Fuentes de datos

Introducción al análisis y minería de textos

Tipos de datos

- Estructurados
  - Números, tablas, renglones, atributos
  - Predecibles, bien organizados
  - Tipicamente se derivan de un sistema transaccional – Estructura repetible
- No estructurados
  - eMail, reportes, documentos, registros médicos, contratos
  - No tienen estructura, no son repetibles, no son predecibles
  - Frecuentemente derivados de actividades cotidianas de comunicación. *Por ello, la mayoría son texto*

Datos no estructurados en una empresa

Unidad	Fuente de datos
Contabilidad	Hojas de cálculo, notas, word, auditoría, conversaciones, respuestas
Call center	Conversaciones, notas, respuestas
Ingeniería	Inventarios, Diseños, Arcivos de producción, Especificaciones
Finanzas	Hojas de cálculo, Reportes anuales, Notas
Recursos humanos	eMail, cartas, ofertas de trabajo
Legal	Contratos, transcripciones telefónicas, propuesas patentes, marcas registradas, NDA
Marketing	Hojas de cálculo, cuentas, pronósticos, conferencias, webinars, notas de contacto con clientes, información de canales
Operaciones	Corridas de manufactura, productos con defectos, inventarios, notas de entrega y despacho
Ventas	Estadísticas, llamadas, pronósticos, evaluaciones de desempeño

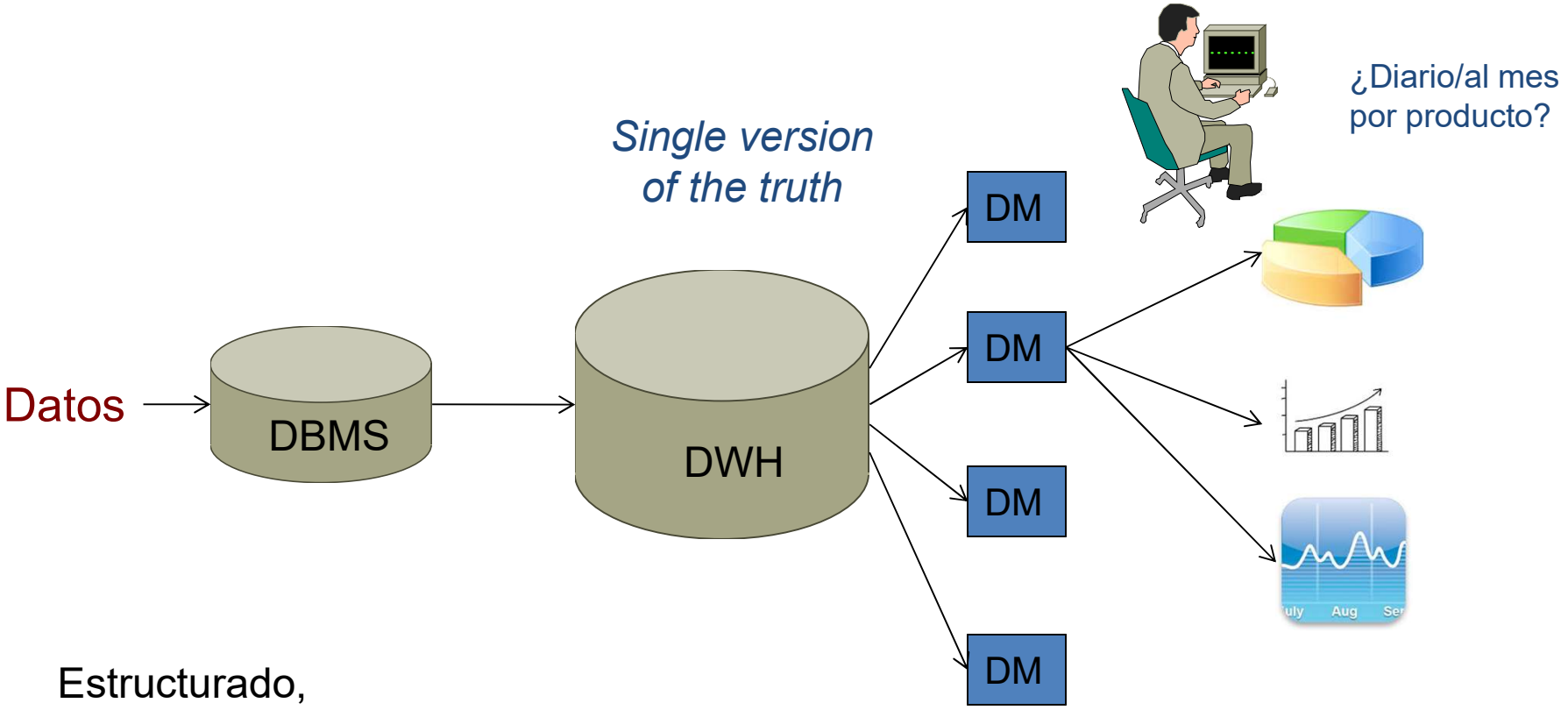
Datos no estructurados

- Muy poca uniformidad entre distintas fuentes (estilo de redacción, lenguaje, repetibilidad, …)
- Generalmente no cambian (no se actualizan) una vez creados
- Muy difíciles de analizar
  - Almacenados en formatos y lugares distintos
  - Terminología heterogénea
  - Grandes volumenes de información

Ejemplos

- eMail
  - Muchos irrelevantes; algunos con información de gran valor. Cortos, con anexos, …
- Hoja de cálculo
  - Gran ubicuidad, pero su formato depende de quién la diseña
- Transcripción conversaciones telefónicas
  - Siempre tienen un porcentaje de error. Se pierde el componente “emotivo”
- Registros médicos
  - Terminología especializada y, desgraciadamente, suele variar de un sujeto a otro

Ambiente de análisis clásico



Estructurado,
orientado a transacciones

DSS

BI

Análisis de textos

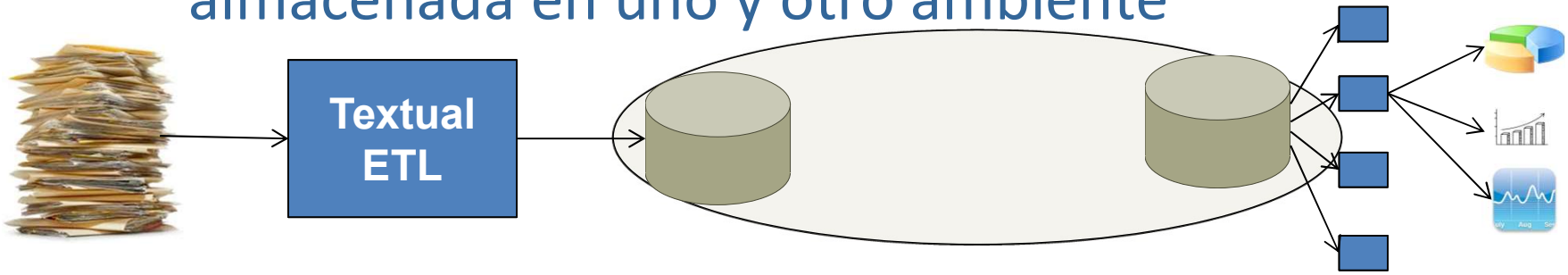
- Busca convertir el texto no estructurado en datos útiles para su análisis
  - Clasificación
  - Categorización, rankeo de páginas
  - Extracción de patrones relevantes
  - Análisis de sentimientos
  - Identificación de tópicos de interés
  - Detección de patrones y tendencias
  - …

Análisis de datos semi/no estructurados

- Todo (el documento) debe ser leído secuencialmente al menos una vez
  - Se pueden ir creando tags o índices para acceso posterior
- ¿Análisis sobre los datos fuente, o preproceso para llevarlos al entorno estructurado?
  - Ambiente no estructurado
    - DBMS capaz de manipular grandes volumenes
    - Indexación a gran variedad de datos
    - Herramientas analíticas especializadas
    - Duplicar/adaptar mucha infraestructura
  - Ambiente estructurado
    - Proceso largo de pre-acondicionamiento e integración
    - Ok para análisis de datos en reposo

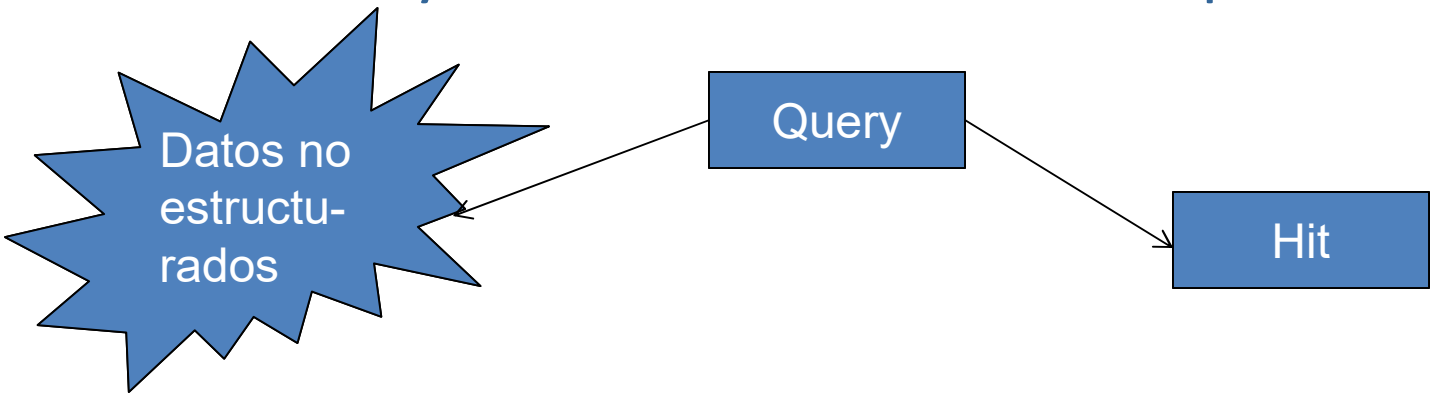
Análisis de datos (texto) en ambiente clásico

- Mucho menor costo que duplicar infraestructura
- Más sencillo
- Permite combinar datos estructurados con no estructurados.
  - Gran potencial para extraer valor e inferir relaciones no triviales en la información almacenada en uno y otro ambiente



Primera generación análisis de texto

- Consultas a través de búsquedas sobre datos no estructurados
  - Acceder y analizar texto buscando palabras



- Algunas tecnologías pueden buscar patrones
  - Número telefónico, RFC, …
  - Datos semiestructurados

Fuentes de texto

- Archivos
  - doc, txt, rtf, pdf, ps, ppt, xls, …
  - Digitalización, OCR
- Campos en BD
  - Comment, misc, …
- Transcripciones
  - Reconocimiento de voz , transcripción automática
- Video
  - Doblajes, subtítulos

Hit

- Si el análisis se hace con base en una consulta, se sabe lo que se está buscando
  - La minería de datos es el “arte de descubrir” información oculta
- ¿Reportar la primer ocurrencia o todas?
- Formato del reporte
  - XML con metadata
  - JSON
  - Excel
  - CSV
  - BD

Consultas

- Búsqueda de texto
  - Análisis sintáctico
    - Relativamente simple, basado en análisis gramatical
    - Puede requerir soporte a permutaciones, expresiones booleanas, comodines
      - Andrés Manuel López Obrador, C. López Obrador, AMLO, Andrés Manuel, Presidente López, …
  - Análisis semántico
    - Permite entender el contexto
    - Muy complejo, procesamiento de lenguaje natural (NLP)
    - Sorprendentes avances en los últimos años

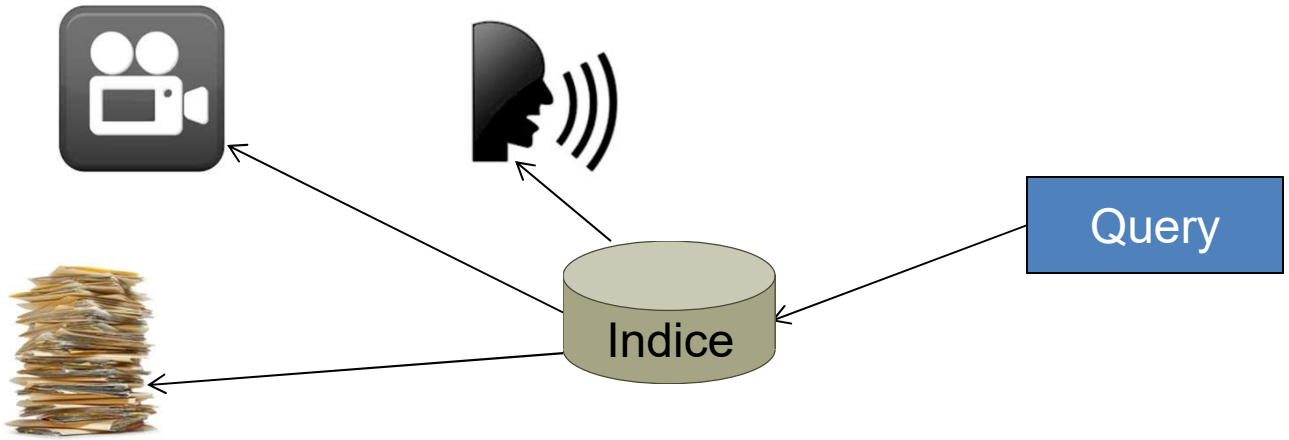
Ejemplo

- “Él se pasó con Laura”
  - ¿Quién es “él”, quién es Laura?
  - ¿Cruzaron juntos un umbral? ¿Le dio algo de gran valor? ¿La ofendió?

- El significado no puede ser derivado de la estructura gramatical, sino del contexto de la frase (probablemente…)

Consultas

- Es frecuente que las consultas se hagan sobre un índice y no sobre los documentos originales (datos crudos)
  - Si se espera que las consultas se hagan repetidamente



Indice

- Puede formarse
  - Con un *crawler* que busca información nueva y actualiza el índice
  - Con etiquetas (Tags)
    - Crowdsourcing: De-li-ci-ous, facebook, televisa
    - Metadata insertada en la fuente (si lo permite) o en el índice
- Creación de hipervínculos hacia secciones del documento y entre documentos si la fuente lo permite
- Generación
  - Diccionario, “bag of words”, rankeo de páginas, …

Definiciones útiles (aunque informales)

- Vocabulario controlado – Conjunto de términos relevantes para la organización, con una definición precisa
- Taxonomía – Jerarquización de un grupo de términos en el vocabulario
- Theasaurus – Relaciones asociativas además de jerárquicas
- Ontología – En AI, lenguaje formal para representar el significado de términos en un dominio específico

Enterprise content management

- Procesos, metodologías, herramientas para gestionar el ciclo de vida de información documental y otros contenidos en la organización
- Inicialmente, ECM contempla tecnologías que permiten identificar texto no estructurado y almacenarlo para simplificar su consulta posterior

Segunda generación análisis de texto

- Integración de datos no estructurados en ambiente estructurado
  - Distintos contextos, tecnologías, organización, estructura, funcionalidad, justificación,…
- Datos estructurados
  - Evolución natural de TI para dar soporte a las operaciones esenciales de la empresa
    - Fuertemente vinculados a sistemas transaccionales
- Datos no estructurados
  - Refleja las comunicaciones entre actores
    - Análisis informal de información

Incorporación de texto a ambiente estructurado

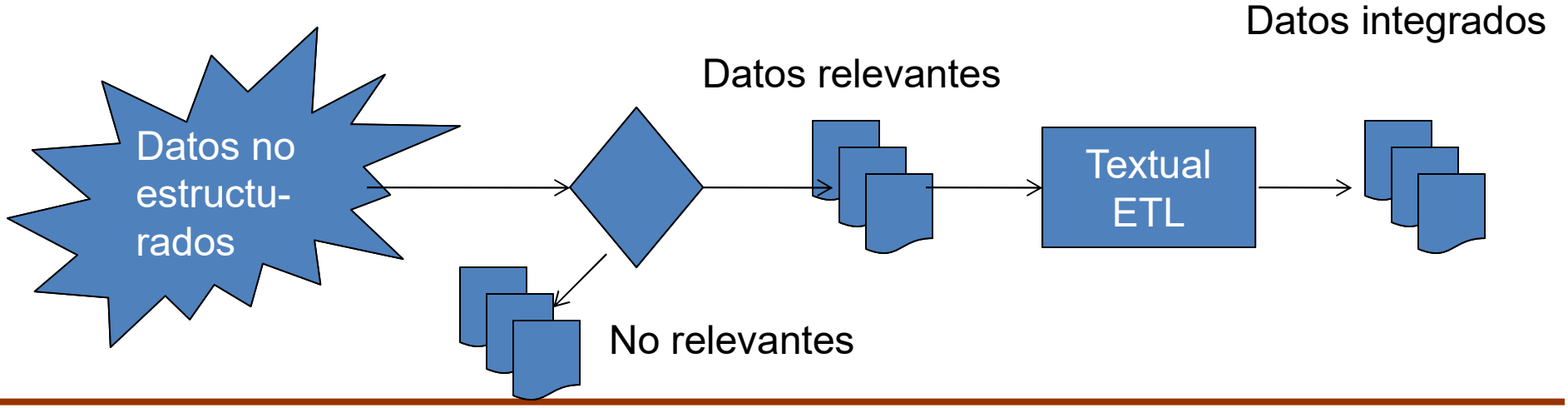
- Se requiere de un procesamiento (integración) para que el texto pueda incorporarse al ambiente de análisis estructurado
  - Texto está “sucio”, contiene errores
    - Ortográficos, gramaticales, mal uso del lenguaje, redundante, irrelevante, …
- Varias estrategias
  - La selección depende en parte del volúmen de texto relevante para la organización, que deba ser incorporado a las herramientas de análisis

Ingesta de texto

- Identificación de las fuentes de datos
  - Documentos físicos, transcripciones de conversaciones, navegación en web interno, …
- Lectura de documentos
  - Documentos físicos requieren de escaneo y OCR
    - Porcentaje no despreciable de errores que deben ser corregidos manualmente
  - Documentos electrónicos
    - Ajustar al diccionario de datos, taxonomías, etc.
  - Fuentes de voz
    - Herramientas VCR -> 5% tasa de error, pérdida de componentes emocionales, …
  - Selección de formato “oficial” de salida

Integración y precondicionamiento

- Determinar si el texto es relevante para la organización
  - Frecuentemente, decisión basada en reglas
- Remover "stop words" que no aportan valor al significado
  - "Un" "que" "el" "y" "para" …



Integración y precondicionamiento

- Reducción a raíces etimológicas
  - Búsqueda más simple. Unifica palabras similares sintácticamente distintas
    - Mover, movió, mueve, movimiento, moviendo, …
- Resolución de sinónimos
  - En función del vocabulario común
  - Remplazo a término general o concatenación
- Resolución de homónimos
  - BD -> Base de datos, Big Data, Business Decision, …
  - Vocabulario común y contexto

Integración y precondicionamiento

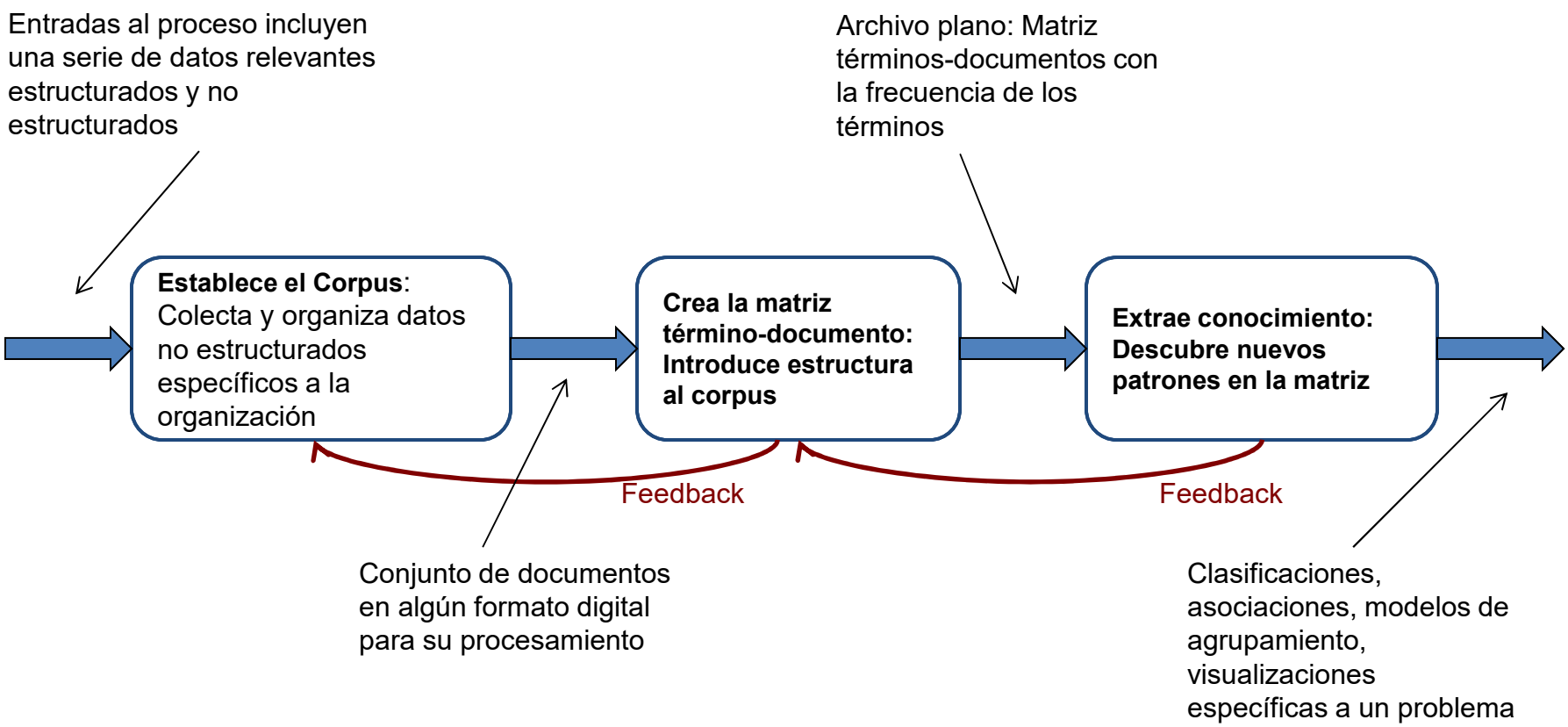
- Soporte a palabras y frases conjuntas
  - "Gran Canal", "Río Mixcoac" "Barranca del Muerto"
- Tolerancia a errores/diferencias ortográficas
  - Acentos, letras invertidas, tildes, mayúsculas, …
- Identificación (y exclusión) de frases negativas
  - "… no es big data…"
- Consolidación de documentos con temáticas similares

Ejemplo: Análisis de documentos para motor de búsqueda

- Leer fuentes (distintos códigos) y representarlas en un formato unificado
- Identificar idioma del texto
- Identificar los tokens (sentencias, párrafos)
- Generar raíces y hacer análisis morfológico para buscar en diccionario
- Actualizar diccionario

Fuente: Hammp, IBM

Text Mining Process



Matriz términos /documento

Terms / Documents	investment risk	project management	software engineering	development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

Requerimientos para extracción de información

- Calidad de las respuestas es esencial!
  - Obtener una precisión alta para la analítica de textos requiere de programas “complejos”
  - Usabilidad: Se requiere de un lenguaje de alto nivel para simplificar el desarrollo de estos programas
- Escalabilidad es crítica, pues el tiempo de ejecución puede variar considerablemente

Número creciente de herramientas, muchas software libre

- Aylen
- Keatext
- KNIME
- Refinitiv
- Apache OpenNLP
- Google NL API
- GATE
- RapidMiner
- KH Coder
- VisualText
- TAMS
- Carrot2
- Apache Mahout
- QDA Miner
- Aika
- LingPipe
- ---

Referencias

- Inmon, W., Nesavich, A., *Tapping into unstructured data Integrating Unstructred Data and Textual Analytics into Business Intelligence*, Prentice Hall, 2008
- Reiss, F. Chiticariu, L., Yunyao Li. *Varias referencias en web,* IBM Research – Almaden
- Witten, I., *Text mining*, U. ov Waikato, New Zeland