

Breaking Cloud

Cognitive

Lab 2: Encuentra respuestas e identifica patrones en datos no estructurados

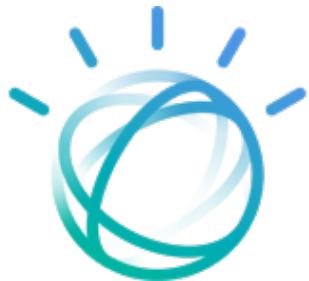


Tabla de contenidos

Introducción al laboratorio	3
1- Creamos el servicio de Watson Discovery.....	4
2- Procesamiento del lenguaje con Watson Discovery News.....	6
3- Búsquedas en Watson Discovery News.....	11



Introducción al laboratorio

En este laboratorio, adquirirás los conocimientos necesarios para utilizar Watson AI para analizar grandes volúmenes de datos, procesamiento del lenguaje natural y realizar búsquedas para identificar patrones. Para ello, utilizaremos en primer lugar el servicio de **Discovery** que nos permitirá:

- Convertir, enriquecer y normalizar los datos
- Explorar el contenido de diferentes fuentes de información
- Incluir información adicional, como conceptos relacionados o realizar un análisis de sentimiento
- Realizar búsquedas vía API

Requisitos:

- Tener cuenta de IBM Cloud
- Acceso a Internet

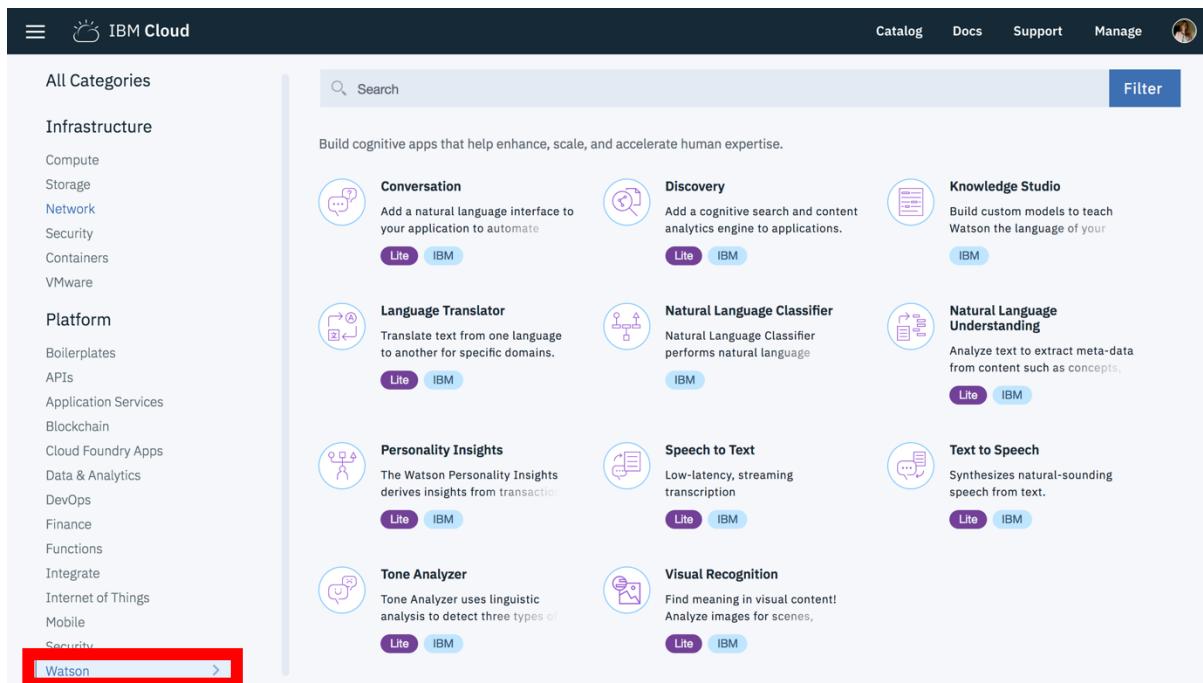
1- Creamos el servicio de Watson Discovery

Para crear el servicio de Watson Discovery, necesitamos acceder al catálogo de **IBM Cloud** desde la siguiente URL: bluemix.net

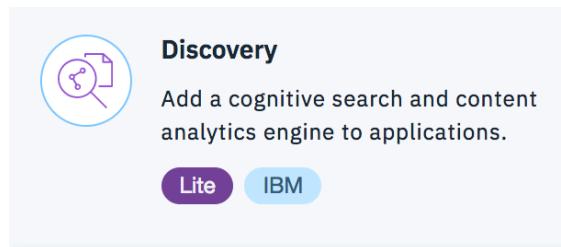
Una vez hemos accedido a **IBM Cloud**, hacemos click en catálogo (barra superior) como se muestra en la imagen:



En el menú de la izquierda, donde se muestran todos los servicios de IBM Cloud, buscamos la categoría **Watson** debajo de **Plataforma** como se muestra en la imagen y hacemos click en ella:


 A screenshot of the IBM Cloud Catalog. On the left, there's a sidebar with "All Categories" and a list of service categories under "Infrastructure" and "Platform". The "Watson" category is located at the bottom of the list and is highlighted with a red box. The main area shows a grid of cognitive services. Each service has a circular icon, a name, a brief description, and two buttons: "Lite" and "IBM". The services include Conversation, Discovery, Knowledge Studio, Language Translator, Natural Language Classifier, Natural Language Understanding, Personality Insights, Speech to Text, Text to Speech, Tone Analyzer, and Visual Recognition.

En este caso, entre todo el conjunto de servicios, vamos a elegir desplegar **Discovery**, así que lo buscamos y hacemos click sobre el mismo:



En la siguiente pantalla, debemos asignarle un nombre al servicio (por ejemplo: lab2), una región (recomendable EEUU), nuestra organización y el espacio de trabajo donde queremos desplegarlo. Hacemos click en **crear**.

IBM Cloud

View all

Discovery

Add a cognitive search and content analytics engine to applications to identify patterns, trends and actionable insights that drive better decision-making. Securely unify structured and unstructured data with pre-enriched content, and use a simplified query language to eliminate the need for manual filtering of results.

Service name: Discovery-xk

Choose a region/location to deploy in: Germany

Choose an organization: maria.borbones@es.ibm.com

Choose a space: cognitive

Pricing Plans

Monthly prices shown are for country or region: Spain

View Docs

AUTHOR IBM
PUBLISHED 02/14/2018
TYPE Service
LOCATION Sydney, Germany, United Kingdom, US South

PLAN	FEATURES	PRICING
Lite	0 - 2,000 documents (or 200 MB) per month 1,000 news queries per month 1 custom model 500 element classification pages per month See documentation for plan details	Free

Need Help? [Contact IBM Cloud Sales](#)

Estimate Monthly Cost [Cost Calculator](#)

Create

iEnhорабуена! Has completado la primera parte del laboratorio. Ya sabes cómo desplegar un servicio de Watson Discovery en IBM Cloud.

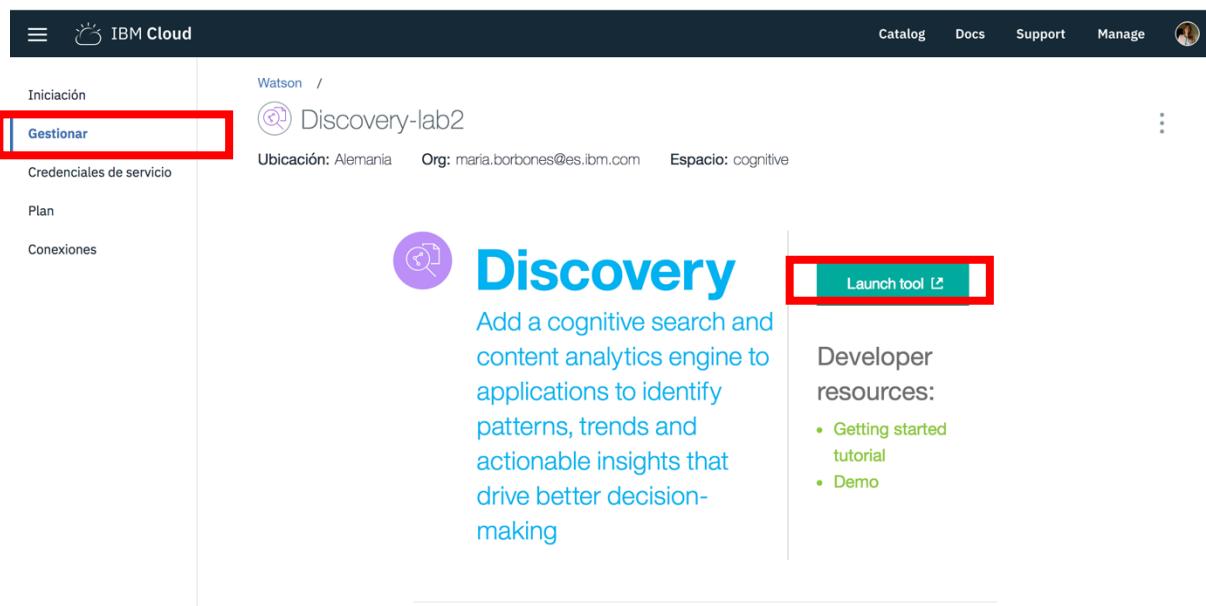
2- Procesamiento del lenguaje con Watson Discovery News

Para esta sección, vamos a utilizar **IBM Watson discovery News**, un conjunto de datos que incluye artículos publicados en internet (noticias, blogs...) y que está incluído en Watson Discovery. Está disponible en inglés, español y coreano y en concreto la versión en español se actualiza diariamente con 60.000 nuevos artículos.

Algunos casos de uso de ejemplo para **Watson Discovery** son:

- **Alertas:** crea nuevas alertas aprovechando las capacidades de discovery para extraer entidades, palabras clave, categorías, y análisis de sentimiento.
- **Detección de eventos:** La extracción del rol semántico en base a sujeto/acción/objeto permite buscar acciones/términos como “adquisición”, “resultados de las elecciones” o “OPV”
- **Trending Topics:** Identifica tópicos más populares y monitoriza como crece o decrece la frecuencia en la que son invocados

Para poder empezar a realizar búsquedas sobre Watson Discovery News desde la pantalla de gestionar del servicio, hacemos click en ‘Launch tool’ como se muestra en la imagen:



Una vez en la consola de IBM Watson Discovery, podemos ver que nos aparece una primera pantalla que nos da acceso a las diferentes colecciones de datos que tenemos disponibles. Como aún no hemos creado ninguna, ya que es la primera vez que accedemos al servicio, sólo tenemos disponible la colección de Watson Discovery News como se muestra en la imagen:

☰ IBM Watson Discovery

Manage data

Collections of your private data and pre-enriched data to configure and query against. [Learn more.](#)

[Create a data collection](#)

Watson Discovery News

PRE-ENRICHED DATA
News sources: Spanish



Como hemos adelantado anteriormente, este va a ser el conjunto de datos sobre el que vamos a trabajar en esta parte del laboratorio y que nos va a servir para entender el funcionamiento de Watson Discovery.

Vamos a acceder a la colección, para ello seleccionamos que queremos utilizar las fuentes en español y posteriormente hacemos click sobre la caja azul de Watson Discovery News como se muestra en la imagen:

☰ IBM Watson Discovery

Manage data

Collections of your private data and pre-enriched data to configure and query against. [Learn more.](#)

[Create a data collection](#)

Watson Discovery News

PRE-ENRICHED DATA
News sources: Spanish



Al acceder a la colección, se nos muestra un resumen a modo 'dashboard' de información relativa al conjunto de datos (entidades clave, sentimiento general, conceptos relacionados, palabras clave...)

Manage data > Watson Discovery News (Spanish)

Document count **3,663,663**

Número documentos en la colección

Collection info
Last updated 3/5/2018 6:40:02 am EST
[Use this collection in API](#)

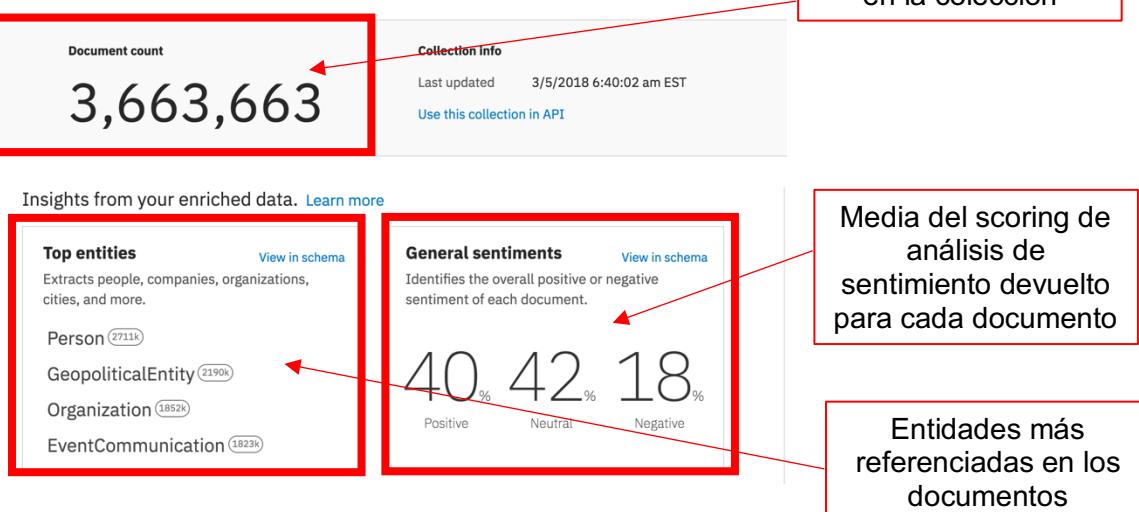
Insights from your enriched data. [Learn more](#)

Top entities [View in schema](#)
Extracts people, companies, organizations, cities, and more.
Person (2713k)
GeopoliticalEntity (2190k)
Organization (1852k)
EventCommunication (1823k)

General sentiments [View in schema](#)
Identifies the overall positive or negative sentiment of each document.
40% Positive, 42% Neutral, 18% Negative

Media del scoring de análisis de sentimiento devuelto para cada documento

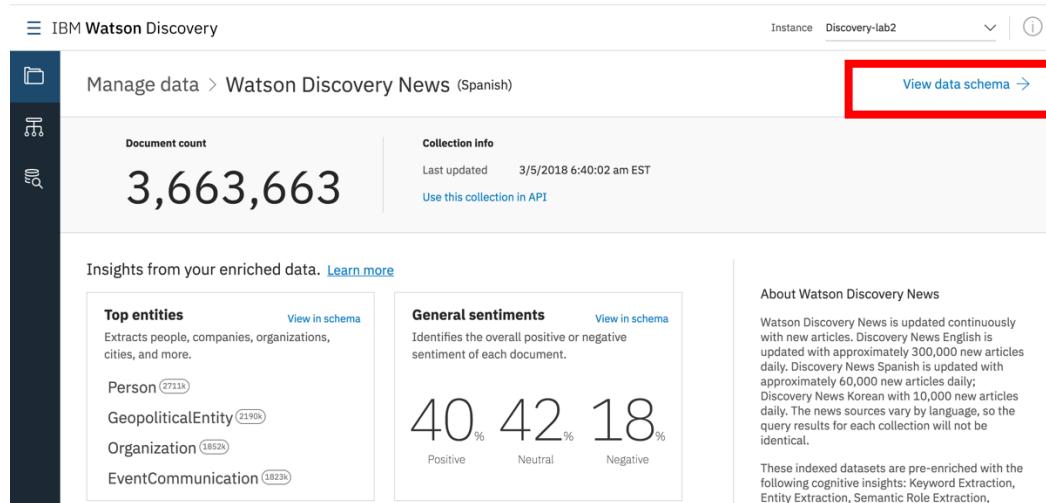
Entidades más referenciadas en los documentos





Una vez analizado el 'dashboard' vamos a ver cómo ha procesado los documentos Discovery utilizando el modelo de procesamiento del lenguaje natural que vimos con **Natural Language Understanding** para extraer entidades, conceptos, relaciones, etc.

Hacemos click sobre [View data schema](#) → como se muestra en la imagen:



This screenshot shows the 'Manage data > Watson Discovery News (Spanish)' section. It includes:

- Document count**: 3,663,663
- Collection info**: Last updated 3/5/2018 6:40:02 am EST, Use this collection in API
- View data schema** button (highlighted with a red box)
- Insights from your enriched data** section:
 - Top entities**: Extracts people, companies, organizations, cities, and more. Examples: Person (2711k), GeopoliticalEntity (2190k), Organization (1852k), EventCommunication (1823k).
 - General sentiments**: Identifies the overall positive or negative sentiment of each document. Data: Positive (40 %), Neutral (42 %), Negative (18 %).
- About Watson Discovery News** section: Watson Discovery News is updated continuously with new articles. English is updated with approximately 300,000 new articles daily; Spanish is updated with approximately 60,000 new articles daily; Korean with 10,000 new articles daily. The news sources vary by language, so the query results for each collection will not be identical.

En la siguiente pantalla vemos el esquema que ha analizado discovery y el resultado obtenido. Podemos seleccionar entre ver el resultado para la colección en general o para un documento específico como se muestra en la imagen (collection view / document view):

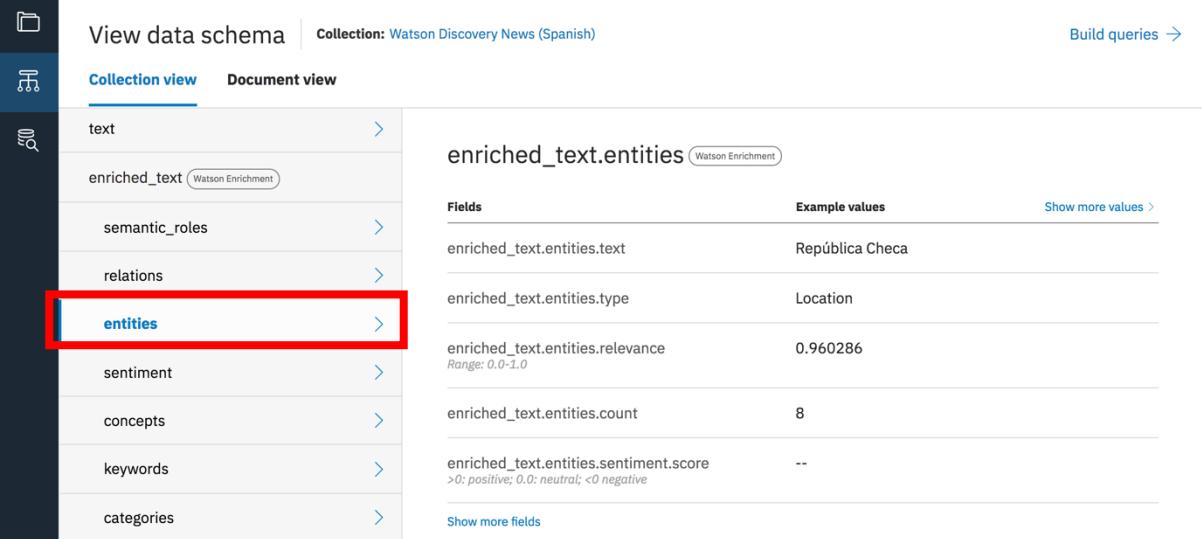


This screenshot shows the 'View data schema' page for the 'Watson Discovery News (Spanish)' collection. It includes:

- View data schema** button (highlighted with a red box)
- Collection view** and **Document view** buttons (with 'Collection view' highlighted)
- Fields** section: Shows a single field named 'text'.
- Build queries** button

Seleccionamos [Collection View](#) y te aconsejo que te detengas a ver el resultado mostrado por discovery para cada una de las categorías o 'enrichments'.

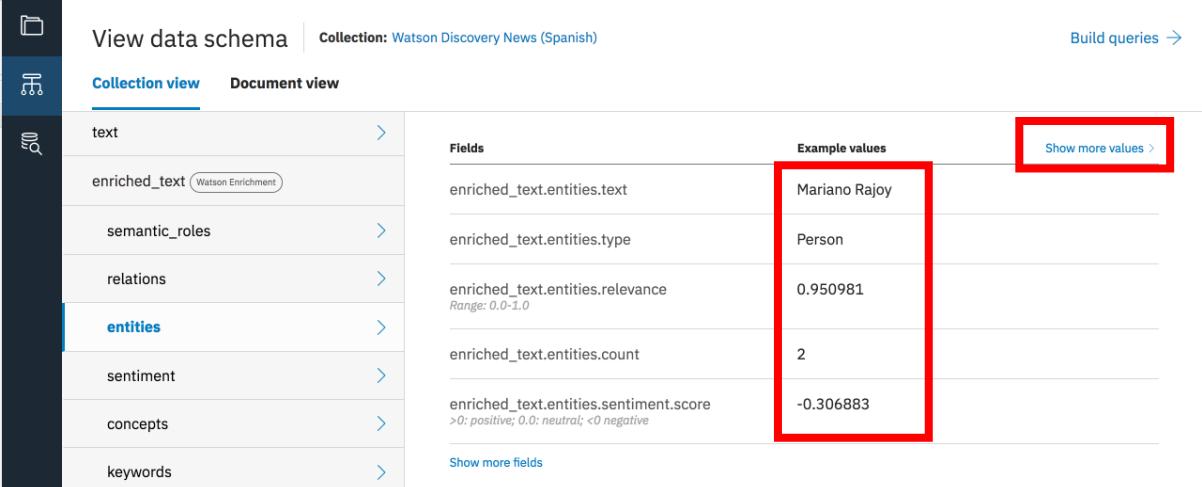
Por ejemplo, si hacemos click en **entities** vemos las entidades que ha extraído de los documentos, como **República Checa** que lo ha anotado como tipo de entidad '**Location**' con un scoring del **0.960286**



Fields	Example values	Show more values >
enriched_text.entities.text	República Checa	
enriched_text.entities.type	Location	
enriched_text.entities.relevance	0.960286 Range: 0.0-1.0	
enriched_text.entities.count	8	
enriched_text.entities.sentiment.score	-- >0: positive; 0.0: neutral; <0 negative	
Show more fields		

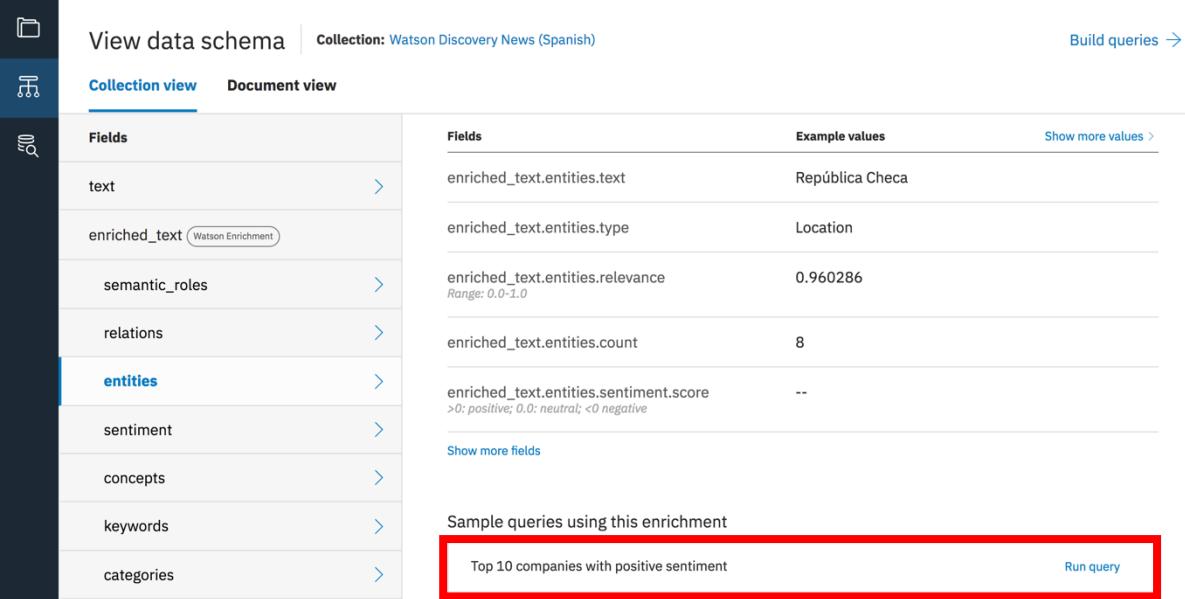
**** Importante**** los datos de vuestra vista de colección no tienen porque coincidir, ya que los documentos se van actualizando y el resultado puede variar

Si hacéis click en [show more values](#) podréis acceder otras entidades que se han identificado en los documentos:



Fields	Example values	Show more values >
enriched_text.entities.text	Mariano Rajoy	
enriched_text.entities.type	Person	
enriched_text.entities.relevance	0.950981 Range: 0.0-1.0	
enriched_text.entities.count	2	
enriched_text.entities.sentiment.score	-0.306883 >0: positive; 0.0: neutral; <0 negative	
Show more fields		

Además, Watson Discovery nos propone consultas que podemos hacer sobre los documentos referentes a las entidades :



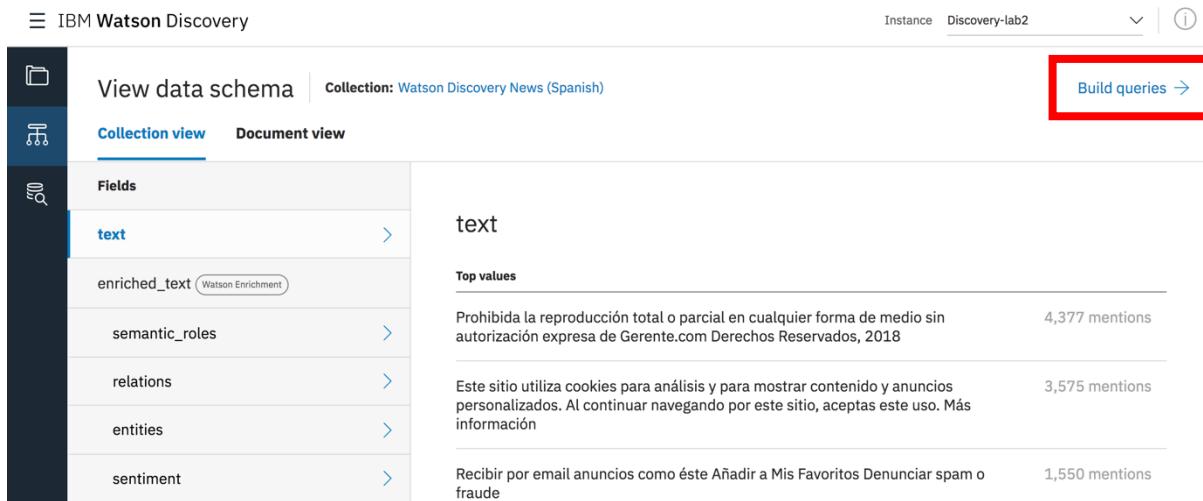
The screenshot shows the 'Collection view' of the 'Watson Discovery News (Spanish)' collection. On the left, there's a sidebar with icons for 'View data schema', 'Collection view' (which is selected), and 'Document view'. Below these are sections for 'Fields' and 'Enrichments'. Under 'Fields', there are links for 'text', 'enriched_text' (labeled 'Watson Enrichment'), 'semantic_roles', 'relations', 'entities', 'sentiment', 'concepts', 'keywords', and 'categories'. Under 'enriched_text', it shows 'enriched_text.entities.text' with an example value 'República Checa', 'enriched_text.entities.type' with an example value 'Location', 'enriched_text.entities.relevance' with a range of '0.0-1.0' and an example value '0.960286', and 'enriched_text.entities.count' with an example value '8'. There's also a link to 'Show more fields'. Below this, a section titled 'Sample queries using this enrichment' contains a button 'Top 10 companies with positive sentiment' which is highlighted with a red border, and a 'Run query' button.

Tómate un tiempo para entender la información extraída y también accede a la opción de [Document view](#) y analiza el resultado extraído pero esta vez por cada uno de los documentos que se incluyen en la colección.

¡Enhorabuena! Has completado la segunda parte del laboratorio. Ahora sabes acceder a Watson Discovery y ver el resultado del análisis.

3- Búsquedas en Watson Discovery News

Ahora vamos a realizar búsquedas sobre el conjunto de documentos, basándonos en la información extraída. Para acceder a la consola de búsquedas hacemos click en [Build queries →](#) como se muestra en la imagen:



IBM Watson Discovery

Instance: Discovery-lab2

View data schema | Collection: Watson Discovery News (Spanish)

[Build queries →](#)

Collection view [Document view](#)

Fields

- text**
- enriched_text (Watson Enrichment)
- semantic_roles
- relations
- entities
- sentiment

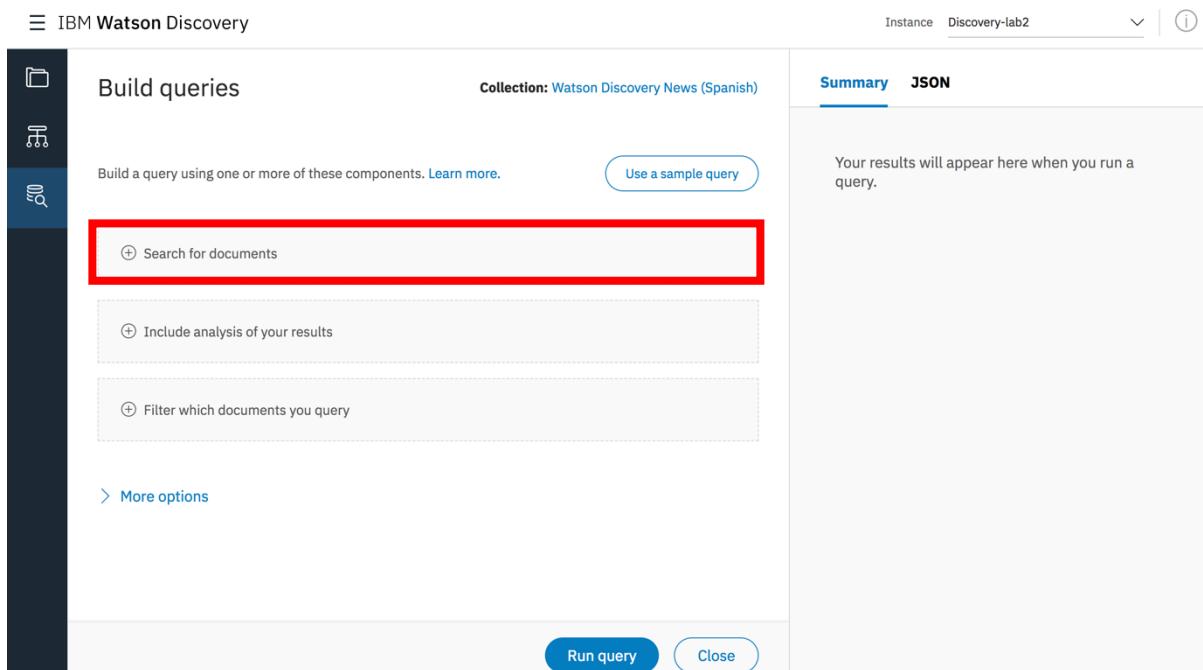
Top values

Value	Mentions
Prohibida la reproducción total o parcial en cualquier forma de medio sin autorización expresa de Gerente.com Derechos Reservados, 2018	4,377 mentions
Este sitio utiliza cookies para análisis y para mostrar contenido y anuncios personalizados. Al continuar navegando por este sitio, aceptas este uso. Más información	3,575 mentions
Recibir por email anuncios como éste Añadir a Mis Favoritos Denunciar spam o fraude	1,550 mentions

La pantalla siguiente nos muestra un constructor de consultas que nos permite realizar búsquedas utilizando lenguaje natural o el lenguaje propio de discovery sobre el conjunto de documentos.

Vamos a realizar algunas consultas para entender el funcionamiento de **Watson Discovery**.

Hacemos click en ‘Search for documents’ como se muestra en la imagen:



IBM Watson Discovery

Instance: Discovery-lab2

Build queries | Collection: Watson Discovery News (Spanish)

Build a query using one or more of these components. [Learn more.](#) [Use a sample query](#)

Search for documents

Summary [JSON](#)

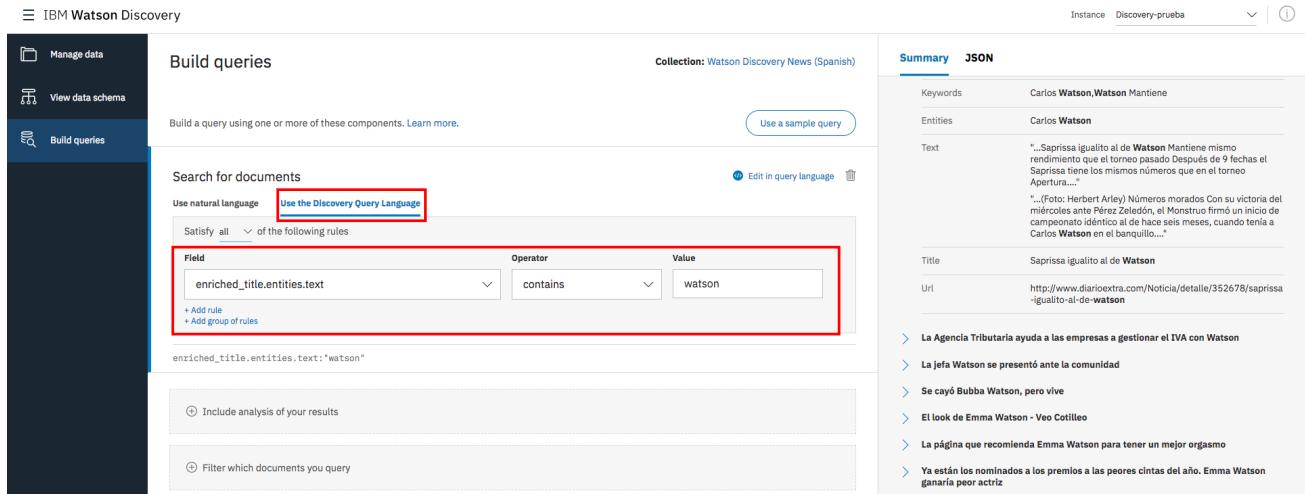
Your results will appear here when you run a query.

More options

[Run query](#) [Close](#)

La primera consulta que vamos a realizar, es buscar todos aquellos artículos en los que aparezca la entidad ‘Watson’ en el título. Si os fijáis en el campo **Field** se pueden utilizar para realizar la búsqueda todos aquellos ‘enrichments’ que hemos obtenido al realizar el procesamiento de los textos en lenguaje natural.

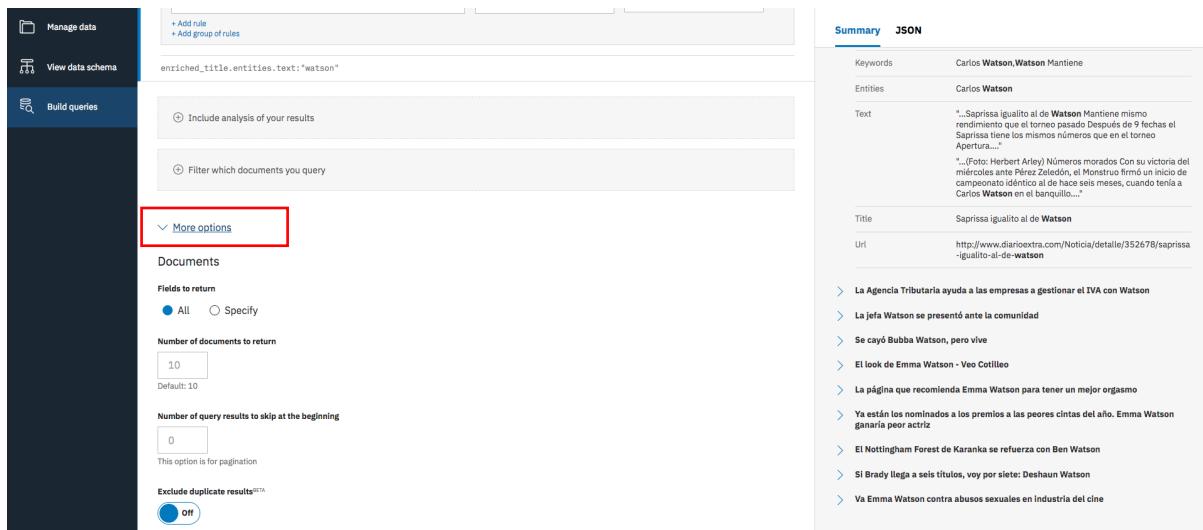
Hacemos click en [Use the Discovery Query Language](#) y completamos los campos con los siguientes valores el editor visual, tal y como se muestra en la imagen:



Keywords	Carlos Watson, Watson Mantene
Entities	Carlos Watson
Text	"...Saprissa igualito al de Watson Mantene mismo rendimiento que el torneo pasado. Después de 9 fechas el Saprissa tiene los mismos números que en el torneo Apertura..." "(Foto: Herbert Arley) Números morados Con su victoria del miércoles ante Pérez Zeledón, el Monstruo firmó un inicio de campeonato idéntico al de hace seis meses, cuando tenía a Carlos Watson en el banquillo..."
Title	Saprissa igualito al de Watson
Url	http://www.diarioextra.com/Noticia/detalle/352678/saprissa-igualito-al-de-watson

enriched_title.entities.text:"watson"

Una vez hemos completado los campos podemos hacer click en more options para delimitar los resultados de nuestra búsqueda:



Keywords	Carlos Watson, Watson Mantene
Entities	Carlos Watson
Text	"...Saprissa igualito al de Watson Mantene mismo rendimiento que el torneo pasado. Después de 9 fechas el Saprissa tiene los mismos números que en el torneo Apertura..." "(Foto: Herbert Arley) Números morados Con su victoria del miércoles ante Pérez Zeledón, el Monstruo firmó un inicio de campeonato idéntico al de hace seis meses, cuando tenía a Carlos Watson en el banquillo..."
Title	Saprissa igualito al de Watson
Url	http://www.diarioextra.com/Noticia/detalle/352678/saprissa-igualito-al-de-watson

En nuestro caso, vamos a elegir simplemente el número de documentos que queremos que devuelva, que van a ser 15 y hacemos click en ‘Run Query’.

Una vez ejecutada la query vemos que se nos devuelve un conjunto de documentos, que se muestran en el panel derecho de la pantalla como se muestra en la imagen:

[Summary](#) [JSON](#)

Results

Showing 15 of 390 matching documents

Saprissa igualito al de Watson

Sentiment	positive
Keywords	Carlos Watson , Watson Mantiene
Entities	Carlos Watson
Text	"...Saprissa igualito al de Watson Mantiene mismo rendimiento que el torneo pasado Después de 9 fechas el Saprissa tiene los mismos n�meros que en el torneo Apertura...." "...(Foto: Herbert Arley) N�meros morados Con su victoria del mi�rcoles ante P�rez Zeled�n, el Monstruo firm� un inicio de campeonato id�ntico al de hace seis meses, cuando ten� a Carlos Watson en el banquillo...."
Title	Saprissa igualito al de Watson
Url	http://www.diarioextra.com/Noticia/detalle/352678/saprissa-igualito-al-de-watson

> La Agencia Tributaria ayuda a las empresas a gestionar el IVA con Watson

> La jefa Watson se present  ante la comunidad

> Se cay  Bubba Watson, pero vive

> El look de Emma Watson - Veo Cotilleo

> La p gina que recomienda Emma Watson para tener un mejor orgasmo

> Si Brady llega a seis t tulos, voy por siete: Deshaun Watson

> Ya est n los nominados a los premios a las peores cintas del a o. Emma Watson ganaría peor actriz

> El Nottingham Forest de Karanka se refuerza con Ben Watson

En el [Summary](#) se nos muestran los documentos que nos proporciona Discovery como resultado de la b squeda y el an lisis resultado del procesamiento del lenguaje natural para cada uno de esos documentos.

Si hacemos click en **JSON**, se nos muestra el resultado del an lisis para cada uno de los documentos pero en formato JSON (resultado que ser a devuelto en la invocaci n de Discovery mediante API Rest)

Summary **JSON**

Query URL <https://gateway.watsonplatform.net/discovery/api/v1/environments/system/col>

```
{
  "matching_results": 399,
  "passages": [],
  "results": [
    {
      "id": "QnI3R7y4c5tsNZurzQt1L8BBBvWduoYl4Tl3abUWtrs6RJ1TQ4VH8gfdI5PEX2",
      "result_metadata": {...},
      "author": "Pedro Retana Cuendis",
      "enriched_title": {...},
      "entities": [
        {
          "count": 1,
          "sentiment": {
            "score": 0.595538,
            "label": "positive"
          },
          "text": "Watson",
          "relevance": 0.978347,
          "type": "Person"
        }
      ],
      "sentiment": {
        "document": {
          "score": 0.595538,
          "label": "positive"
        }
      },
      "semantic_roles": [],
      "concepts": [],
      "categories": [],
      "relations": [],
      "keywords": [
        {
          "text": "Saprissa igualito",
          "relevance": 0.99364
        }
      ],
      "crawl_date": "2018-02-09T17:08:30Z",
      "url": "http://www.diarioextra.com/Noticia/detalle/352678/saprissa-igualito-al-de-watson",
      "host": "diarioextra.com",
      "text": "Saprissa igualito al de Watson Mantiene mismo rendimiento que el torneo pasado. Después de 9 fechas el Saprissa tiene los mismos números que en el torneo Apertura. (Foto: Herbert Arley) Números morados Con su victoria del miércoles ante Pérez Zeledón, el Monstruo firmó un"
    }
  ]
}
```

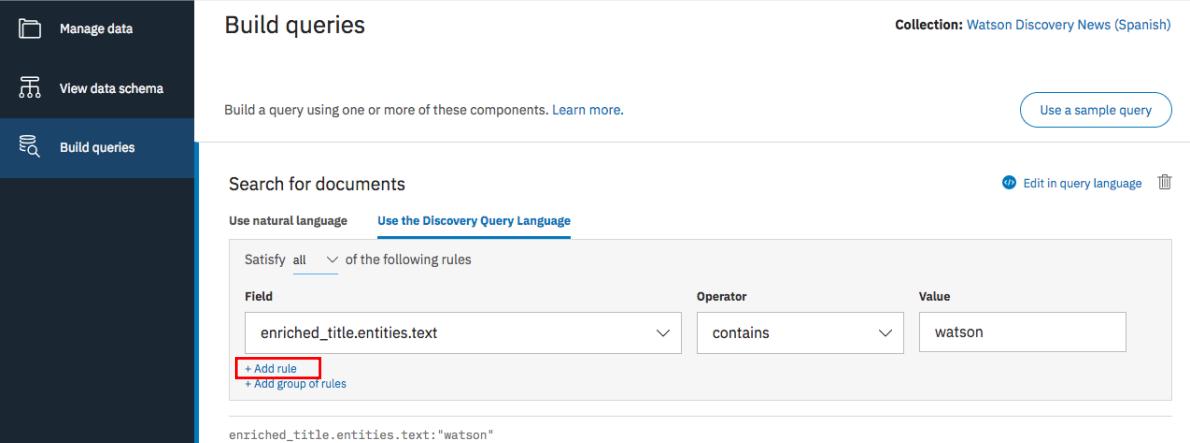
Si os fijáis, hemos hecho una consulta muy general, en la que le hemos indicado a Discovery que nos muestre simplemente aquellos documentos en los que haya encontrado una entidad con el texto '**Watson**'. Como consecuencias, nos ha devuelto artículos tanto que hacen referencia a IBM Watson como otros artículos que hablan de actores o jugadores de fútbol.

Para poder filtrar mejor las búsquedas, vamos a utilizar los **conceptos**. Como os explicábamos al inicio de este laboratorio, los conceptos son términos que no tienen que aparecer de forma explícita en el texto pero que tienen relación con el contenido de los documentos.

Así que vamos a mejorar nuestra búsqueda para que sólo nos muestre aquellos documentos que giren en torno al concepto de **tecnología**. Para ello, actualizamos nuestra búsqueda.

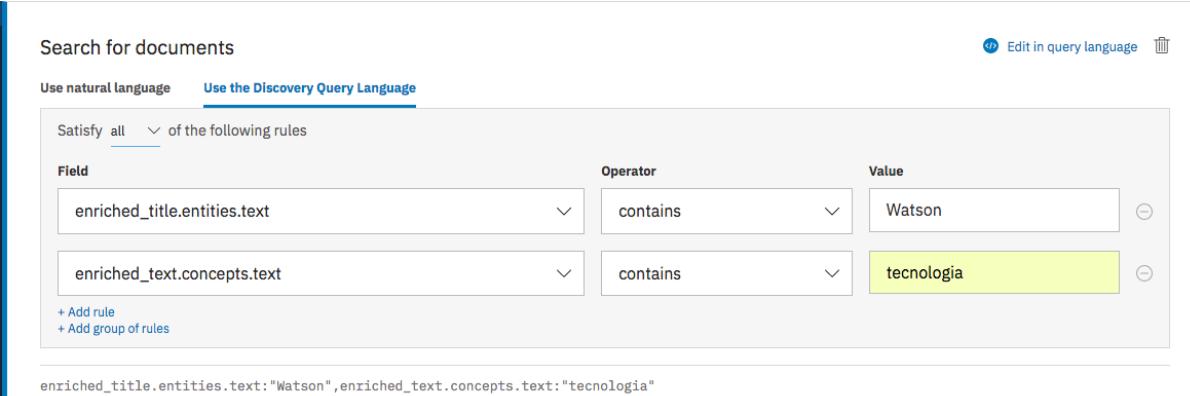
Primero, añadimos una nueva regla haciendo click en **+Add rule**

☰ IBM Watson Discovery



The screenshot shows the 'Build queries' interface for Watson Discovery. On the left sidebar, 'Build queries' is selected. The main area is titled 'Build queries' and shows a collection named 'Watson Discovery News (Spanish)'. Below the title, it says 'Build a query using one or more of these components. [Learn more.](#)' and a 'Use a sample query' button. Under 'Search for documents', there are two tabs: 'Use natural language' (selected) and 'Use the Discovery Query Language'. A rule is defined under 'Satisfy all of the following rules': 'enriched_title.entities.text' contains 'watson'. There are buttons for '+ Add rule' and '+ Add group of rules'.

Y completamos la nueva regla indicándole que sólo busco aquellos documentos en cuyo contenido aparezcan conceptos de tecnología de la siguiente forma:



The screenshot shows the 'Build queries' interface with two search rules defined. The first rule is 'enriched_title.entities.text' contains 'Watson'. The second rule is 'enriched_text.concepts.text' contains 'tecnologia'. Both rules are under the 'Satisfy all of the following rules' section. The 'enriched_text.concepts.text' rule has 'tecnologia' highlighted in yellow. The bottom of the interface shows the resulting query string: 'enriched_title.entities.text:"Watson",enriched_text.concepts.text:"tecnologia"'.

`enriched_title.entities.text:"watson",enriched_text.concepts.text:"tecnologia"`

Y vuelvo a hacer click en '**Run Query**'. Pero esta vez observamos que como resultado sólo tenemos documentos que hablan de IBM Watson:

Summary JSON

Query URL: <https://gateway.watsonplatform.net/discovery/api/v1/environments/system/collections>

Results

Showing 1 of 1 matching documents

La Agencia Tributaria ayuda a las empresas a gestionar el IVA con Watson

Sentiment	positive
Keywords	tecnología watson,servicio watson conversation
Concepts	Tecnología
Entities	Watson
Text	"...Buscando una forma innovadora de ayudar a estos profesionales a resolver sus dudas sobre el SII, la Agencia Tributaria ha desarrollado un asistente virtual con tecnología Watson capaz de responder las preguntas que les puedan surgir..." "...Buscando una forma innovadora de ayudar a estos profesionales a resolver sus dudas sobre el SII, la Agencia Tributaria ha desarrollado un asistente virtual con tecnología Watson Actualmente, los profesionales de los servicios financieros y contables de las empresas pueden entablar una conversación – gracias al servicio Watson Conversation- en lenguaje natural con el asistente virtual durante las 24..."
Title	La Agencia Tributaria ayuda a las empresas a gestionar el IVA con Watson

A modo introductorio, en este laboratorio hemos realizado búsquedas utilizando el editor visual, pero Discovery nos permite también realizar búsquedas más complejas e incluso agregaciones de datos que nos devuelvan los valores clave de los documentos, el sentimiento general de las entidades, encontrar todos los conceptos relacionados con tu colección, entre otros datos.

¡Enhorabuena! Has completado la tercera parte del laboratorio. Ahora sabes realizar búsquedas con Watson Discovery y crear tus propias reglas.

4- Creamos nuestra propia colección

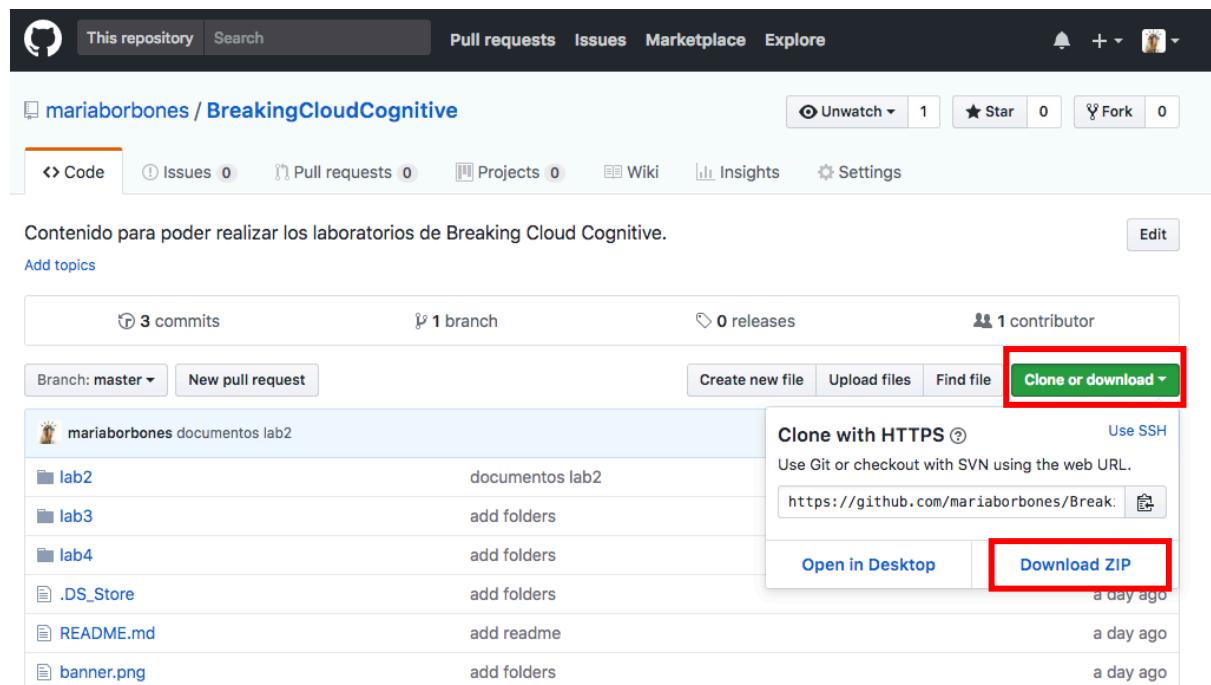
Una vez hemos visto como funciona Discovery con una colección ya existente, vamos a ver como se realiza la ingesta de datos.

La ingesta puede realizarse a través de la API, con la interfaz de usuario del propio discovery o utilizando un data crawler. En este caso, por simplicidad, vamos a utilizar la propia interfaz del servicio.

Accedemos al repositorio de github donde encontramos los materiales necesarios para poder completar los laboratorios:

<https://github.com/mariaborbones/BreakingCloudCognitive>

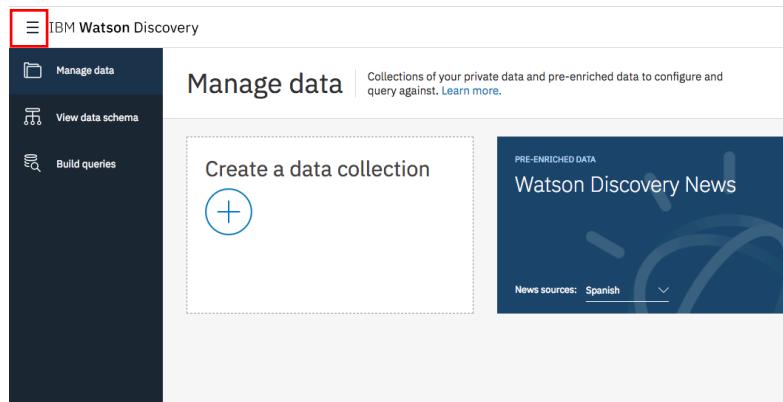
y hacemos click en el botón **Clone or download** y hacemos click en el menú en **Download ZIP** para descargarnos el material de todos los laboratorios:



The screenshot shows a GitHub repository page for 'mariaborbones / BreakingCloudCognitive'. The top navigation bar includes 'This repository', 'Search', 'Pull requests', 'Issues', 'Marketplace', 'Explore', and icons for notifications, user profile, and more. Below the header, there's a summary: 3 commits, 1 branch, 0 releases, 1 contributor. A dropdown menu shows 'Branch: master' and 'New pull request'. On the right, there's a 'Clone or download' button, which is highlighted with a red box. A sub-menu for 'Clone with HTTPS' is open, showing the URL 'https://github.com/mariaborbones/BreakingCloudCognitive'. Below this, there are two options: 'Open in Desktop' and 'Download ZIP', with 'Download ZIP' also highlighted with a red box. The main content area lists the repository's contents: 'lab2' (documents lab2), 'lab3' (add folders), 'lab4' (add folders), '.DS_Store' (add folders), 'README.md' (add readme), and 'banner.png' (add folders). Each item has a timestamp indicating it was added 'a day ago'.

Descomprimimos el fichero zip en el escritorio de nuestro ordenador para utilizar los ficheros que contiene más adelante.

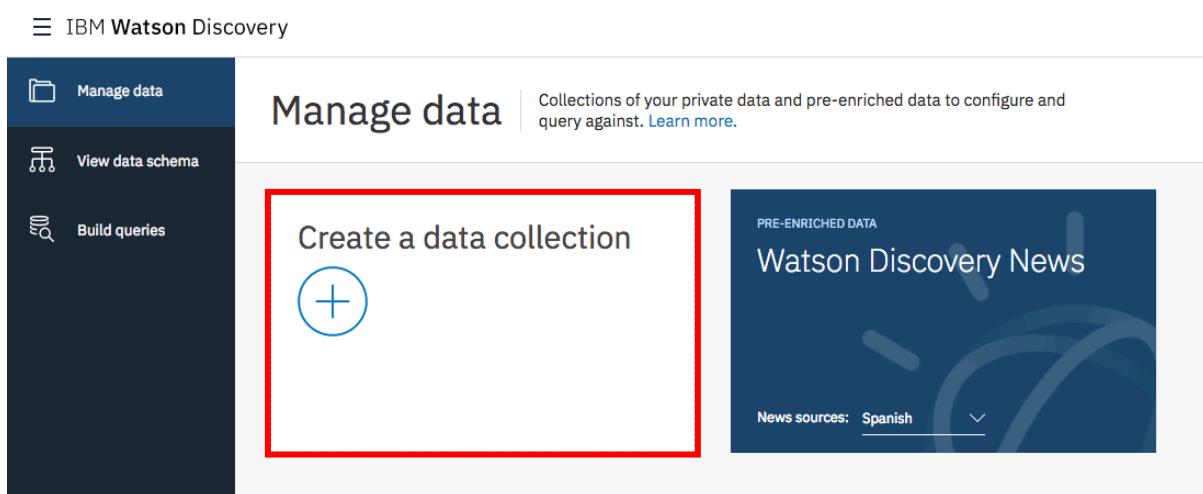
Volvemos a nuestro servicio de **Watson Discovery** y hacemos click en las tres barritas en la parte superior izquierda para acceder al menú:



The screenshot shows the 'Manage data' section of the IBM Watson Discovery interface. On the left, there's a sidebar with three options: 'Manage data' (highlighted with a red box), 'View data schema', and 'Build queries'. The main area has a heading 'Manage data' and a sub-section 'Create a data collection' with a large blue plus sign button. To the right, there's a preview card for 'Watson Discovery News' showing 'PRE-ENRICHED DATA' and 'News sources: Spanish'.

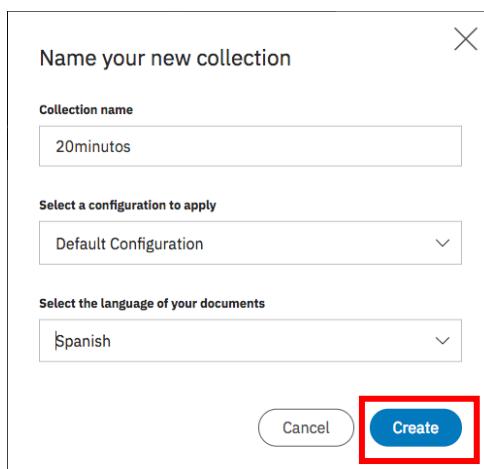
Y elegimos Manage data para acceder a las colecciones que tenemos disponible dentro de nuestro servicio.

Vamos a crear una nueva colección, así que hacemos click en **Create a data collection +**



The screenshot shows the 'Create a data collection' dialog box. It has a title 'Name your new collection' and a 'Collection name' input field containing '20minutos'. Below it are two dropdown menus: 'Select a configuration to apply' set to 'Default Configuration' and 'Select the language of your documents' set to 'Spanish'. At the bottom are 'Cancel' and 'Create' buttons, with the 'Create' button highlighted by a red box.

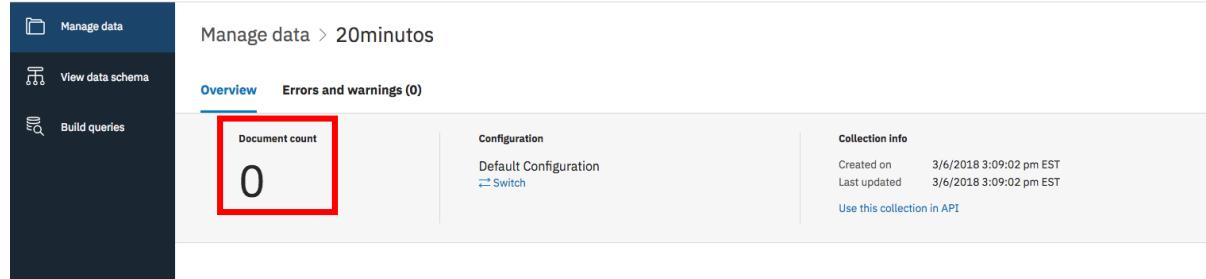
Seleccionamos un nombre (por ejemplo: 20minutos), la configuración dejamos la de por defecto y el idioma seleccionamos español y hacemo click en **create**



This is a detailed view of the 'Create a data collection' dialog box. It includes fields for 'Collection name' (20minutos), 'Select a configuration to apply' (Default Configuration), 'Select the language of your documents' (Spanish), and 'Cancel' and 'Create' buttons. The 'Create' button is highlighted with a red box.

Ya tenemos creada nuestra colección, pero de momento está vacía. De hecho, en el panel de control el número de documentos aparece a 0

IBM Watson Discovery



The screenshot shows the 'Manage data > 20minutos' section. On the left sidebar, there are three options: 'Manage data', 'View data schema', and 'Build queries'. The 'Overview' tab is selected. In the center, there's a large red box highlighting the 'Document count' field, which displays the number '0'. To the right of this, under 'Configuration', it says 'Default Configuration' with a 'Switch' link. Below that is 'Collection Info' with details: Created on 3/6/2018 3:09:02 pm EST, Last updated 3/6/2018 3:09:02 pm EST, and a 'Use this collection in API' link.

El siguiente campo nos muestra la configuración elegida para hacer la convertir, enriquecer y normalizar los documentos. ¿Y qué significa esto?

Convertir: permite definir las reglas que quieras utilizar para convertir cada uno de los formatos soportados (PDF, WORD, HTML, JSON) en el momento de la ingestión. Por ejemplo, eliminar tags HTML, combinar campos JSON o asignar tags HTMLs a fuentes de PDF o docs.

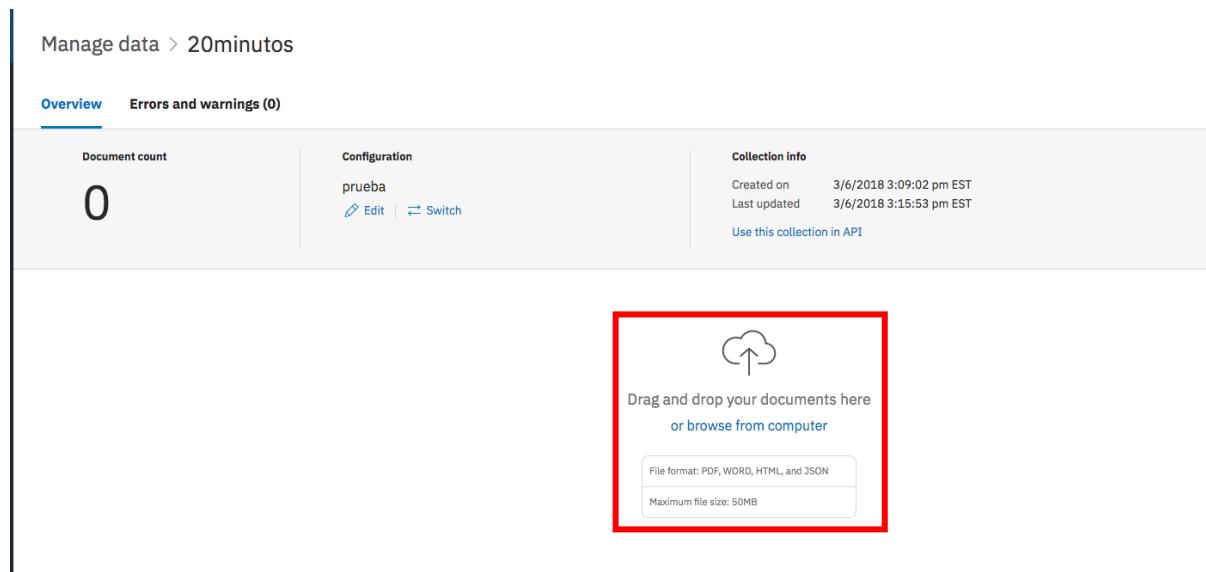
Enriquecer: permite elegir que reglas aplicar en el momento de enriquecer los documentos. Por ejemplo, utilizar un modelo propio de machine learning desplegado en Watson Knowledge Studio para detectar entidades y sus relaciones.

Normalizar: cómo hacer el proceso de refinamiento de los documentos.

En nuestro caso vamos a dejar la configuración por defecto, pero si tienes interés pregunta al instructor/a por una demo de cómo utilizar un modelo concreto de machine learning.

Y por último, se nos muestra la información necesaria para invocar Watson Discovery via API en **Collection Info**

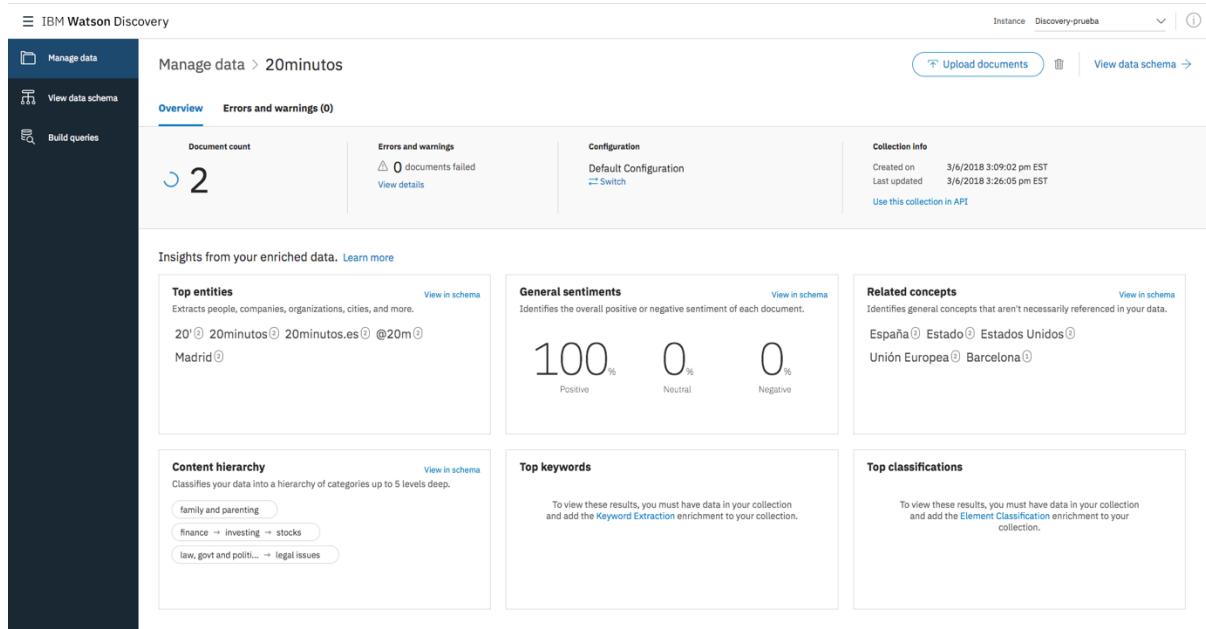
Para esta parte del laboratorio, vamos a utilizar una serie de PDFs que contienen los periódicos 20minutos del año 2018. Así que hacemos click en [browse from computer](#)



The screenshot shows the 'Manage data > 20minutos' section. The 'Overview' tab is selected. In the center, there's a large red box highlighting the 'Document count' field, which displays the number '0'. To the right of this, under 'Configuration', it says 'prueba' with a 'Edit' and 'Switch' link. Below that is 'Collection Info' with details: Created on 3/6/2018 3:09:02 pm EST, Last updated 3/6/2018 3:15:53 pm EST, and a 'Use this collection in API' link. At the bottom, there's a red box highlighting a file upload area with a cloud icon, the text 'Drag and drop your documents here or browse from computer', and a note: 'File format: PDF, WORD, HTML, and JSON' and 'Maximum file size: 50MB'.

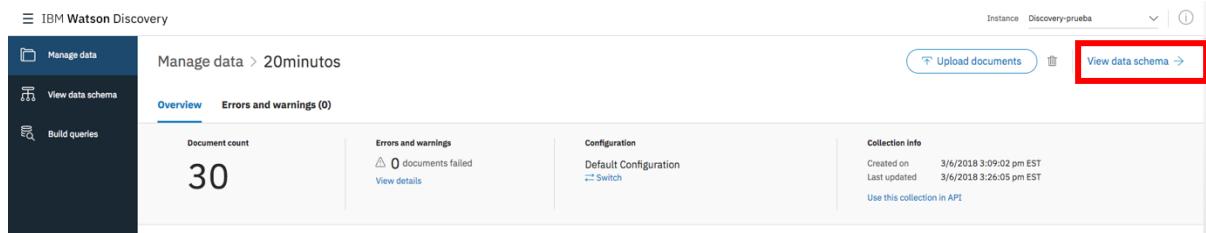
Y accedemos en la carpeta que hemos descomprimido en el escritorio, al directorio lab2 y después a la carpeta documentos. En esta carpeta encontramos todos los periódicos de 20 minutos, así que los seleccionamos todos y hacemos click en abrir. Automáticamente vemos como comienzan a subirse

nuestros documentos y como se va actualizando la información del panel de control sobre nuestra colección:



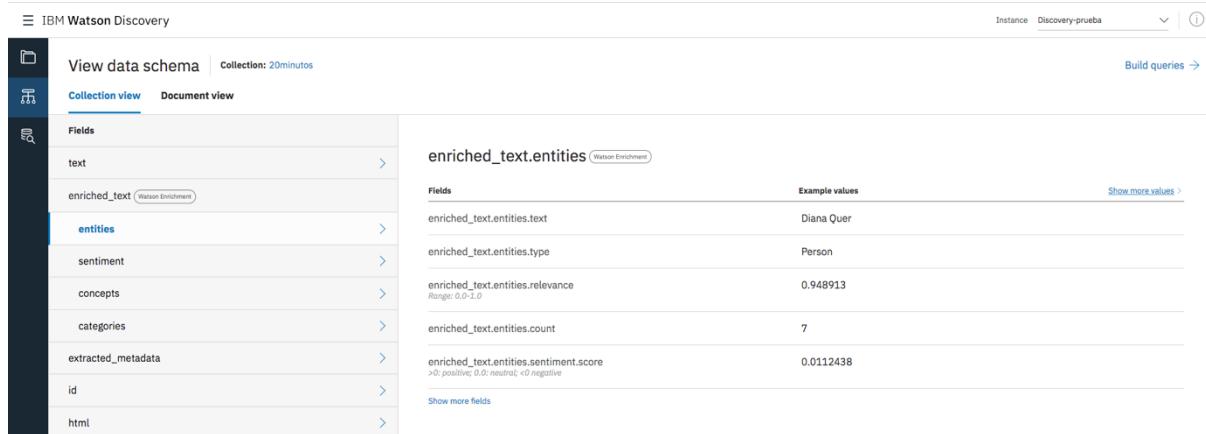
The screenshot shows the 'Overview' section of the IBM Watson Discovery interface. Key statistics include 2 documents, 0 failed documents, and a 100% positive sentiment score. Insights sections show top entities like '20minutos', '20minutos.es', and 'Madrid'; general sentiments (all positive); related concepts (like 'España', 'Estado', 'Estados Unidos', 'Unión Europea', 'Barcelona'); content hierarchy; top keywords; and top classifications.

Una vez terminado el proceso de ingestión de documentos, podemos acceder como hemos hecho anteriormente en el laboratorio, al análisis realizado de la colección. Hacemos click en **view data schema** →



The screenshot shows the 'Overview' section after the collection has been updated to 30 documents. The overall analysis remains largely the same, with 100% positive sentiment and no failed documents.

Y podemos visualizar los **enrichments** obtenidos en este caso para nuestra colección de documentos de manera global o para cada uno de los documentos analizados:



The screenshot shows the 'View data schema' page for the '20minutos' collection. It displays the 'Collection view' and highlights the 'enriched_text.entities' field under the 'Fields' section. The right pane shows detailed information about this enrichment, including fields like 'text', 'entities', 'sentiment', 'concepts', 'categories', 'extracted_metadata', 'id', and 'html'. For 'enriched_text.entities', it lists 'enriched_text.entities.text' (Diana Quer), 'enriched_text.entities.type' (Person), 'enriched_text.entities.relevance' (0.948913), 'enriched_text.entities.count' (7), and 'enriched_text.entities.sentiment.score' (0.0112438). There is also a link to 'Show more values'.

Haz click en **Build queries** → y prueba a realizar algunas búsquedas sobre los documentos como por ejemplo:

- Busca todos los documentos que hablen de la Unión Europea
- Documentos con sentimiento negativo que se hable de Puigdemont

¡Enhorabuena! Has completado la cuarta parte del laboratorio. Ahora sabes realizar ingestar tus propios documentos al servicio de discovery

5- Visualización con grafos

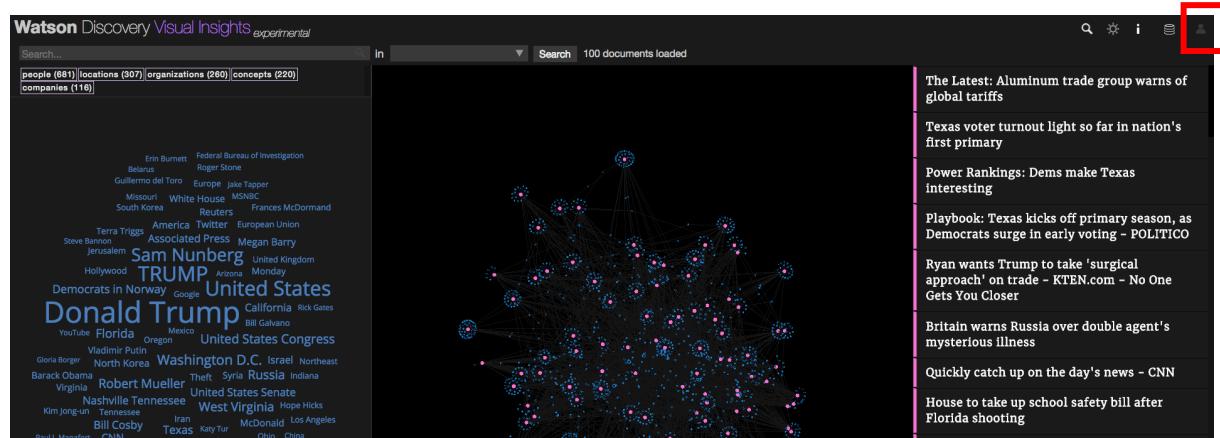
En este último paso vamos a utilizar una herramienta desarrollada por Research y que está en fase experimental para visualizar a modo de grafo nuestra colección de documentos. **Watson Discovery Visual Insights** nos permite explorar conexiones indentificadas por Discovery para entender los elementos semánticos, relaciones, conceptos y mucho más

Puedes acceder a Watson Discovery Visual Insights desde:

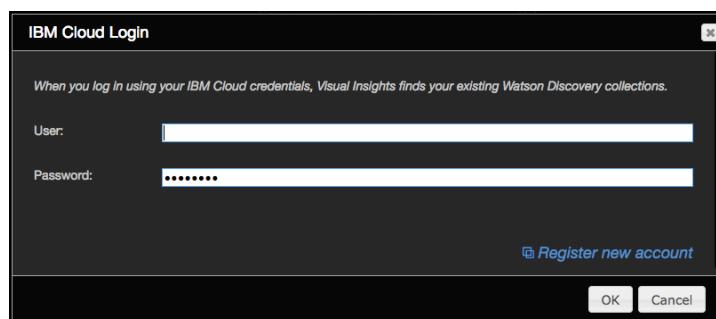
<https://visual-insights.bluemix.net>

Al acceder vemos que ya aparecen datos, porque nos muestra la representación de Watson Discovery News, pero nosotros queremos representar nuestra propia colección de documentos.

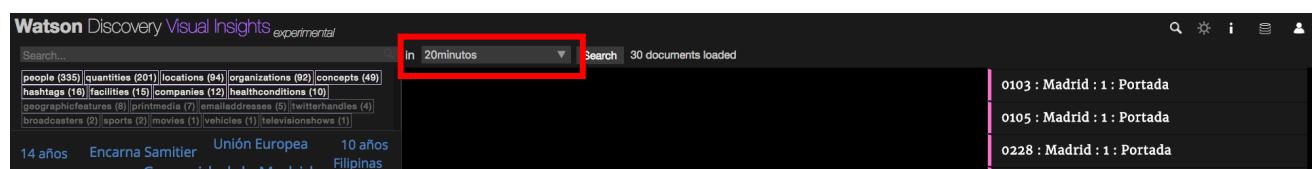
Para ello, hacemos click en el icono de login que se encuentra en la parte superior derecha de la pantalla:



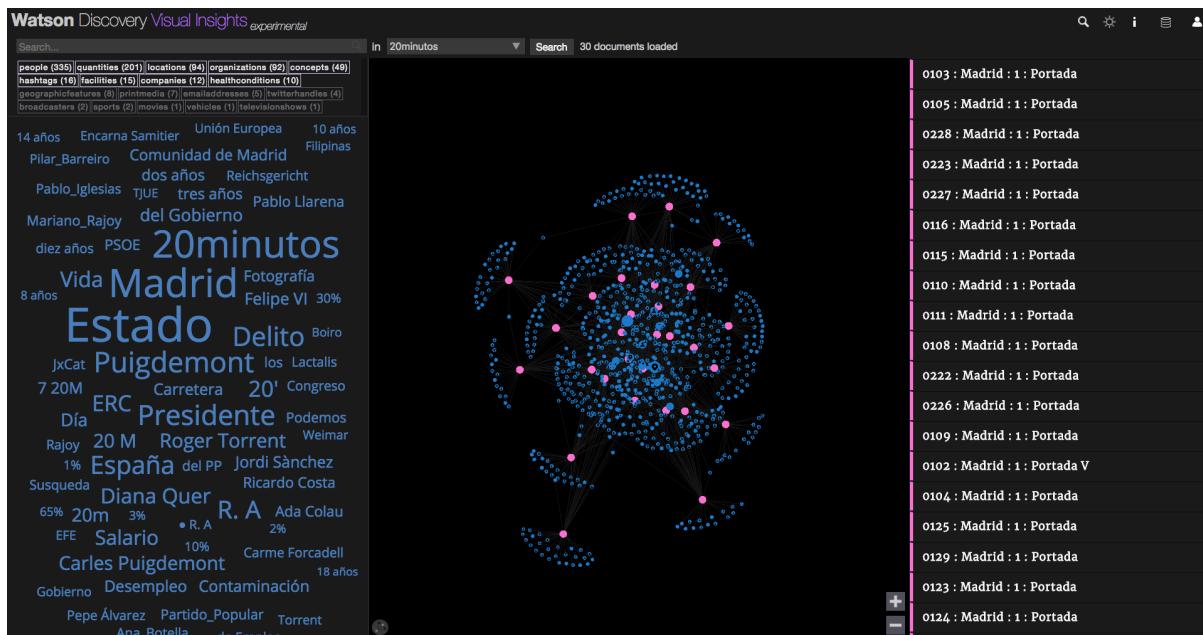
Y nos logueamos con nuestro usuario y contraseña de IBM Cloud



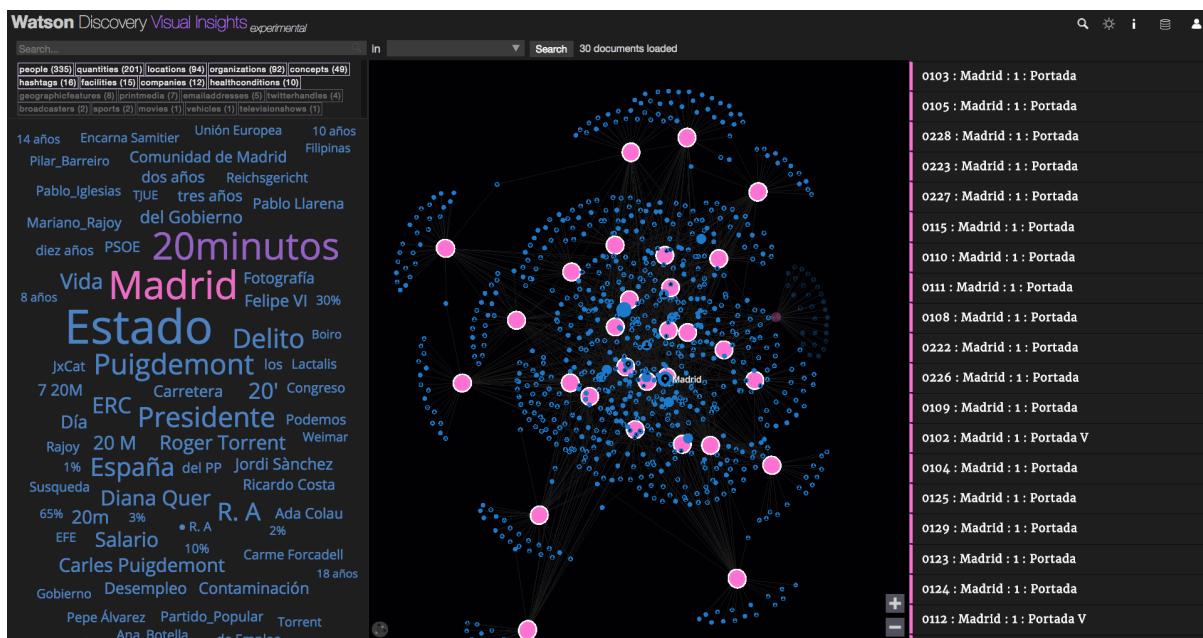
Y una vez logueados, elegimos nuestra colección de documentos 20 minutos



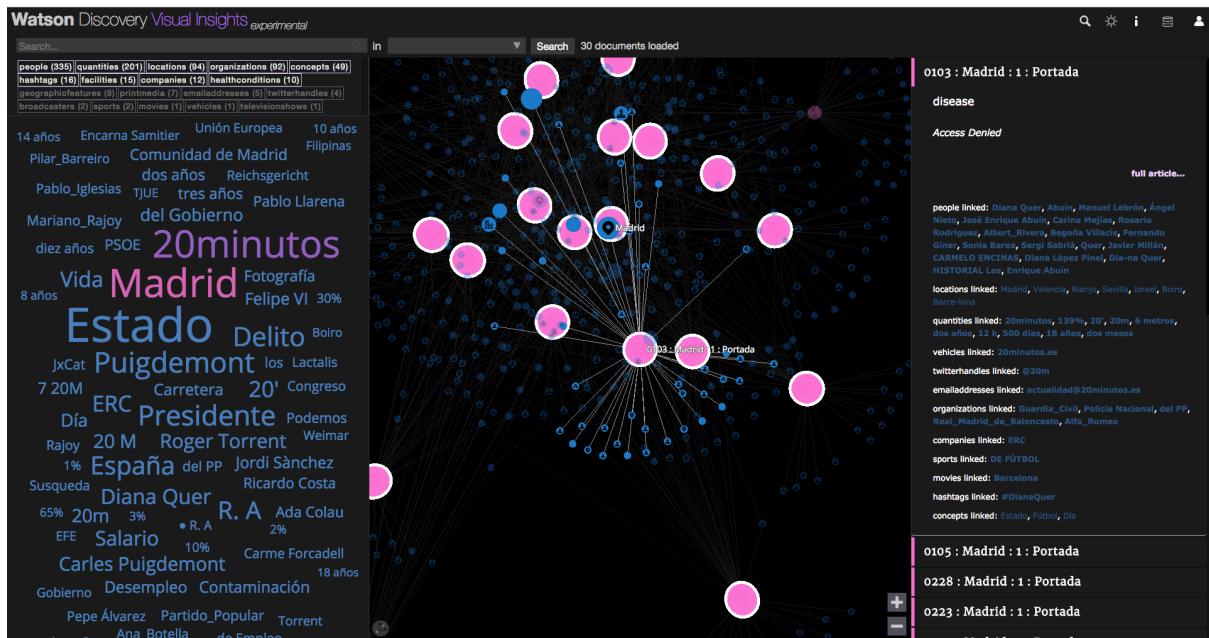
Una vez seleccionada, hacemos click en **Search** y se nos muestra el resultado del grafo para nuestra colección



En la parte izquierda de la pantalla se muestran todas las entidades que han sido extraídas de los documentos. Si hacemos click en cualquiera de ellas o seleccionamos varias vemos que documentos hacen referencia a dicha entidad/es de forma gráfica en el grafo:



Si en cambio haces click en uno de los documentos que se muestran en el panel de la derecha, se despliega toda la información específica de ese documento y además se visualizan las relaciones de forma gráfica en el grafo:



Toma tu tiempo para hacer algunas búsquedas, activar/desactivar entidades, recorrer el grafo y entender como funciona la representación.

Si quieres saber más sobre la librería desarrollada por IBM Research para crear esta visualización basada en grafos puedes acceder en:

<https://ndapi.res.ibm.com/landing/start>

¡Enhorabuena! Has completado el segundo laboratorio. Ahora sabes encontrar respuestas e identificar patrones en datos no estructurados. Además eres capaz de crear visualizaciones con grafos a partir de tus datos.