
Bitext – NaturalExtractor - Extracción de información

Las herramientas de extracción de información desarrolladas por Bitext proporcionan todas las capacidades que hoy ofrece la tecnología lingüística a cualquier aplicación que gestione grandes cantidades de texto:

- inteligencia de negocio
- seguimiento de prensa
- análisis de mercado
- vigilancia competitiva y tecnológica
- reputación empresarial
- detección de fraude, etc.

Funcionalidad

Estas herramientas permiten extraer diferentes tipos de información de grandes bases de datos de texto:

- Entidades: nombres propios de personas, empresas, productos, lugares, etc.
 - "Barack Obama", "Agencia Española de Cooperación Internacional", "Avenida del Mar Mediterráneo"...
- Conceptos: ideas o asuntos de los que trata un texto
 - "calentamiento global", "países en vías de desarrollo", "fuentes principales de niveles de ruido urbano"...
- Eventos o relaciones entre entidades y conceptos
 - de la frase "el presidente Barack Obama ha visitado recientemente los países aliados de EEUU en el Golfo Pérsico" se extraen estas relaciones:
 - autor "Barack Obama (presidente)"
 - acción "visitar"
 - objeto "los países aliados de EEUU en el Golfo Pérsico"

Estas aplicaciones se adaptan para resolver problemas concretos como:

- detectar todos los nombramientos de nuevos cargos aparecidos en diferentes fuentes, como la prensa especializada (para sector privado) o los boletines oficiales (para sector público)
- crear una lista de todas las personas físicas y jurídicas que contraen relaciones civiles, mercantiles, etc., en documentos legales, como escrituras o contratos de compra-venta
- reunir todas las noticias de prensa sobre una empresa o un tema concretos

Además, las aplicaciones permiten la creación y mantenimiento de bases de conocimiento (diccionarios especializados) específicas para cada cliente, para abordar problemas como la catalogación automática o la creación de ontologías. Estas herramientas ofrecen multitud de posibilidades de personalización y configuración.

Bitext incrementa así su oferta de soluciones de tecnología lingüística en el campo de la extracción de información y el análisis de contenidos (conocido también como "information extraction", "text analytics" o "business intelligence").

A continuación, mostramos diferentes ejemplos de extracción de entidades, conceptos y eventos.

Ejemplos de extracción de entidades

Las aplicaciones de extracción de entidades permiten analizar un texto y obtener:

- nombres propios de lugares (Sevilla, Paseo de los Reyes), de personas (José Pérez, El Lagartija), de organizaciones (FMI, Ministerio de Fomento), etc.
- atributos de estas personas (juez, Capitán General)
- relaciones entre personas y organizaciones (José Pérez preside el FMI)

Además, la aplicación puede vincular estas entidades con diccionarios, tesauros, ontologías y otras fuentes de conocimiento para así proporcionar un contexto completo para una entidad concreta.

A continuación se muestra un fragmento perteneciente a un artículo de prensa sobre el virus H1N1:

El objetivo es aumentar esta cifra hasta los 1200 millones de dosis, de manera que se pueda atender a la vez la demanda de la gripe estacional y la del nuevo virus H1N1. Otra cuestión que habrá que debatir con los laboratorios es el precio del tratamiento. La vacuna no tiene por qué ser cara (la vacuna estacional puede costar 20 euros por persona), pero aun así habrá países que no podrán pagarla. La OMS afirma que su objetivo es conseguir por lo menos 110 millones de dosis gratis de dos laboratorios para los más pobres, o, si no, fondos del Banco Mundial o de las organizaciones de la ONU que tienen capacidad para financiarlo: Unicef y la Organización Panamericana de la Salud (PAHO), según explicó Marie-Paule Kieny, quien advirtió que todos los planes pueden variar si el virus cambia. Porque día a día se sabe más del patógeno, y no todo son malas noticias. La expansión parece haberse frenado, de acuerdo con los últimos datos de la OMS. Además, los expertos cada vez conocen mejor el agente y su comportamiento. Los Centros de Control de Enfermedades de EEUU anunciaron que han desarrollado equipos de diagnóstico rápido y que ya los habían enviado a todos los estados.

A partir del fragmento anterior, la aplicación de extracción de entidades de Bitext generará la siguiente lista de entidades, señalando si están o no presentes en el diccionario de la aplicación:

ENTIDADES DETECTADAS	PRESENTE EN EL DICCIONARIO	NO PRESENTE EN EL DICCIONARIO
<i>H1N1</i>	-	SÍ
<i>OMS</i>	SÍ	-
<i>Banco Mundial</i>	SÍ	-
<i>ONU</i>	SÍ	-
<i>Unicef</i>	-	SÍ
<i>Organización Panamericana de la Salud (PAHO)</i>	-	SÍ
<i>Marie-Paule Kieny</i>	-	SÍ
<i>OMS</i>	SÍ	-
<i>Centros de Control de Enfermedades de EEUU</i>	-	SÍ

La aplicación también puede identificar formas alternativas de nombrar a la misma entidad (por ejemplo, sugerirá que "Kieny" y "Marie-Paule Kieny" son la misma entidad).

Ejemplos de extracción de conceptos

Las aplicaciones de extracción de conceptos analizan un texto y extraen conceptos basados en estructuras lingüísticas (sintagmas nominales, preposicionales, etc.). La extracción de conceptos incluye también extracción de entidades.

Como en el caso anterior, la aplicación puede categorizar textos mediante el enlace entre conceptos y diccionarios, tesauros, ontologías y otras fuentes de conocimiento.

A continuación se muestra otro fragmento perteneciente a un artículo de prensa sobre el virus H1N1:

En contra de la situación en México, foco de la gripe, donde según el ministro de Sanidad, José Ángel Córdovas, el ritmo de transmisión está bajando. Las autoridades sanitarias de EEUU advirtieron ayer que esperan más casos, más hospitalizaciones y más muertes. La buena noticia llega de un estudio hecho a partir de los datos de los brotes de Estados Unidos, el país con más casos después de México. Un estudio hecho por Ira Longini, epidemióloga de la Universidad de Washington, en Seattle, indica que la tasa de infectividad (el número de personas que se contagian a partir de una afectada) es de 1.

A partir de dicho fragmento, la aplicación de extracción de conceptos de Bitext extraerá la siguiente lista de conceptos (en negrita las entidades):

CONCEPTOS - SINTAGMAS NOMINALES
<i>situación</i>
México
<i>foco de la gripe</i>
<i>ministro de Sanidad</i>
José Ángel Córdovas
<i>ritmo de transmisión</i>
<i>autoridades sanitarias de EEUU</i>
<i>casos</i>
<i>hospitalizaciones</i>
<i>muertes</i>
<i>buena noticia</i>
<i>estudio</i>
<i>datos de los brotes de Estados Unidos</i>
<i>país</i>
Ira Longini
<i>epidemióloga de la Universidad de Washington</i>
Seattle
<i>tasa de infectividad</i>
<i>número de personas</i>

Ejemplos de extracción de eventos

Las aplicaciones de extracción de eventos analizan un texto y extraen no sólo entidades y conceptos, sino también los eventos, es decir, los hechos o acciones realizados por las entidades y los conceptos, como por ejemplo

“José Pérez presidió la reunión de ministros de finanzas”

- Evento
 - Entidad: José Pérez
 - Tipo de evento: presidir
 - Concepto: reunión de ministros de finanzas

A continuación se muestra otro fragmento perteneciente a un artículo de prensa sobre el virus H1N1:

En España, los últimos datos del Ministerio de Sanidad indican que 9 de los 81 infectados no han estado en México. Así, la ministra Trinidad Jiménez dijo ayer en Zaragoza que la situación está muy controlada. Mientras, la guerra diplomática continúa por todo el mundo. La OMS ha pedido explicaciones a los países, como China, que han prohibido la entrada de ciudadanos de Estados afectados. España sigue su batalla para que se levante el veto a los productos porcinos en Rusia.

A partir de dicho fragmento, la aplicación de extracción de eventos de Bitext producirá los siguientes análisis de las oraciones del fragmento incluyendo entidades, conceptos e información sobre el quién, qué, cómo, cuándo, etc.:

Oración 1: “En España, los últimos datos del Ministerio de Sanidad indican que 9 de los 81 infectados no han estado en México.”

- ⇒ **Entidades:** España, Ministerio de Sanidad, México
- ⇒ **Conceptos:** últimos datos, 9 de los 81 infectados
- ⇒ **Tipo de evento:** indicar
- ⇒ **Sujeto:** los últimos datos del Ministerio de Sanidad
- ⇒ **Objeto:** que 9 de los 81 infectados no han estado en México
 - **Tipo de evento:** estar(negado)
 - **Sujeto:** 9 de los 81 infectados
 - **Lugar:** México
- ⇒ **Lugar:** España

Oración 2: “Así, la ministra Trinidad Jiménez dijo ayer en Zaragoza que la situación está muy controlada.”

- ⇒ **Entidades:** Trinidad Jiménez (ministra), Zaragoza
- ⇒ **Conceptos:** situación
- ⇒ **Tipo de evento:** decir
- ⇒ **Sujeto:** la ministra Trinidad Jiménez
- ⇒ **Objeto:** que la situación está muy controlada
 - **Tipo de evento:** estar
 - **Sujeto:** la situación
 - **Calificativo:** muy controlada
- ⇒
- ⇒ **Modo:** así
- ⇒ **Tiempo:** ayer
- ⇒ **Lugar:** Zaragoza

Oración 3: "Mientras, la guerra diplomática continúa por todo el mundo"

- ⇒ **Entidades:** ---
- ⇒ **Conceptos:** guerra diplomática, mundo
- ⇒ **Tipo de evento:** continuar
- ⇒ **Sujeto:** la guerra diplomática
- ⇒ **Tiempo:** mientras
- ⇒ **Lugar:** por todo el mundo

Oración 4: "La OMS ha pedido explicaciones a los países, como China, que han prohibido la entrada de ciudadanos de Estados afectados."

- ⇒ **Entidades:** OMS, China, Estados
- ⇒ **Conceptos:** explicaciones, países, entrada de ciudadanos de Estados afectados
- ⇒ **Tipo de evento:** pedir
- ⇒ **Sujeto:** OMS
- ⇒ **Objeto:** explicaciones
- ⇒ **Paciente:** países, como China, que han prohibido la entrada de ciudadanos de Estados afectados
 - **Ejemplo:** China
 - **Tipo de evento:** prohibir
 - **Sujeto:** países como China
 - **Objeto:** la entrada de ciudadanos de Estados afectados

Oración 5: "España sigue su batalla para que se levante el veto a los productos porcinos en Rusia."

- ⇒ **Entidades:** España, Rusia
- ⇒ **Conceptos:** batalla, veto, productos porcinos
- ⇒ **Tipo de evento:** seguir
- ⇒ **Sujeto:** España
- ⇒ **Objeto:** batalla
- ⇒ **Finalidad:** para que se levante el veto a los productos porcinos en Rusia
 - **Tipo de evento:** levantar(impersonal)
 - **Sujeto:** ---
 - **Objeto:** el veto
 - **Paciente:** a los productos porcinos
 - **Lugar:** Rusia

Bitext
The Bits and Text Company
Edificio Prisma, 1-1 - Cólquide 6
28230 - Las Rozas (Madrid)
<http://www.bitext.com>
info@bitext.com