

# Bitext - The Bits and Text Company

**Tecnología de lenguaje natural  
y análisis de contenido generado por usuarios**

*Enrique Torrejón  
Antonio S. Valderrábanos  
info@bitext.com  
<http://www.bitext.com>*



# Índice

- Tecnología de lenguaje natural y extracción de información
- Extracción de información en contenido generado por usuarios
- Análisis de opiniones para inteligencia de negocio
  
- Producto para inglés
- Producto para castellano
  
- Bitext: la industria de buscadores y análisis de contenidos



# Tecnología de lenguaje natural y extracción de información

- DataSuite:
  - SDK para procesamiento de lenguaje natural
  - Procesamiento basado en reglas, no en estadística
  - Disponible para castellano, inglés, francés, catalán y euskera
- DataSuite consiste en:
  - DataLexica: información morfológica y de categorías
  - DataGrammar: analizador sintáctico/semántico
  - DataSpell: corrector ortográfico
  - DataNet: relaciones semánticas



# Tecnología de lenguaje natural y extracción de información

- Características técnicas de DataSuite:
  - Desarrollado en C/C++ para Linux/UNIX y Windows
  - API flexible
  - Integración con cualquier aplicación de terceros
- Rendimiento de DataSuite (Intel Pentium 4, 2.8GHz):
  - Procesamiento léxico: 1.000.000 palabras por segundo
  - Procesamiento sintáctico: 1.000 frases por segundo



## Contenido generado por usuarios

- Contenido generado por usuarios:
  - Chats, foros, RSS, aplicaciones, podcasts, redes sociales y blogs
- Un número de blogs en crecimiento constante
  - Technorati rastrea 112,8 millones de blogs (febrero de 2008)
- Cada vez más consumidores comentan sobre marcas y productos en sus blogs y posts
- Los estudios muestran que los consumidores cambian de opinión tras leer blogs que comentan sobre un producto
- Inteligencia de negocio: las instituciones necesitan saber qué se está diciendo sobre su marca/producto para definir estrategias de marketing, mejorar productos, etc.



## Análisis de opiniones para inteligencia de negocio

- Enfoque de Bitext: tecnología de lenguaje natural
  - DataSuite, adaptado especialmente para entender el estilo libre de escritura de los blogs
- Determinación de la polaridad de la opinión
  - Entidad: lista de nombres de marcas en diccionarios
  - Atributos de la entidad: determinados por reglas gramaticales
  - Voz: qué se dice sobre las entidades o atributos
  - Polaridad: directa o negada
- Resultado: opiniones clasificadas según
  - Relevancia, Polaridad, Voz y Atributos



## Ejemplos de determinación de la polaridad de opiniones

- **Ejemplo:** "Creo que los coches de Honda fallan a menudo"
- **Salida de DataSuite:**
  - entity: "Honda"
  - entity-type: "MARCA"
  - entity-component: "coches"
  - component-attribute: "PRODUCTOS"
  - polarity: "DIRECTA"
  - voice: "fallan a menudo"
  - verb: "Creo"
  - direct-complement: "que los coches de Honda fallan a menudo"
  - subord-subject: "los coches de Honda"
  - subord-verb: "fallan"
  - subord-adverbial-complement: "a menudo"



## Ejemplos de determinación de la polaridad de opiniones

- **Ejemplo:** "...por eso nunca me han gustado los anuncios de Nike"
- **Salida de DataSuite:**
  - entity: "Nike"
  - entity-type: "MARCA"
  - entity-component: "anuncios"
  - component-attribute: "IMAGEN"
  - polarity: "NEGADA"
  - voice: "nunca me han gustado"
  - verb: "me han gustado"
  - direct-complement: "los anuncios de Nike"
  - adverbial-complement: "por eso"
  - adverbial-complement: "nunca"





# Producto para castellano

The screenshot shows a software window titled "Business Intelligence" with a text input field containing the sentence "por eso nunca me han gustado los anuncios de Nike". Below the input are four buttons: "Analyze Sentence", "Next Sentence", "Load Corpus", and "EXIT".

The left pane displays the following analysis results:

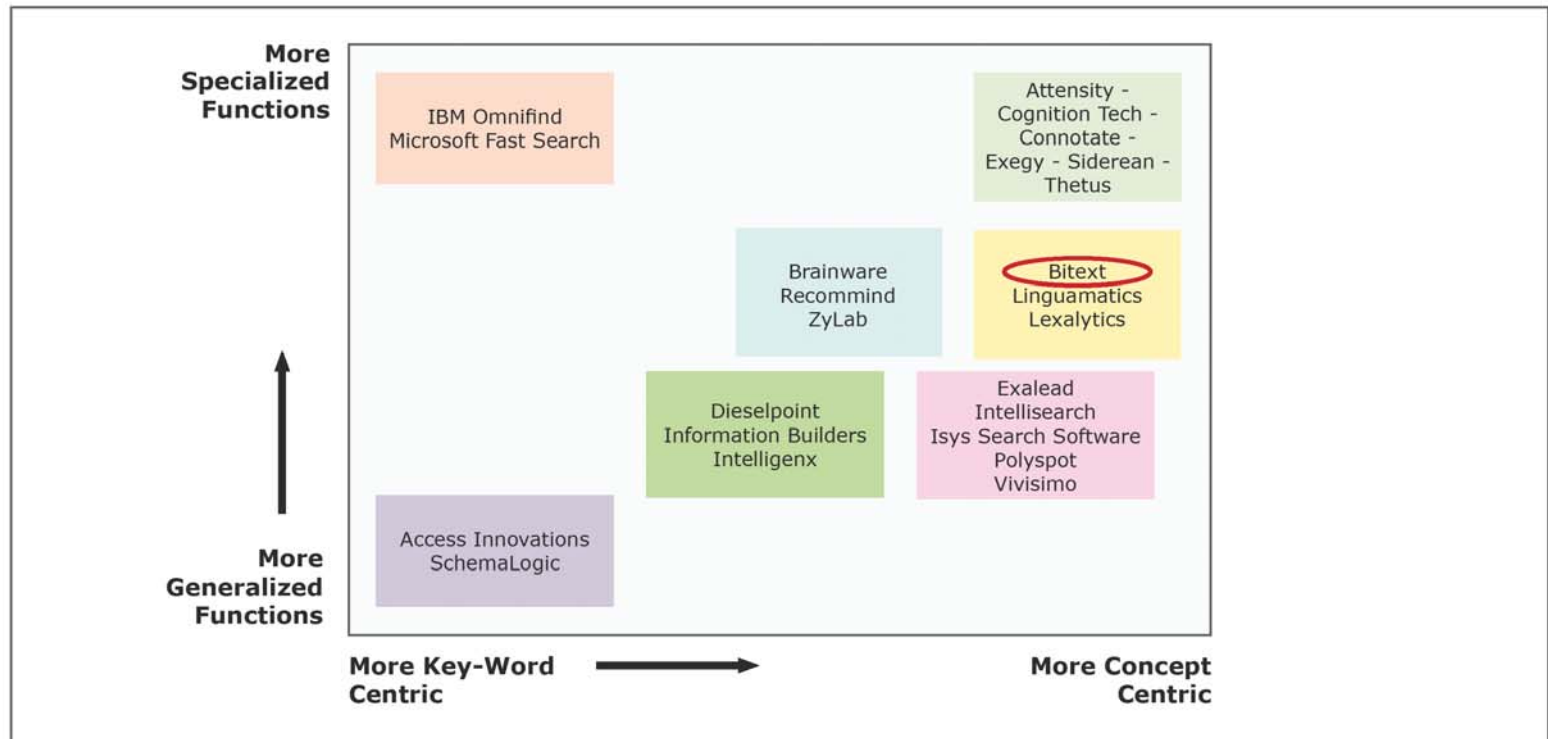
```
entity: "Nike"  
entity-type: "BRAND"  
entity-component: "los anuncios"  
component-attribute: "IMAGE"  
polarity: "DIRECT"  
opinion: "nunca me han gustado"
```

The right pane displays a hierarchical tree diagram of the sentence analysis. The root node is VOID, which branches into unknown (containing 'eso') and VVOID. VVOID branches into NVOID, which further branches into GS\_NED. GS\_NED branches into POLARITY-DIRECT-IMAGE-BRAND, which then branches into VOICE-direct. VOICE-direct branches into SADV, which contains ADV (containing 'nunca') and Voicetype3tr. Voicetype3tr branches into ADV (containing 'me') and Verbs. Verbs branches into Verbsref, which contains verb (containing 'han') and verb (containing 'gustado').

Below the VOICE-direct branch is the ATTR-brand-IMAGE node, which branches into Brand-image-PART-OF. Brand-image-PART-OF branches into GDet (containing det 'los') and sust (containing 'anuncios'). Below this is GPrep, which contains unknown 'de'. Finally, SN-ENTITY-brand branches into sust (containing 'Nike').

Key nodes and their children are highlighted with colored boxes: POLARITY-DIRECT-IMAGE-BRAND (cyan), VOICE-direct (red), SADV (red), ADV (red), me (red), Verbs (red), Verbsref (red), verb (red), han (red), verb (red), gustado (red), ATTR-brand-IMAGE (green), Brand-image-PART-OF (green), GDet (green), det (green), los (green), sust (green), anuncios (green), GPrep (green), unknown (green), de (green), SN-ENTITY-brand (yellow), sust (yellow), Nike (yellow).

# Bitext: la industria de buscadores y análisis de contenidos



Quadrant for Market Sector Vendors

Gilbane Report – “Beyond Search – What to do when your search engine does not work”

**Bitext.com**  
THE BITS AND TEXT COMPANY



¡Gracias!

Bitext - The Bits and Text Company  
Edificio Prisma, 1, 1  
Cólquide 6  
E-28230 Las Rozas Madrid  
Teléfono: +34 911461660

[info@bitext.com](mailto:info@bitext.com)

<http://www.bitext.com>

