Hervorheben von interessanten Auffälligkeiten und Abweichungen.
- Offene und standardisierte UIMA-Textanalyse für höchste Flexibilität und Erweiterung.
- Einfaches flexibles Werkzeug um Annotatoren, Regeln und Wörterbücher anzupassen.
- Erweitert das Content Management von Filenet P8 um die Analyse.
- Analyse bei der Vorgangsbearbeitung für Advanced Case Management
- Erweitert IBM Cognos 8 BI und seine Reports und Dashboards
- Integrierbar in jede Anwendung – von PC bis Mainframe – mit WebServices oder Java APIs.



**Business Analytics**



**Industry Solutions**



**ECM**

## Woher kommt die „Intelligenz“ bei der Analyse?

### Ein neuer OASIS-Standard für die Verarbeitung und Analyse von Inhalten

- In Deutschland setzt z.B. die **Fraunhofer Gesellschaft** auf UIMA und entwickelt Annotatoren
- UIMA definiert eine einheitliche Schnittstelle zur Integration von Analyseschritten
  - Ermöglicht Interoperabilität verschiedener Analyselösungen und Unternehmensanwendungen
  - DARPA (Forschungseinrichtung des amerikanischen Verteidigungsministeriums) benutzt UIMA
- UIMA Working Group: IBM und DARPA sind Co-Sponsoren
  - Gegründet im Januar 2005 zur Weiterentwicklung von UIMA
  - Wichtige Vertreter aus Wissenschaft und akademischer Forschung
  - Teilnahme von Partnern, Forschungsabteilungen von Unternehmen



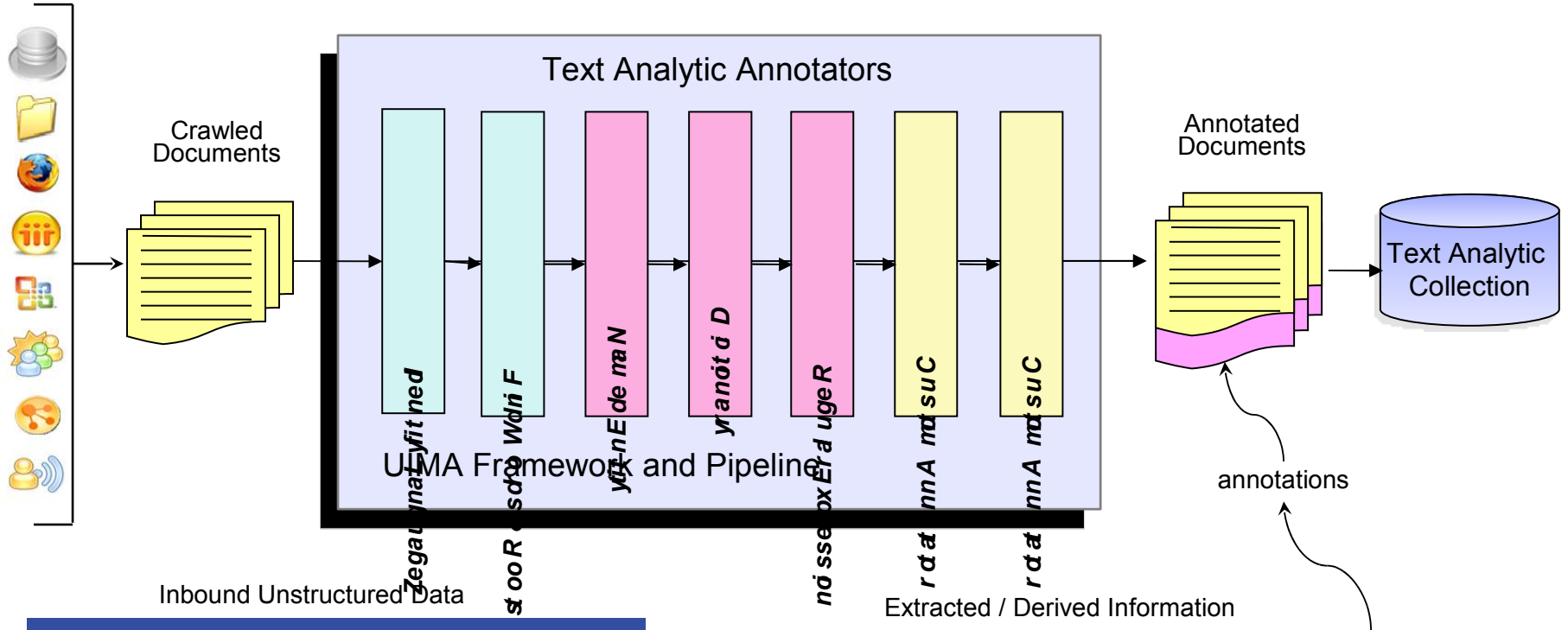
Unstructured Information Management Architecture

*An Apache Incubator Project.*



Federführung der UIMA Entwicklung im IBM Labor Böblingen (bei Stuttgart)

# Unstructured Information Management Architecture

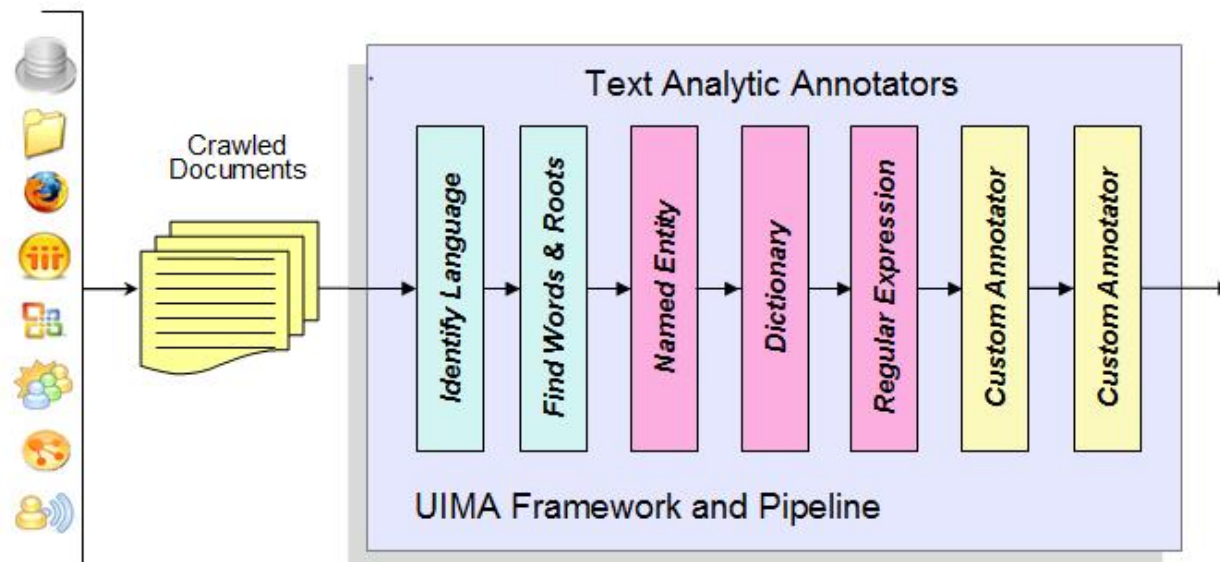
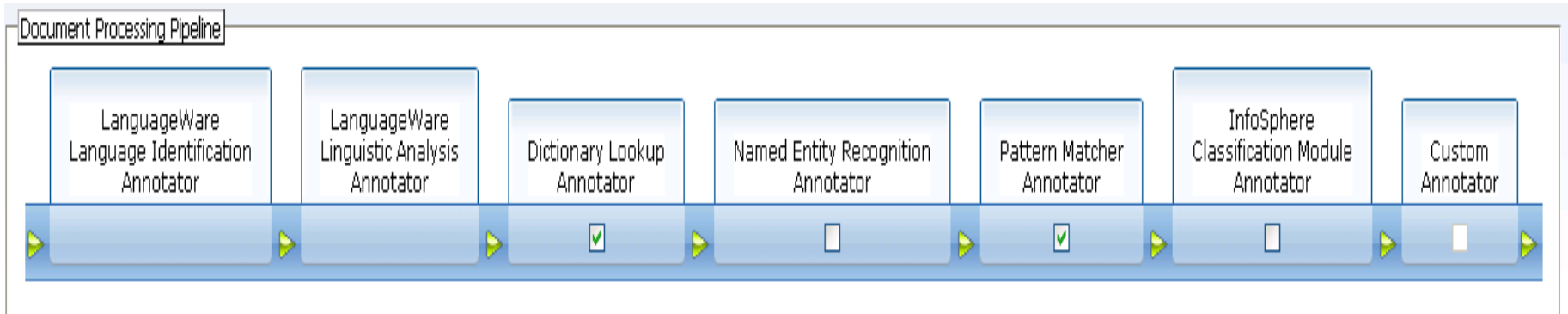


Device Malfunction Description:  
 It was reported that during a **gastric bypass** roux-en-y procedure, on the 3<sup>rd</sup> firing with a blue load on the **stomach** there was **bleeding**. Not sure if staples were formed properly. They over sewed the staple line. There was no PT consequence.



Involved Body Part	<b>stomach</b>
Type of Injury	<b>bleeding</b>
Procedure Performed	<b>gastric bypass surgery</b>

# Umsetzung der UIMA Pipe im ICA



## Analysefunktionen

Dokumente

24507 results matched

Keywords	Frequency	Correlation
HOUSTON	196	1.0
DALLAS	139	1.4
CHICAGO	119	0.8
MIAMI	99	0.8
BALTIMORE	91	0.7
JACKSONVILLE		
WASHINGTON		

Abweichungen

576752 results matched

Date range: 2000 - 2009

Charts per page: 1

Index Amount: Least Most

HOUSTON (3563)

CHICAGO (2638)

SAN DIEGO (2424)

Facetten Paare

576752 results matched

Rows:Part of Speech	Columns:COMPONENT DES...	Frequency	Correlation
transmission	POWER TRAIN:AUTOMATIC TRANS	28523	10.7
brake	SERVICE BRAKES, HYDRAULIC:ANT	27847	5.1
be	POWER TRAIN:AUTOMATIC TRANS	24898	1.1
and	POWER TRAIN:AUTOMATIC TRANS	23393	1.1
AK	SERVICE BRAKES, HYDRAULIC:ANT	19943	1.5
have	POWER TRAIN:AUTOMATIC TRANS	18711	1.2
be	ENGINE AND ENGINE COOLING:EH	18572	1.0
be	SERVICE BRAKES, HYDRAULIC:ANT	18230	0.9
and	SERVICE BRAKES, HYDRAULIC:ANT	18162	1.0
and	ENGINE AND ENGINE COOLING:EH	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:EH	16044	4.7
vehicle	SERVICE BRAKES, HYDRAULIC:ANT	15273	1.2
AK	POWER TRAIN:AUTOMATIC TRANS	14806	1.0

Trends

576752 results matched

Date range: 2000 - 2009

Charts per page: 1

Index Amount: Least Most

be (363699)

and (331219)

Time Series

24507 results matched

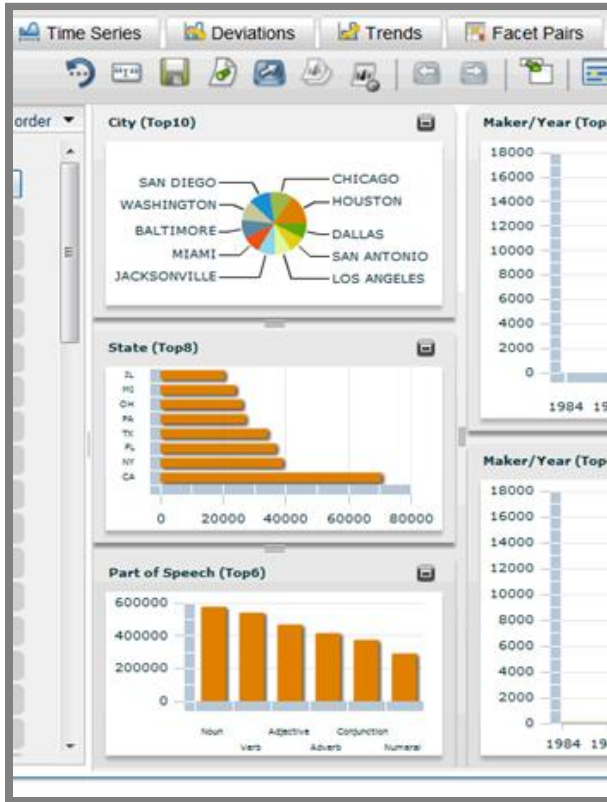
Date range: 2000 - 2009

Charts per page: 1

Time scale: Month

Filter: CITY

# What's new in IBM Content Analytics v2.2?



- **New Visualizations in Content Analytics Text Miner**
  - Connections View links highly correlated terms to one another
  - Dashboard view to see 1 or more analytics views in a single window.
  - Query Builder to easily create and save queries.
  - Ability to add custom views
- **Easier integration with Cognos BI reports and models**
  - Quick Cognos® BI report generation
  - Tighter integration with Cognos data models
  - Cognos reports can link from and back to Content Analytics
- **Speed Time to Value: Enhanced analytics configuration tools**
  - Tighter integration with LanguageWare® Resource Workbench (LRW)
  - Parametric dates and numerical range support in Facet Tree Editor
  - Support to auto-detect and add-on new languages
- **Document Analysis Support**
  - Mapping file metadata to auto-generate Facets
  - Documents flagging support
  - Near duplicated document detection
  - Support for Linux® (Redhat) on IBM System z® for file system, databases and web pages
  - Enhanced import/export document analysis to CSV, RDB, etc.

---

## Weitere Unterlagen

- IBM Redbook: <http://www.redbooks.ibm.com/abstracts/sg247877.html?Open>
- Online ICA Doku: <http://publib.boulder.ibm.com/infocenter/analytic/v2r2m0/index.jsp>
- YouTube Video: <http://www.youtube.com/user/IBMECM#p/u/4/cM-sYcYlhP4>
- ICA Web Seite: <http://www-01.ibm.com/software/data/content-management/analytics/>

