Extracción de información compleja

Cálamo Educación S.L.
María Teresa Tijeras Pascual
Cálamo Educación S.L.
María Teresa Tijeras Pascual

Indice

Extracción de información compleja

El objetivo de esta unidad es describir los fundamentos y técnicas para abordar las tareas de procesamiento del lenguaje natural dedicadas a la extracción de piezas de conocimiento más elaboradas (llamadas *"insights"*).

0. Objetivos de la unidad

El objetivo de esta unidad es describir los fundamentos y técnicas para abordar las tareas de procesamiento del lenguaje natural dedicadas a la extracción de piezas de conocimiento más elaboradas (llamadas *"insights"*), en forma de grafos semánticos de objetos y relaciones entre objetos. Se trata de una tarea de una gran complejidad, que se podría considerar como la evolución del reconocimiento de entidades.

En el **apartado 1** se presentan los fundamentos y conceptos más importantes.

El **apartado 2** se centra en describir el modelo de grafo semántico para la representación de información, ilustrando sus ventajas con un ejemplo y describiendo los diferentes estándares propuestos para su representación textual y su explotación en sistemas de bases de datos semánticas.

Por último, el **apartado 3** presenta los diferentes enfoques que se pueden aplicar para abordar la tarea de extracción de información avanzada: los modelos basados en reglas de extracción o las técnicas de aprendizaje automático, en particular, *deep learning*.

1. Introducción

El objetivo de esta tarea es la detección de información estructurada compleja o relacional, más allá del mero reconocimiento de entidades o la información aportada por etiquetas de clasificación automática. En esta tarea, además de la lista de entidades con nombre, se extraen también las relaciones entre dichas entidades y su significado específico en el texto.

En ocasiones, esta tarea se conoce como "extracción de *insights*" *(insight extraction)*, utilizando el término *insight*, muy empleado en marketing y publicidad, que se refiere a un dato clave, esencial y de gran valor que permite comprender o encontrar la solución a un problema.

Desde el punto de vista de la salida, el objetivo de la tarea es convertir el texto de entrada en un grafo semántico que contenga las entidades mencionadas en el texto (igual que en el reconocimiento de entidades, con su tipo semántico y relevancia) y las relaciones entre ellas (habitualmente incluyendo también un valor de relevancia o confianza).

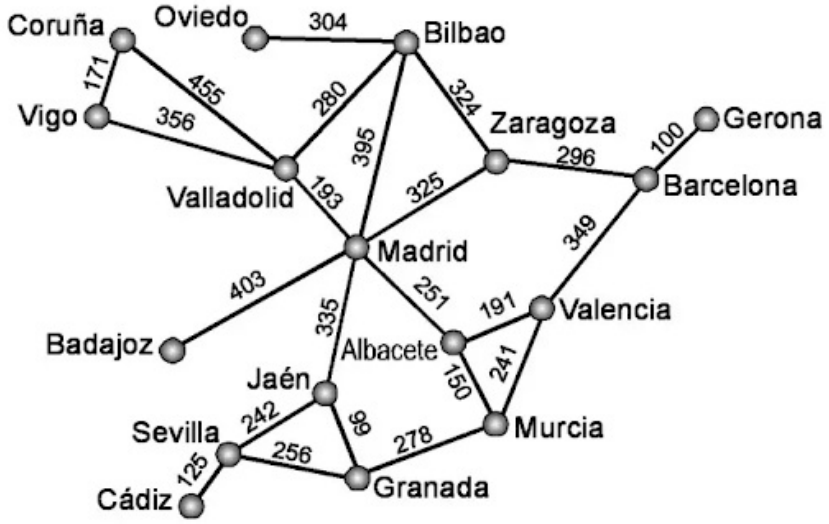
2. Modelo de grafo semántico

En matemáticas y ciencias de la computación, un **grafo** (en inglés, *graph*) es un conjunto de objetos llamados vértices o nodos (en inglés, *vertex* -plural *vertices*- o *nodes*) unidos por enlaces, también llamados aristas o arcos (*edges* o *links*), que permiten representar relaciones entre ellos.

En los grafos dirigidos, los enlaces indican expresamente que el sentido de la relación es del nodo origen al nodo destino, pero no al revés (por ejemplo, la relación "autor" en "Pablo Picasso es autor del Guernica"). En los grafos no dirigidos, la relación va en ambos sentidos (por ejemplo, la relación "hermano").

Un ciclo es un recorrido por el grafo que se inicia en un nodo y permite regresar al mismo nodo de partida sin repetir enlaces. Se denomina grafo cíclico cuando el grafo incluye algún ciclo, y grafo acíclico en caso contrario.

Por ejemplo, la figura siguiente presenta un grafo que representa un conjunto de ciudades de España (nodos) y las distancias por carretera entre ellas (aristas). Se trata de un grafo no dirigido (la distancia de A a B es igual a la distancia de B a A) cíclico (hay ciclos, por ejemplo, Coruña-Vigo-Valladolid-Coruña).

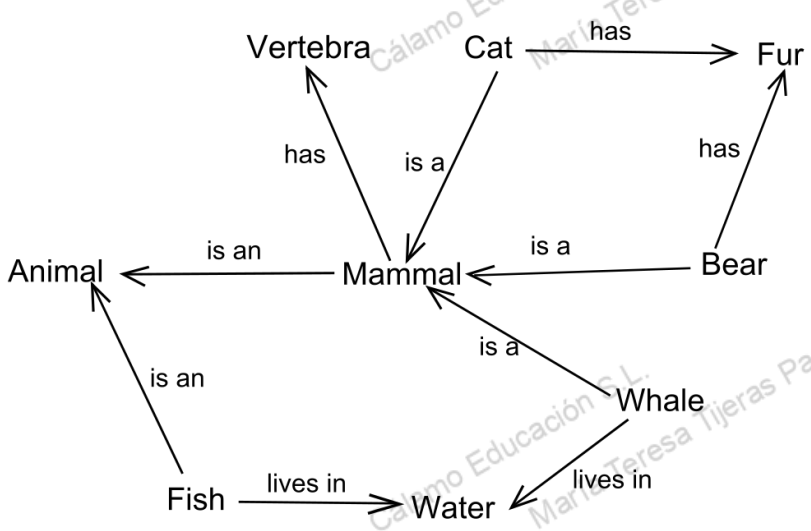


Grafo de distancias entre ciudades de España

Se considera que prácticamente cualquier problema puede representarse mediante un grafo, ya que se trata de la estructura de datos más general. Por ejemplo, un árbol se puede definir como un grafo dirigido acíclico (no hay un ciclo o camino de enlaces que permita llegar al nodo de partida).

La teoría de grafos es la rama de las matemáticas y las ciencias de la computación que estudia las propiedades de los grafos. Es muy importante ya que tiene una gran cantidad de aplicaciones en problemas de optimización, búsqueda de soluciones, gestión de procesos y flujos, etc.

Generalizando este concepto, un **grafo semántico** (o red semántica, en inglés *semantic network*) es una base de conocimientos en forma de grafo que representa las relaciones semánticas entre las entidades/conceptos de una ontología. Es decir, es un grafo que consta de nodos, que representan conceptos, y enlaces, que representan relaciones semánticas entre ellos, y puede ser dirigido o no dirigido, según el tipo de relaciones que incluya.

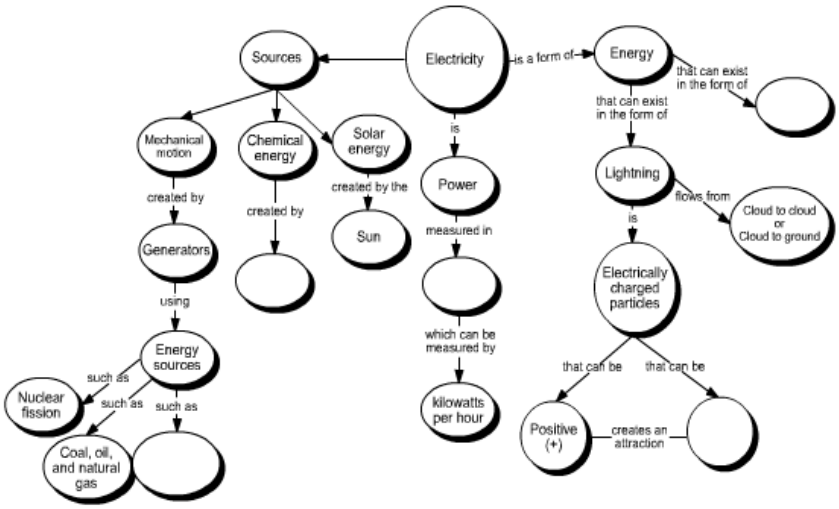


Grafo semántico
Fuente de la imagen: Wikipedia.org

Los grafos semánticos se utilizan para representar información en los sistemas de inteligencia artificial en general y en muchas de las tareas de procesamiento del lenguaje natural, como el análisis sintáctico y la desambiguación semántica. Permiten representar de forma sencilla los datos complejos que modelan el mundo real.

Gracias a este grafo se puede representar toda la información significativa de un escenario de manera estructurada, lo que permitirá: descubrir conocimiento nuevo, enriquecer la información actual, utilizarse de base de conocimiento para responder a preguntas, etc.

En contextos no técnicos, a veces se denominan mapas conceptuales (o diagramas conceptuales), y se usan para representar información o ideas conectadas entre sí.



Ejemplo de mapa conceptual
Fuente de la imagen: Wikipedia.org

La representación gráfica de los grafos semánticos mostrada en las figuras anteriores no es apta para ser manipulada y utilizada por los sistemas informáticos, sino que, en su lugar, se utilizan formatos más apropiados basados en texto.

El W3C (World Wide Web Consortium) ha definido una familia de estándares específicos para la representación e intercambio de información de grafos semánticos en forma legible por la máquina.

Entre ellos, RDF (Resource Description Framework), estándar del W3C desde 2004 (RDF 1.0), plantea un enfoque de modelado conceptual clásico similar a los diagramas entidad-relación, que se basa en la idea de declarar los valores de los recursos con triplas (listas de tres elementos) de la forma objeto-atributo-valor (para este objeto, este atributo toma este valor). RDF define un modelo de datos que se adapta de forma óptima a la información almacenada en el grafo semántico.

Para representar en formato textual esta información semántica, el W3C ha definido también diferentes formatos estándar para facilitar la interoperabilidad, como RDF/XML (ficheros con extensión .rdf), N-Triples (extensión .nt), Turtle (ficheros con extensión .ttl) o JSON-LD (extensión .jsonld).

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://en.wikipedia.org/wiki/Tony_Benn">
  <dc:title>Tony Benn</dc:title>
  <dc:publisher>Wikipedia</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

Ejemplo de RDF
Fuente de la imagen: Wikipedia.org

```
<http://example.org/person/Mark_Twain>
   <http://example.org/relation/author>
   <http://example.org/books/Huckleberry_Finn> .
```

Ejemplo de Turtle
Fuente de la imagen: Wikipedia.org

Para profundizar más sobre estos aspectos, puedes pinchar en los siguientes enlaces:

- Estándar RDF: https://www.w3.org/RDF/
- Notation3 o N3: https://en.wikipedia.org/wiki/Notation3
- N-Triples: https://en.wikipedia.org/wiki/N-Triples
- Turtle: https://en.wikipedia.org/wiki/Turtle_(syntax)
- JSON-LD: https://en.wikipedia.org/wiki/JSON-LD

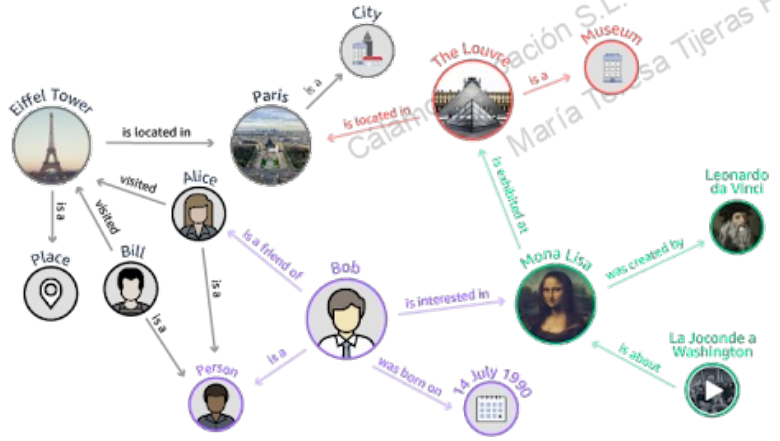
Además de estos estándares, el W3C también ha definido un lenguaje estándar de consulta de información estructurada basada en el modelo RDF, llamado SPARQL (SPARQL Protocol and RDF Query Language), con la misma idea que el lenguaje estándar SQL (Structured Query Language) para consulta de bases de datos relacionales.

People who were born in Berlin before 1900

```
SELECT ?name ?birth ?death ?person WHERE {
   ?person dbo:birthPlace :Berlin .
   ?person dbo:birthDate ?birth .
   ?person foaf:name ?name .
   ?person dbo:deathDate ?death .
   FILTER (?birth < "1900-01-01"^^xsd:date) .
} ORDER BY ?name
```

Ejemplo de consulta con SPARQL

Todos los sistemas que se basan en grafos semánticos utilizan el modelo RDF y el lenguaje SPARQL para consulta. Una base de datos orientada a grafos semánticos muy popular es Virtuoso y también hay un servicio de Amazon Web Services llamado Amazon Neptune.

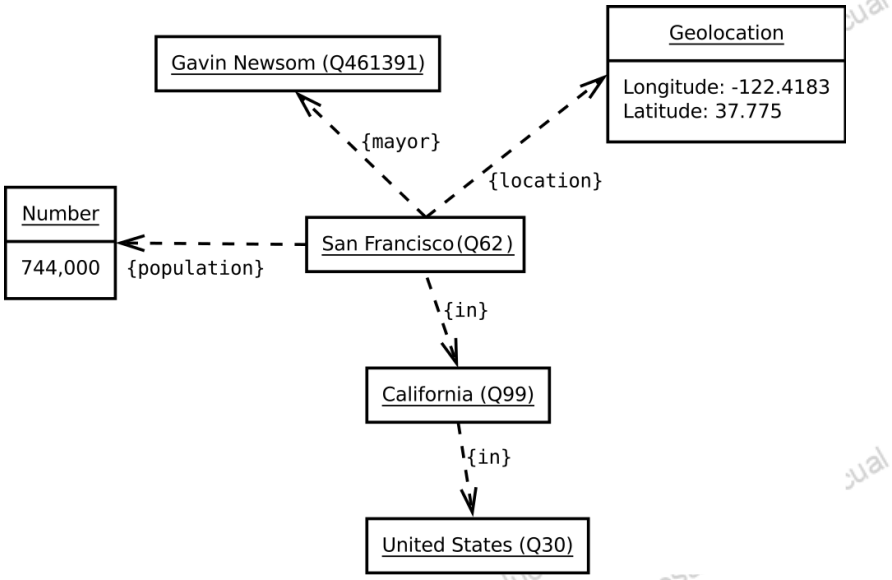


Amazon Neptune
Fuente de la imagen: Amazon.com

Como es bien sabido, Wikipedia es una enciclopedia web multilingüe de contenido libre basado en un modelo de edición abierta que se ha convertido en el mayor proyecto de recopilación de conocimiento de la historia de la humanidad.

Wikidata consiste en una base de conocimiento libre y abierta que actúa como almacenamiento central para los datos estructurados de Wikipedia y otros proyectos hermanos como Wikivoyage, Wiktionary, Wikisource y otros.

Wikidata es un grafo semántico de los contenidos estructurados de Wikipedia.



Fragmento de grafo de Wikidata
Fuente de la imagen: Wikidata.org

Google y otros sistemas de búsqueda se basan en la información de Wikidata para dar respuestas específicas a consultas de los usuarios.

Por ejemplo, es posible preguntar en Google:

> ¿Cuál es la profundidad máxima del mar Mediterráneo?

Y el buscador responde que es 5.267 metros, como se muestra en la figura siguiente.

Resultado de la consulta en Google

Esto es posible porque la página de Wikipedia sobre el mar Mediterráneo incluye esta información como parte de la información estructurada de la ficha de atributos (llamada *infobox* en la jerga de Wikipedia), que se muestra en la parte derecha de la pantalla.

Mar Mediterráneo

Coordenadas: 38°N 17°E (mapa)

«Mediterráneo» redirige aquí. Para otras acepciones, véase Mediterráneo (desambiguación).

El **Mediterráneo** es un mar continental que conecta con el océano Atlántico a través del estrecho de Gibraltar. Rodeado por Europa, África y Asia, fue testigo de la evolución de varias civilizaciones antiguas como los egipcios, los fenicios, hebreos, griegos, cartagineses y romanos. Con aproximadamente 2,5 millones de km² y 3860 km de longitud, es el segundo mar interior más grande del mundo, después del Caribe.[1] Se trata de un mar relativamente hondo, con una profundidad media de 1370 metros, siendo su punto más profundo la fosa de Calipso, al oeste de Grecia. Sus aguas, que bañan las tres grandes penínsulas del sur de Europa (ibérica, itálica y balcánica) y una de Asia (Anatolia), se comunican, además de con el Atlántico por el estrecho de Gibraltar, con el mar Negro por los estrechos del Bósforo y de los Dardanelos y con el mar Rojo por el canal de Suez.[2]

Se trata de un mar cálido con un clima característico homónimo y un estilo culinario propio en toda su cuenca. Estos factores atraen a millones de turistas de otros puntos del mundo. Por otro lado, es el mar con las tasas más elevadas de hidrocarburos y contaminación del mundo.[3]

Índice [ocultar]

1 Nombre
2 Situación
2.1 Límites
2.2 Subdivisiones
2.3 Islas de mayor tamaño
2.4 Clima
2.4.1 Temperatura del mar
3 Historia
4 Naturaleza
4.1 Geología
4.2 Oceanografía
4.3 Flora

Mar Mediterráneo

Océano o mar de la IHO (n.º id.: 28)

El Mediterráneo y sus subdivisiones.

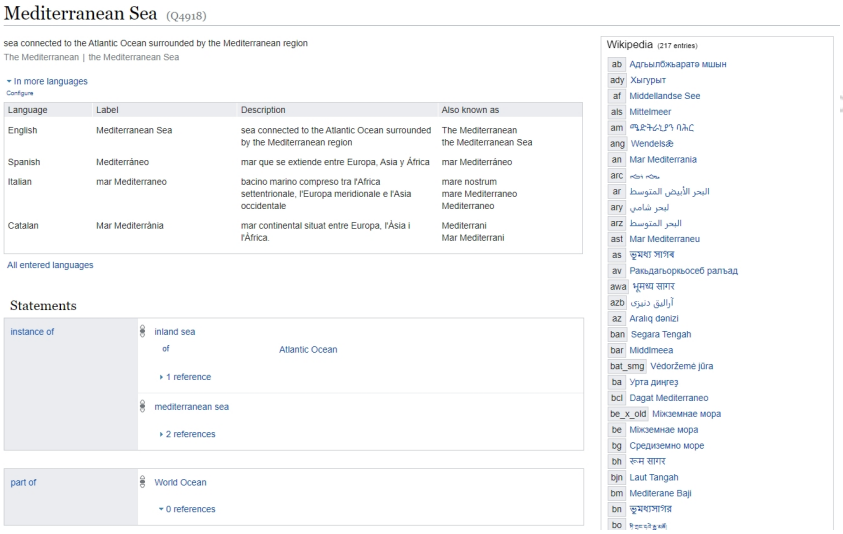
Ubicación geográfica	
Continente	Europa meridional / Asia Occidental / África del Norte
Océano	Océano Atlántico
Cuenca	cuenca hidrográfica del Mediterráneo
Coordenadas	38°N 17°E
Ubicación administrativa	
País	Albania, Argelia, Bosnia y Herzegovina, Chipre, Croacia, Egipto, Eslovenia, España, Francia, Grecia, Israel, Italia, Líbano, Libia, Malta, Marruecos, Mónaco, Montenegro, Palestina, Reino Unido (Gibraltar, Acrotiri y Dhekelia), Siria, Túnez y Turquía
Accidentes geográficos	

Captura de pantalla de la página de Wikipedia sobre mar Mediterráneo

Volumen	3 735 000 km³
Longitud de costa	46 000 km
Profundidad	Media: 1430 m Máxima: 5267 m (Calypso 36°34' N, 21°08' E, GR)
Altitud	0

Detalle de la información sobre profundidad

Esta entidad (mar Mediterráneo) corresponde a la entidad de Wikidata con identificador Q4918 (se puede ver en la página de Wikipedia, en el enlace del menú lateral izquierdo “Elemento de Wikidata”), accesible en la URL: https://www.wikidata.org/wiki/Q4918



Mediterranean Sea (Q4918)

sea connected to the Atlantic Ocean surrounded by the Mediterranean region
The Mediterranean | the Mediterranean Sea

▾ In more languages
Configure

Language	Label	Description	Also known as
English	Mediterranean Sea	sea connected to the Atlantic Ocean surrounded by the Mediterranean region	The Mediterranean the Mediterranean Sea
Spanish	Mediterráneo	mar que se extiende entre Europa, Asia y África	mar Mediterráneo
Italian	mar Mediterraneo	bacino marino compreso tra l'Africa settentrionale, l'Europa meridionale e l'Asia occidentale	mare nostrum mare Mediterraneo Mediterraneo
Catalan	Mar Mediterrània	mar continental situat entre Europa, l'Àsia i l'Àfrica.	Mediterrani Mar Mediterrani

All entered languages

Statements

instance of	inland sea of Atlantic Ocean ▸ 1 reference mediterranean sea ▸ 2 references
part of	World Ocean ▾ 0 references

Wikipedia (217 entries)

ab	Адлыпбжьаратә мшын
ady	Хыгурыт
af	Middellandse See
als	Mittelmeer
am	ሜዲትራኒያን ባሕር
ang	Wendelsæ
an	Mar Mediterrania
arc	ܝܡܐ ܪܒܐ
ar	البحر الأبيض المتوسط
ary	لبحر شامي
arz	البحر المتوسط
ast	Mar Mediterraneu
as	ভূমধ্য সাগৰ
av	Ракьдагьоркьосеб ралъад
awa	भूमध्य सागर
azb	آرالیق دنیزی
az	Aralıq dənizi
ban	Segara Tengah
bar	Middlmeea
bat_smg	Vėdoržemė jūra
ba	Урта диңгеҙ
bcl	Dagat Mediterraneo
be_x_old	Міжземнае мора
be	Міжземнае мора
bg	Средиземно море
bh	रूम सागर
bjn	Laut Tangah
bm	Mediterane Baji
bn	ভূমধ্যসাগর
bo	ཨུར་ཏི་རྒྱམ་མཚོ

Información de Wikidata sobre el mar Mediterráneo

Si en vez de a la dirección anterior, se accede a: https://www.wikidata.org/wiki/Special:EntityData/Q4918.json

Se obtiene la información en formato JSON. También es posible obtenerla en otros formatos cambiando la extensión en la URL: .ttl (formato Turtle), .nt (formato N-Triples) y .jsonld (formato JSON-LD).

Fragmento de información en formato JSON-LD


```
{
    "@graph": [
        {
            "@id": "data:Q4918",
            "@type": "schema:Dataset",
            "about": "wd:Q4918",
            "license": "http://creativecommons.org/publicdomain/zero/1.0/",
            "softwareVersion": "1.0.0",
            "version": 1487545015,
            "dateModified": "2021-08-25T18:53:27Z",
            "statements": 361,
            "sitelinks": 243,
            "identifiers": 59
        },
        {
            "@id": "wd:Q4918",
            "@type": "wikibase:Item"
        },
        {
            "@id": "https://de.wikivoyage.org/wiki/Mittelmeer",
            "@type": "schema:Article",
            "about": "wd:Q4918",
            "inLanguage": "de",
            "isPartOf": "https://de.wikivoyage.org/",
            "name": {
                "@language": "de",
                "@value": "Mittelmeer"
            }
        },
        {
            "@id": "https://de.wikivoyage.org/",
            "wikiGroup": "wikivoyage"
        },
        {
            "@id": "https://en.wikivoyage.org/wiki/Mediterranean_Sea",
            "@type": "schema:Article",
            "about": "wd:Q4918",
            "inLanguage": "en",
            "isPartOf": "https://en.wikivoyage.org/",
            "name": {
                "@language": "en",
                "@value": "Mediterranean Sea"
            }
        },
        {
            "@id": "https://en.wikivoyage.org/",
            "wikiGroup": "wikivoyage"
        },
        {
            "@id": "https://commons.wikimedia.org/wiki/Mediterranean_Sea",
            "@type": "schema:Article",
            "about": "wd:Q4918",
            "inLanguage": "en",
            "isPartOf": "https://commons.wikimedia.org/",
            "name": {
                "@language": "en",
                "@value": "Mediterranean Sea"
            }
        },
```


Con esta información disponible es posible devolver la respuesta a la pregunta de la profundidad máxima.

1

El primer paso es identificar que el usuario está haciendo una pregunta, por ejemplo, al detectar la palabra clave “cuál”.

2

El segundo paso es identificar que se pregunta por el “mar Mediterráneo”. En este caso, mediante un sistema de reconocimiento de entidades enlazadas con las entidades de Wikidata (en la unidad 2 se presentó la tarea de *entity linking*).

3

Luego es necesario identificar por qué propiedad (o atributo) está preguntando el usuario. Con un diccionario de nombres en diferentes idiomas de los atributos o las relaciones del grafo semántico, se obtiene con relativa facilidad que “profundidad máxima” corresponde con el atributo de Wikidata llamado “vertical depth - maximum”.

4

Por último, solo hay que presentar al usuario el valor de dicho atributo para esa entidad, que en este caso es 5.267 metros.

3. Tarea de extracción de información

La extracción de información elaborada, entendida como se ha definido anteriormente como la conversión del texto de entrada en un grafo semántico, resulta una tarea compleja que hay que adaptar a cada escenario en concreto.

Por ejemplo, supongamos el escenario de extracción de información sobre la situación de las empresas de un país, a partir de los informes anuales que todas las empresas deben entregar al Registro Mercantil.

Así ocurre en EE. UU., donde la Comisión del Mercado de Valores (SEC, Securities and Exchange Commision) exige a las empresas el informe 10K (Form 10-K) que contiene un resumen exhaustivo de los resultados financieros de una empresa, en un formato estándar, con una serie de apartados obligatorios.

Toda la información está disponible públicamente en el sitio web de la SEC.

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**

For the fiscal year ended December 31, 2008

OR

☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**

For the transition period from to

Commission File No. 1-2217

The Coca-Cola Company

(Exact name of Registrant as specified in its charter)

DELAWARE (State or other jurisdiction of incorporation or organization)	**58-0628465** (IRS Employer Identification No.)
One Coca-Cola Plaza Atlanta, Georgia (Address of principal executive offices)	**30313** (Zip Code)

Registrant's telephone number, including area code: (404) 676-2121

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Name of each exchange on which registered
COMMON STOCK, $0.25 PAR VALUE	NEW YORK STOCK EXCHANGE

Securities registered pursuant to Section 12(g) of the Act: None

Indicate by check mark if the Registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes ☒ No ☐

Indicate by check mark if the Registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Exchange Act. Yes ☐ No ☒

Indicate by check mark whether the Registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months and (2) has been subject to such filing requirements for the past 90 days. Yes ☒ No ☐

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K is not contained herein, and will not be contained, to the best of Registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K. ☒

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a smaller reporting company. See the definitions of "large accelerated filer," "accelerated filer" and "smaller reporting company" in Rule 12b-2 of the Exchange Act. (Check one):

Large accelerated filer ☒ Accelerated filer ☐ Non-accelerated filer ☐ Smaller reporting company ☐
(Do not check if a smaller reporting company)

Indicate by check mark if the Registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). Yes ☐ No ☒

The aggregate market value of the common equity held by non-affiliates of the Registrant (assuming for these purposes, but without conceding, that all executive officers and Directors are "affiliates" of the Registrant) as of June 27, 2008, the last business day of the Registrant's most recently completed second fiscal quarter, was $113,780,250,547 (based on the closing sale price of the Registrant's Common Stock on that date as reported on the New York Stock Exchange).

The number of shares outstanding of the Registrant's Common Stock as of February 23, 2009 was 2,314,658,162.

DOCUMENTS INCORPORATED BY REFERENCE

Portions of the Company's Proxy Statement for the Annual Meeting of Shareowners to be held on April 22, 2009, are incorporated by reference in Part III.

Ejemplo de Form 10-K

Para saber más sobre *Form 10-K*, puedes pinchar en este enlace.

El objetivo de un sistema de extracción de información elaborada aplicado a este escenario sería, dado el texto de un Form 10-K, extraer toda la información estructurada posible que resulte de interés a un analista para evaluar la empresa en cuestión y poder tomar decisiones de mercado. Por ejemplo: ingresos y gastos totales, cuenta de resultados, clientes más importantes, adquisiciones o fusiones con otras empresas, información de los accionistas de la empresa, etc.

Los sistemas se basan en tecnologías de procesamiento del lenguaje natural, que deben abarcar el nivel léxico, gramatical y semántico del análisis de textos, incluyendo, por supuesto, todos los aspectos de desambiguación y análisis del contexto.

Esta tarea se puede considerar la evolución del reconocimiento de entidades y, en cierta manera, se aborda con técnicas similares.

3.1. Modelos de reglas de extracción

Por un lado, están las técnicas basadas en modelos de reglas de extracción. A diferencia de las reglas de reconocimiento de entidades (descritas en la unidad 2) y las reglas de clasificación de textos (descritas en la unidad 3), que se orientan principalmente a la detección de contextos (detección de entidades con nombre o condiciones por las que clasificar un texto en una categoría), las reglas de extracción de información se centran tanto en la detección como en la extracción de información.

Detección de entidades

Las reglas de extracción incluyen expresiones y operadores que permiten encontrar patrones avanzados en todos los niveles de análisis del lenguaje, con palabras, formas, estructuras sintácticas, etc. dentro del texto.

En el anterior ejemplo de extracción de información en Form 10-K, para identificar los ingresos y ventas totales, hay que extraer cantidades numéricas que indiquen cifras en dólares (USD); para identificar a los clientes más importantes, hay que extraer nombres de empresas; etc.

Extracción de nodos

Este paso consiste en decidir cuáles de las entidades detectadas en la fase anterior van a añadirse como nodos al grafo semántico.

No todas las entidades que salen del reconocimiento de entidades son aptas, sino que, en este caso, es muy importante realizar una desambiguación adecuada de las entidades detectadas, según el escenario de trabajo.

Es decir, no sirve con detectar expresiones en dólares (por ejemplo, asignando un tipo semántico MONEY) o nombres de empresa (tipo semántico COMPANY), como haría un detector de entidades, sino que es importante "darles un apellido", identificar su rol dentro del escenario, esto es, asegurarse de que esas cifras son realmente ingresos o gastos, o que las compañías detectadas son clientes de la empresa en cuestión.

En este ejemplo, la desambiguación (que siguiendo la misma estrategia se haría mediante reglas) consistiría en decidir cuál de esas entidades de tipo MONEY son INCOME o EXPENSE, o cuál de esos nombres de empresas de tipo COMPANY son CUSTOMER.

Una regla posible podría ser algo como "si una entidad de tipo COMPANY está cerca -en el mismo párrafo, o en la misma sección, o a tantas palabras de distancia…- de las palabras "customer" o "client", añadirla al grafo semántico como un nodo de tipo CUSTOMER:

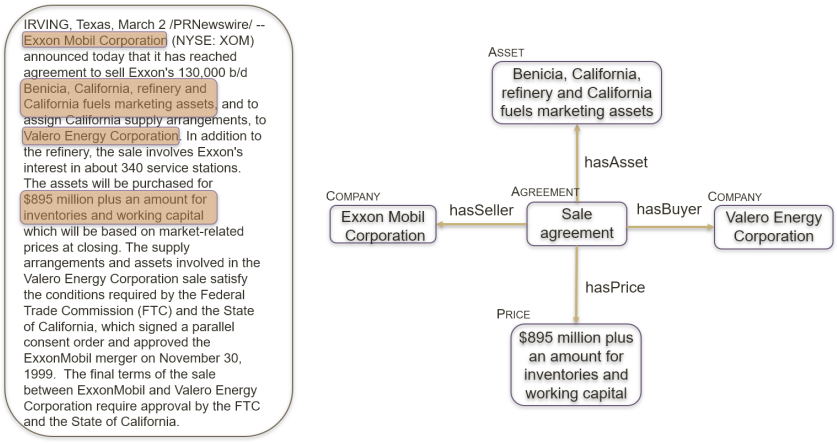
NEAR(S@COMPANY, customer|client) → ENTITY(CUSTOMER)

Extracción de propiedades y relaciones

Además de las entidades, también es necesario (según el escenario) incluir las propiedades (atributos) de los nodos y relaciones entre nodos.

Por ejemplo, para cada cliente detectado anteriormente, puede ser útil añadir como propiedad de la entidad el valor de su cifra de ventas, si está presente en el Form 10-K, o añadir una relación entre ellos si esos clientes forman parte de un mismo grupo de empresas.

La figura siguiente muestra un ejemplo de texto (un artículo) y su representación como grafo semántico que incluye nodos y relaciones entre ellos. Se puede leer que existe un acuerdo de compra (entidad de tipo AGREEMENT) entre las empresas Exxon y Valero (entidades de tipo COMPANY), la primera como vendedor ("hasSeller") y la segunda como comprador ("hasBuyer"), de una refinería (entidad de tipo ASSET, conectada por la relación "hasAsset" con el acuerdo) por 895 millones de dólares (entidad de tipo PRICE, conectada por la relación "hasPrice" con el acuerdo).



Ejemplo de grafo semántico con relaciones

La salida final del sistema consistiría en la lista de nodos, propiedades y relaciones extraídas del texto, representadas como se considere más adecuado en el proyecto, por ejemplo, en un formato estándar como Turtle, o directamente como listas en un fichero CSV.

El empleo de un formato estándar permite su explotación directa por un sistema de base de datos semántica, consultando con SPARQL.

3.2. Técnicas de aprendizaje automático

Otra estrategia, como se estudió en la unidad 2 sobre reconocimiento de entidades, es abordar el problema de detección mediante técnicas de aprendizaje automático.

En principio, serviría cualquiera de los enfoques de aprendizaje automático allí estudiados, en particular, de forma preferente, los basados en *deep learning* por sus mayores capacidades.

El proceso se podría abordar de forma similar al descrito en el apartado anterior (primero la detección de entidades candidatas y luego la incorporación al grafo semántico como nodos, propiedades o relaciones entre nodos), o se podría intentar abordar el problema directamente de forma global, es decir, en un solo paso. El texto se representaría en forma de secuencia y directamente se anotaría sobre cada token si forma parte de un nodo, una propiedad o una relación, indicando además en estos dos últimos casos la entidad o entidades a la(s) que se refiere.

Sin embargo, hoy por hoy, la dificultad de la tarea hace poco factible este enfoque directo y tampoco hay que perder de vista que sería necesario un volumen ingente de datos anotados para realizar el entrenamiento, que no está disponible en la mayoría de los escenarios.

Una estrategia que se ha seguido en ciertos escenarios para anotar masivamente textos de entrenamiento es utilizar la información estructurada de las *infoboxes* de Wikipedia para anotar el propio texto del artículo de Wikipedia.

3.3. Tarea del lingüista

Como sucedía en la tarea de reconocimiento de entidades y otras varias, la principal tarea del lingüista es incorporar al sistema el conocimiento necesario para identificar la información relevante en un determinado escenario de aplicación. El desarrollo de recursos lingüísticos con diccionarios de entidades y/o reglas de extracción es la aportación esencial para la buena marcha del proyecto.

Otra de las tareas que suele estar a cargo del lingüista es la evaluación de los resultados del sistema, utilizando las métricas que se consideren oportunas para el proyecto (que suelen estar basadas en las habituales precisión y cobertura).

Con las conclusiones obtenidas, el lingüista colabora en la definición de la estrategia de optimización del sistema para mejorar los resultados obtenidos, hasta alcanzar el nivel deseado.

RESUMEN

En esta unidad se han presentado los fundamentos y técnicas para abordar la tarea de procesamiento del lenguaje natural dedicadas a la **extracción de información elaborada**. Desde el punto de vista de la salida, el objetivo de la tarea es convertir el texto de entrada en un **grafo semántico** que contenga las entidades mencionadas en el texto (igual que en el reconocimiento de entidades, con su tipo semántico y relevancia) y las relaciones entre ellas (habitualmente incluyendo también un valor de relevancia o confianza).

Un grafo semántico (o red semántica, en inglés *semantic network*) es una **base de conocimientos en forma de grafo** que representa las relaciones semánticas entre las entidades/conceptos de una ontología. Permite representar de forma sencilla los **datos complejos** que modelan el mundo real en los sistemas de inteligencia artificial en general y en muchas de las tareas de procesamiento del lenguaje natural, como el análisis sintáctico y la desambiguación semántica.

El W3C (World Wide Web Consortium) ha definido una familia de estándares específicos para la **representación e intercambio de información de grafos semánticos** en forma legible por la máquina, en concreto, para el modelado conceptual (RDF), la representación textual (RDF/XML, N-Triples, Turtle, JSON-LD) y la explotación mediante un lenguaje de consulta (SPARQL).

Para abordar la tarea de extracción de información elaborada se emplean modelos basados en **reglas de extracción** y las **técnicas de aprendizaje automático**, en particular *deep learning.*

Recursos

Enlaces de Interés

- **Semantic network**
  https://en.wikipedia.org/wiki/Semantic_network
- **Concept map**
  https://en.wikipedia.org/wiki/Concept_map
- **Resource Description Framework**
  https://es.wikipedia.org/wiki/Resource_Description_Framework
- **Turtle (syntax)**
  https://en.wikipedia.org/wiki/Turtle_(syntax)
- **Estándar RDF**
  https://www.w3.org/RDF/
- **Notation3 o N3**
  https://en.wikipedia.org/wiki/Notation3
- **N-Triples**
  https://en.wikipedia.org/wiki/N-Triples
- **JSON-LD**
  https://en.wikipedia.org/wiki/JSON-LD
- **Virtuoso**
  https://virtuoso.openlinksw.com/
- **Amazon Neptune**
  https://aws.amazon.com/neptune
- **Amazon Neptune**
  https://aws.amazon.com/es/neptune/
- **Wikidata**
  https://www.wikidata.org/
- **Wikidata: Introduction**
  https://www.wikidata.org/wiki/Wikidata:Introduction
- **Mar Mediterráneo**
  https://es.wikipedia.org/wiki/Mar_Mediterr%C3%A1neo
- **Mediterranean Sea**
  https://www.wikidata.org/wiki/Q4918
- **SEC**
  https://www.sec.gov/edgar/searchedgar/companysearch.html
- **Form 10-K**
  https://en.wikipedia.org/wiki/Form_10-K