



PROCESAMIENTO DE LENGUAJE NATURAL

DRA CECILIA REYES PEÑA

CREYES@UPMH.EDU.MX



La PNL se refiere a un conjunto de técnicas que implican la aplicación de métodos estadísticos, con o sin conocimientos de lingüística, para comprender textos con el fin de resolver tareas del mundo real. Esta "comprensión" del texto se obtiene principalmente transformando los textos en representaciones informáticas utilizables, que son estructuras combinatorias discretas o continuas como vectores o tensores, grafos y árboles.

WEB SEMÁNTICA

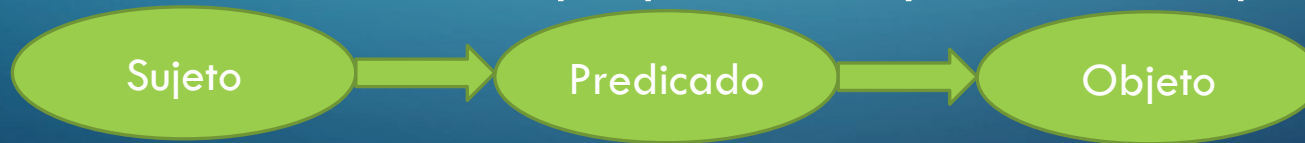
- Web **semántica** es un conjunto de actividades desarrolladas por el World Wide Web Consortium (<https://www.w3.org>) que buscan la creación de tecnologías para publicar **datos legibles por aplicaciones informáticas**.
- Se basa en la idea de añadir **metadatos semánticos y ontológicos** a la *World Wide Web*, *los cuales* describen el contenido, el significado y la relación de los datos
- Los metadatos deben proporcionarse **de manera formal**, para que así sea posible evaluarlas automáticamente por máquinas de procesamiento.

- XML significa lenguaje de marcado extensible
- XML es un lenguaje de marcado muy parecido a HTML
- XML fue diseñado para almacenar y transportar datos
- XML fue diseñado para ser autodescriptivo
- XML es una recomendación W3C

```
<note>  
  <to>Tove</to>  
  <from>Jani</from>  
  <heading>Reminder</heading>  
  <body>Don't forget me this weekend!</body>  
</note>
```

RDF

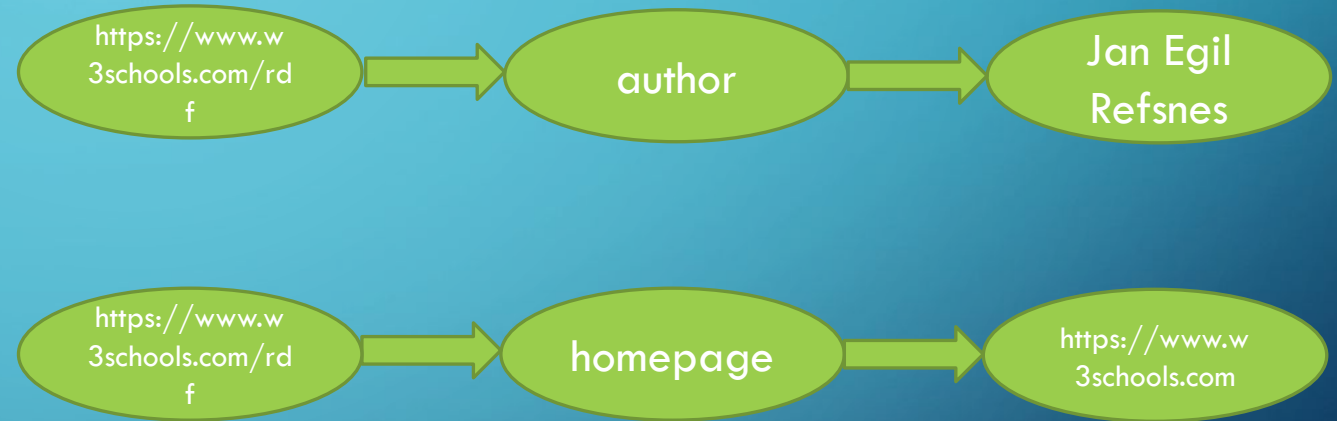
- RDF significa **R**esource **D**escription **F**ramework
- RDF es un marco para describir recursos en la web
- RDF está escrito en XML
- RDF utiliza identificadores web (URI) para identificar recursos.
- RDF describe recursos con propiedades y valores de propiedad.



RDF

```
<?xml version="1.0"?>

<RDF>
  <Description about="https://www.w3schools.com/rdf">
    <author>Jan Egil Refsnes</author>
    <homepage>https://www.w3schools.com</homepage>
  </Description>
</RDF>
```



RDFS

- RDF Schema proporciona el marco para describir clases y propiedades específicas de la aplicación.
- Las clases en el esquema RDF son muy parecidas a las clases en los lenguajes de programación orientados a objetos. Esto permite que los recursos se definan como instancias de clases y subclases de clases.

RDFS

```
<?xml version="1.0"?>

<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xml:base="http://www.animals.fake/animals#">

  <rdf:Description rdf:ID="animal">
    <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  </rdf:Description>

  <rdf:Description rdf:ID="horse">
    <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
    <rdfs:subClassOf rdf:resource="#animal"/>
  </rdf:Description>

</rdf:RDF>
```

CLASSES RDF Y RDFS

Element	Class of	Subclass of
rdfs:Class	All classes	
rdfs:Datatype	Data types	Class
rdfs:Resource	All resources	Class
rdfs:Container	Containers	Resource
rdfs:Literal	Literal values (text and numbers)	Resource
rdf:List	Lists	Resource
rdf:Property	Properties	Resource
rdf:Statement	Statements	Resource
rdf:Alt	Containers of alternatives	Container
rdf:Bag	Unordered containers	Container
rdf:Seq	Ordered containers	Container
rdfs:ContainerMembershipProperty	Container membership properties	Property
rdf:XMLLiteral	XML literal values	Literal

10/27/2022

PROPIEDADES RDF Y RDFS

Element	Domain	Range	Description
rdfs:domain	Property	Class	The domain of the resource
rdfs:range	Property	Class	The range of the resource
rdfs:subPropertyOf	Property	Property	The property is a sub property of a property
rdfs:subClassOf	Class	Class	The resource is a subclass of a class
rdfs:comment	Resource	Literal	The human readable description of the resource
rdfs:label	Resource	Literal	The human readable label (name) of the resource
rdfs:isDefinedBy	Resource	Resource	The definition of the resource
rdfs:seeAlso	Resource	Resource	The additional information about the resource
rdfs:member	Resource	Resource	The member of the resource
rdf:first	List	Resource	
rdf:rest	List	List	

PROPIEDADES RDF Y RDFS

Element	Domain	Range	Description
rdf:rest	List	List	
rdf:subject	Statement	Resource	The subject of the resource in an RDF Statement
rdf:predicate	Statement	Resource	The predicate of the resource in an RDF Statement
rdf:object	Statement	Resource	The object of the resource in an RDF Statement
rdf:value	Resource	Resource	The property used for values
rdf:type	Resource	Class	The resource is an instance of a class

ATRIBUTOS RDF

Attribute	Description
rdf:about	Defines the resource being described
rdf:Description	Container for the description of a resource
rdf:resource	Defines a resource to identify a property
rdf:datatype	Defines the data type of an element
rdf:ID	Defines the ID of an element
rdf:li	Defines a list
rdf:_n	Defines a node
rdf:nodeID	Defines the ID of an element node
rdf:parseType	Defines how an element should be parsed
rdf:RDF	The root of an RDF document
xml:base	Defines the XML base
xml:lang	Defines the language of the element content

10/27/2022

OWL

- El lenguaje de ontología web OWL tiene como objetivo proporcionar un lenguaje que se pueda usar para describir las clases y las relaciones entre ellas que son inherentes a los documentos y aplicaciones web.
- OWL es usado para:
 - formalizar un dominio definiendo clases y propiedades de esas clases,
 - definir individuos y afirmar propiedades sobre ellos, y
 - razonar sobre estas clases e individuos en la medida permitida por la semántica formal del lenguaje OWL.

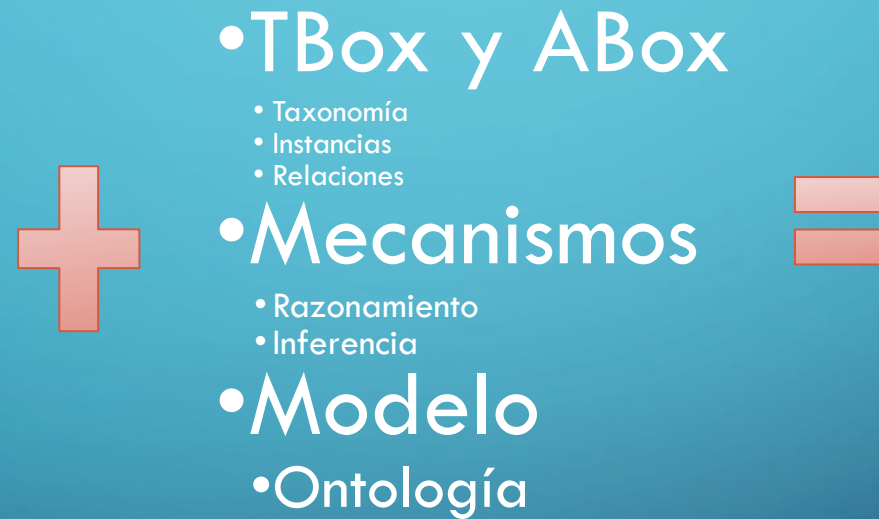
ONTOLOGÍA

- “Una especificación explícita de una conceptualización compartida” (Gruber, 1993)
- Representa un dominio por medio de un vocabulario legible (instancias, propiedades, axiomas) y de una estructura (Hlomani, 2014; Tello, 2001; Luna et al., 2012; Martínez-Gil et al., 2008).
- Basadas en lógicas descriptivas poseen dos mecanismos de razonamiento (Nardi & Brachman, 2003):
 - ❖ Satisfactibilidad del modelo
 - ❖ Clasificación de instancias

ELEMENTOS DE LAS ONTOLOGÍAS

- **Instancias (Objetos):** es la representación de los principales objetos del dominio a representar.
- **Propiedades de tipos de Dato:** indica la información acerca del dominio en un formato maquina (valores enteros, flotante, texto, entre otros).
- **Conceptos (Clases):** son las ideas principales acerca del dominio que serán formalizadas. Visto desde la lógica, los conceptos pueden ser descritos usando propiedades específicas que deberán ser satisfechas por ellos mismos.
- **Propiedades de objeto:** son ligas entre los conceptos con la finalidad de representar la estructura de la ontología lo mas fiel posible a la realidad, ya sea que dicha estructura sea o no taxonómica.
- **Axiomas:** los axiomas son las restricciones, reglas y definiciones de correspondencias lógicas que deben cumplirse en las relaciones entre los elementos de la ontología. También pueden ser vistos como la mas pequeña unidad de conocimiento dentro de una ontología.

ONTOLOGÍAS DESDE LA LÓGICA



TBOX

- El TBox (*Terminology Box*) es la parte de la representación de **conocimiento intencional** o general del dominio a través de las lógicas descriptivas. La terminología tiene forma de taxonomía e incluye operaciones y declaraciones que describen las propiedades de los conceptos.
- Dentro del TBox se realizan las siguientes tareas del razonamiento:
 - **Subsunción:** indica que un subconjunto A está contenido en un conjunto B ($A \sqsubseteq B$)
 - **Satisfactibilidad:** indica que existe un conjunto no vacío que satisface de forma verdadera una definición.
 - **Consistencia:** verifica que no existan contradicciones entre la terminología
 - **Equivalencia (\equiv):** donde dos definiciones son equivalentes si obtienen los mismos valores de verdad para el mismo conjunto.

ABOX

- El ABox (*Assertion Box*) es la parte de la representación que contiene el **conocimiento extendido** por medio de afirmaciones (conocimiento afirmativo), el cual especifica a los individuos del dominio.
- Las tareas básicas de razonamiento que se lleva a cabo en ABox son:
 - **Comprobación de instancias (*Instance Checking*):** verifica si un individuo dado es una instancia de (pertenece a) un concepto específico
 - **Consistencia de la base de conocimiento:** verifica si cada concepto en la base de conocimiento admite por lo menos un individuo
 - **Realización:** encuentra el concepto más específico de un objeto individual, el cual es una instancia de
 - **Recuperación:** que encuentra a los individuos en la base de conocimiento que son^{10/27/2022} instancias de un concepto dado.

MECANISMO DE RAZONAMIENTO

- El mecanismo de razonamiento tiene como propósito **generar conocimiento** basado en los propios **principios de diseño** de la ontología. Para esto se realizan las siguientes tareas:

- **Verificación de consistencia:** contradicciones en las definiciones de todo el TBox.

Pertenencia de la clase: si una clase A tiene como subclase a una clase B y a su vez una instancia x pertenece a la clase B , entonces la instancia x también tiene pertenencia a la clase A .

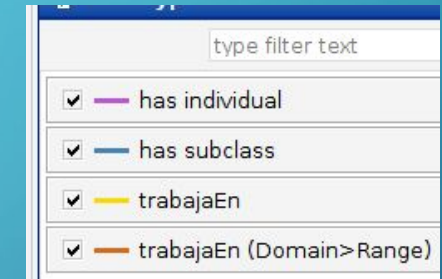
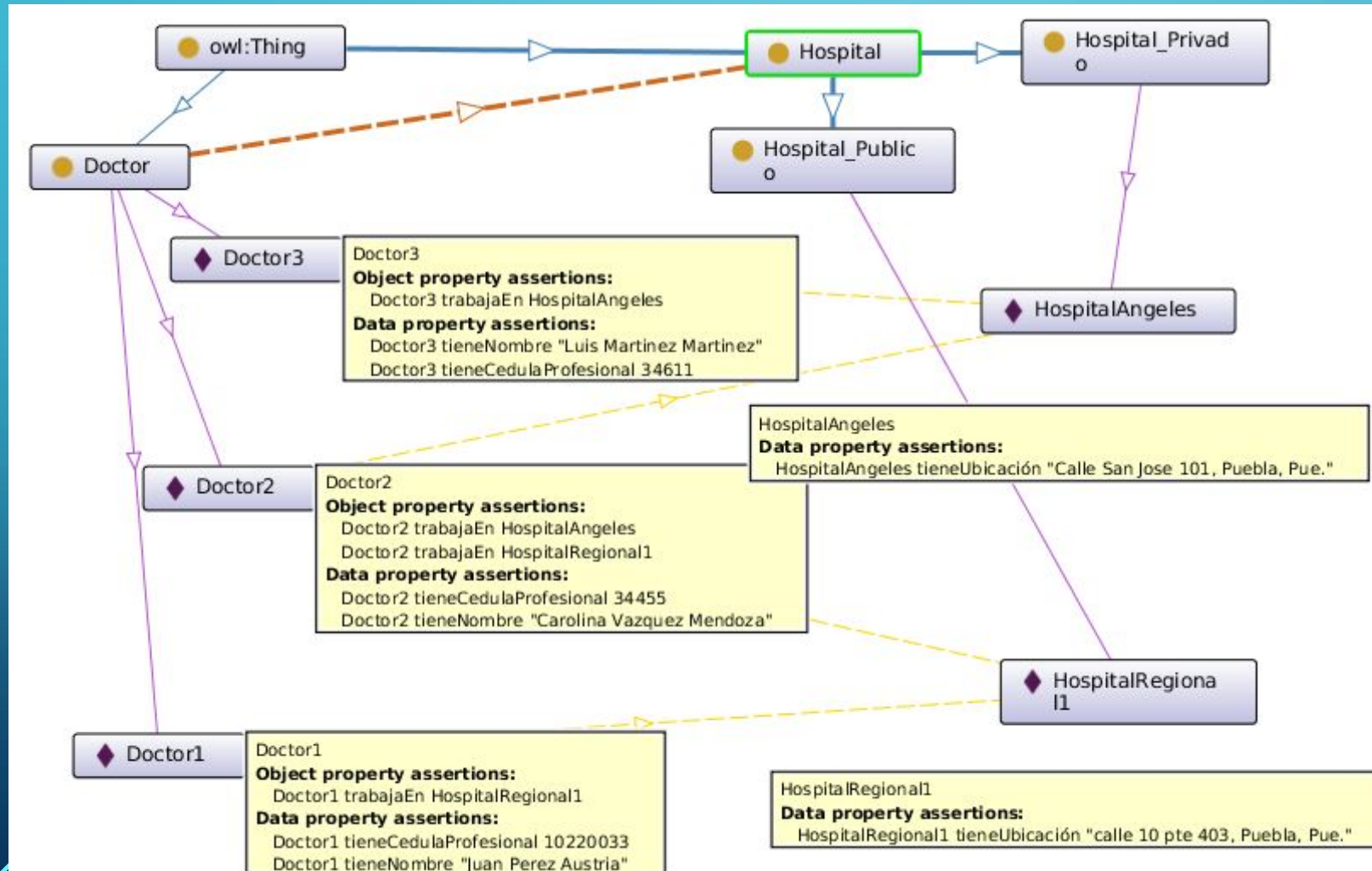
Equivalencia de clases: si una clase A es equivalente a una clase B , a su vez la clase B es equivalente a una clase C , entonces A y C son equivalentes.

Clasificación: se define dentro del TBox que ciertos pares **propiedad-valor** son condiciones suficientes para pertenecer a una clase A , entonces si una instancia x satisface tales condiciones, entonces x debe ser una instancia de la clase A .

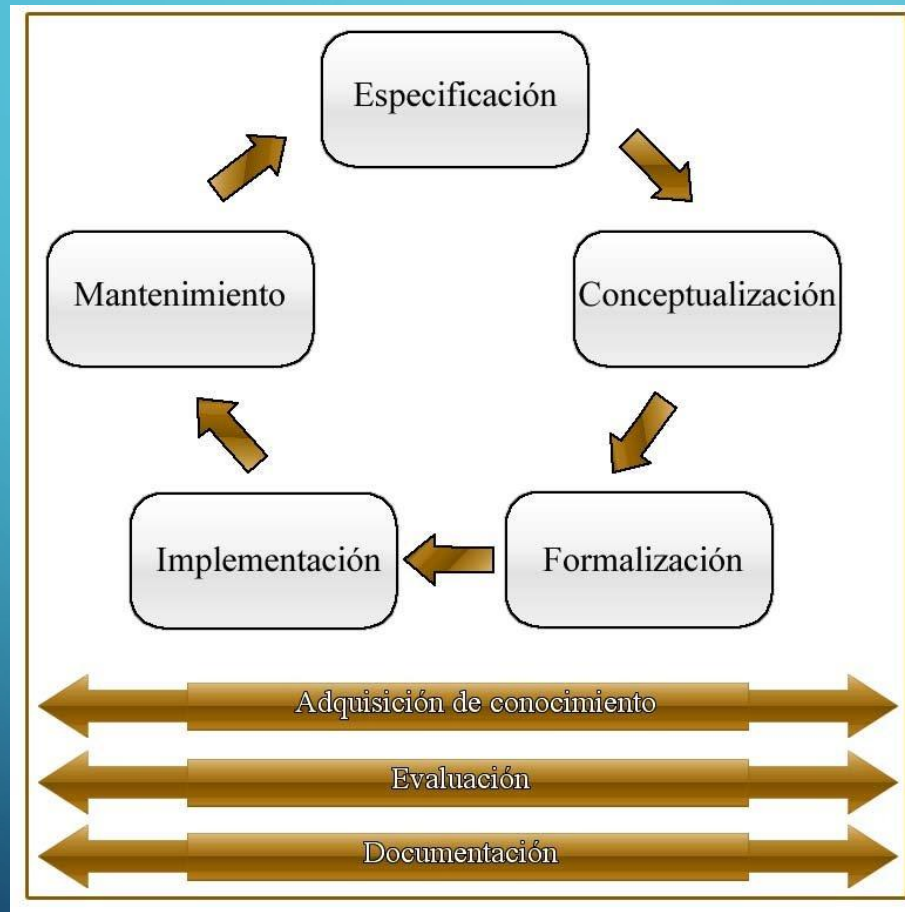
MECANISMO DE INFERENCIA

- El mecanismo de inferencia tiene la tarea de hacer inferencias al **decidir cuales reglas** definidas en el diseño **satisfacen a las instancias del ABox**, asignarles preferencias y ejecutar la de mayor prioridad, además de proveer el razonamiento necesario para formular conclusiones acerca de la información representada mediante lógicas.
- Para formular las conclusiones, el mecanismo de inferencia tiene un proceso recursivo a partir de tres etapas:
 - Durante el estado de **Match**, los datos acerca de las instancias se comparan con las reglas lógicas definidas en el TBox. Todas las reglas que se pueden ejecutar se almacenan en un conjunto denominado *conjunto de conflictos*.
 - Después, se selecciona una regla del conjunto de conflictos (**Select**) según algunos criterios, principalmente la precedencia, con la finalidad de que dicha regla será la primera en ejecutarse (o única en ejecutarse, dado que no exista otra regla que se pueda satisfacer).
 - Finalmente, los resultados de ejecución de la regla se almacenan en el ABox con los datos iniciales ^{10/27/2022} (**Execute**).

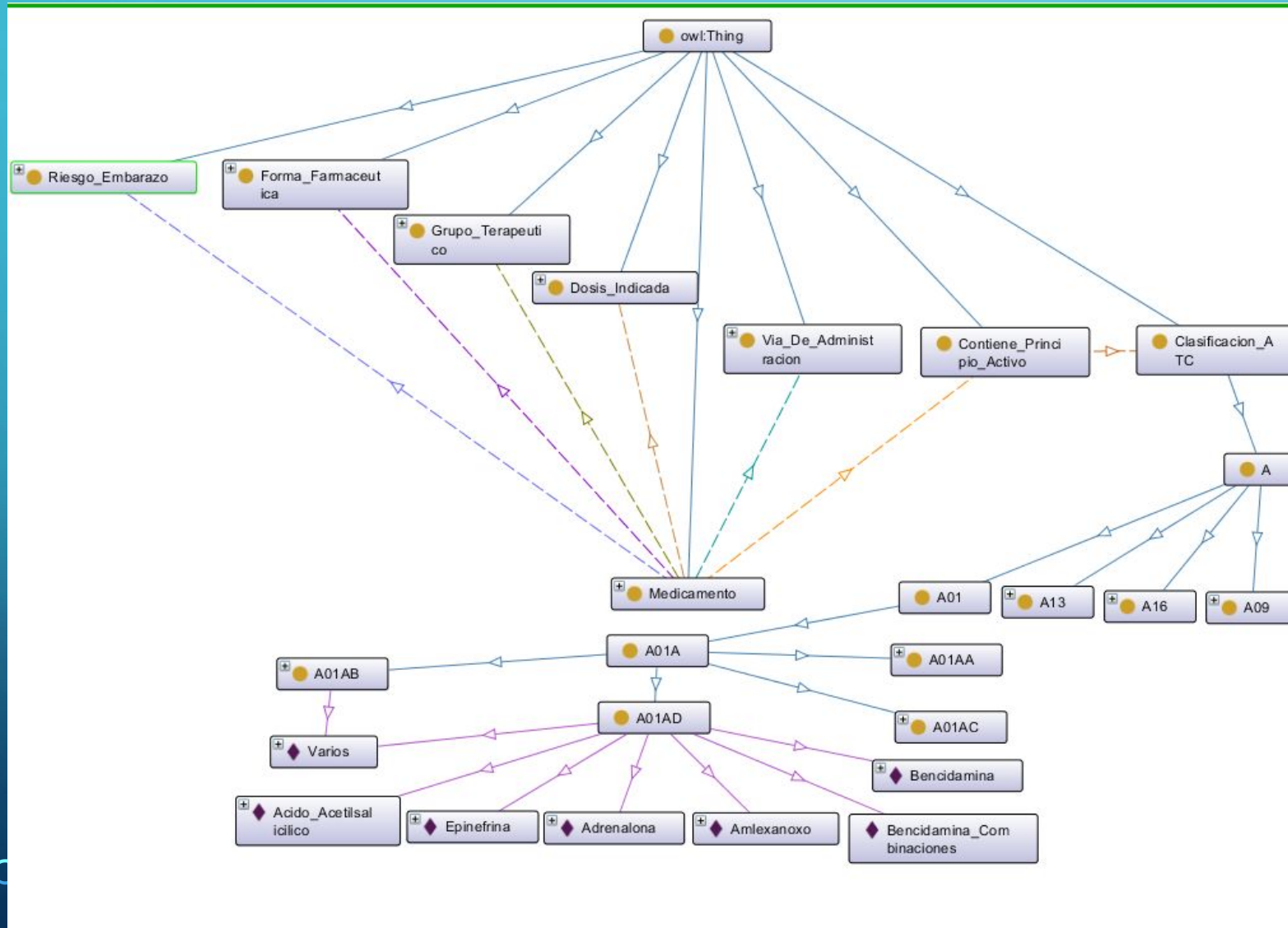
ONTOLOGÍA HOSPITAL



METODOLOGÍA DE DISEÑO



ONTOLOGÍA MEDICAMENTOS-ATC



- ☒ — tieneFormulaFarmaceutica (Domain>Range)
- ☒ — tieneGrupoTerapeutico (Domain>Range)
- ☒ — tienePrincipioActivoPorPorcion (Domain>Range)
- ☒ — tieneDosisIndicada (Domain>Range)
- ☒ — tieneViaDeAdministracion (Domain>Range)
- ☒ — provocaRiesgoDuranteElEmbarazo (Domain>Range)
- ☒ — tienePrincipioActivo (Domain>Range)
- ☒ — tienePrincipioActivoPorPorcion (Domain>Range)

PROPIEDADES DE OBJETO ONTOLOGÍA MEDICAMENTO-ATC

Propiedad de objeto	Dominio	Rango	Cardinalidad
provocaRiesgoDuranteElEmbarazo	Medicamento	Riesgo_Embarazo	1:1
tieneDosisIndicada	Medicamento	Dosis_Indicada	1:N
tieneFormulaFarmaceutica	Medicamento	Forma_Farmaceutica	1:N
tieneGrupoFarmaceutico	Medicamento	Grupo_Terapeutico	1:1
tienePrincipioActivo	Medicamento	Principio_Activo	1:N
tieneViaDeAdministracion	Medicamento	Via_De_Administracion	1:N

PROPIEDADES DE DATO ONTOLOGÍA MEDICAMENTO-ATC

Propiedad de tipo de Dato	Dominio	Rango
tieneActivo	Principio_Activo	string
tieneCantidadDeActivo	Principio_Activo	float
tieneCantidadMaxima	Dosis_Indicada	float
tieneCantidadMinima	Dosis_Indicada	float
tieneContenidoPorEnvase	Medicamento	integer
tieneDescripcion	Riesgo_Embarazo	string
tieneFrecuenciaEnHoras	Dosis_Indicada	float
tieneFrecuenciaEnHorasMaxima	Dosis_Indicada	float
tieneFrecuenciaEnHorasMinima	Dosis_Indicada	float
tieneIndicacionAdicional	Dosis_Indicada	string
tieneIndicacionParaPadecimiento	Dosis_Indicada	string
tieneLimiteInferiorDeEdad	Dosis_Indicada_Para_Ninos	integer
tieneLimiteSuperiorDeEdad	Dosis_Indicada_Para_Ninos	integer
tieneMedida	Principio_Activo	string
tieneMedidaDeEdad	Dosis_Indicada_Para_Ninos	string
tieneMedidaIndicada	Dosis_Indicada_Para_Ninos	string
tieneNombre	Medicamento	string

10/27/2022

AXIOMAS ONTOLOGÍA MEDICAMENTO-ATC

Axioma	Entidades involucradas	Descripción
Tableta-Oral	Medicamento, tiene- FormulaFarmaceutica, Tableta, tieneViaDeAd- ministracion y Oral	Medicamento(?x) ^ tieneFormulaFarma- ceutica(?x, Tableta) -> tieneViaDeAdministra- cion (?x, Oral)
Tableta_Soluble-Oral	Medicamento, tiene- FormulaFarmaceutica, Tableta_Soluble, tiene- ViaDeAdministracion y Oral	Medicamento(?x) ^ tie- neFormulaFarmaceuti- ca(?x, Tableta_Soluble) -> tieneViaDeAdmi- nistracion (?x, Oral)
Tableta_Efervescente-Oral	Medicamento, tiene- FormulaFarmaceutica, Tableta_Efervescente, tieneViaDeAdministra- cion y Oral	Medicamento(?x) ^ tieneFormulaFarma- ceutica(?x, Table- ta_Efervescente) -> tieneViaDeAdministra- cion (?x, Oral)
Parche-Transdermica	Medicamento, tiene- FormulaFarmaceutica, Parche, tieneVia- DeAdministracion y Transdermica	Medicamento(?x) ^ tieneFormulaFarma- ceutica(?x, Parche) -> tieneViaDeAdministra- cion (?x, Transdermica)

SPARQL

- es un lenguaje de consulta basado en la sintaxis SQL (*Structured Query Language*) que permite estructurar documentos de tipo RDF y RDFS como grafos etiquetados (*SPARQL Protocol and RDF Query Language*, 2013). Al realizar una consulta, los grafos son analizados y se devuelve un conjunto de tripletas que satisfagan a la consulta realizada

PREGUNTAS DE COMPETENCIA

Pregunta en lenguaje Natural	Traducción DL-Query	Traducción SPARQL
1.-¿Cuales son los medicamentos para niños que tienen vía de administración oral?	(tieneDosisIndicada some Dosis Indicada Para Ninos) and (tieneViaDeAdministracion value Oral)	SELECT ?nombre WHERE{ ?medi med:tieneDosisIndicada ?dosis. ?dosis a med:Dosis Indicada Para Ninos. ?medi med:tieneViaDeAdministracion med:Oral. ?medi med:tieneNombre ?nombre}
2.-¿Cuales medicamentos que tienen vía de administración oral pertenecen al grupo terapéutico anestesia?	(tieneGrupoTerapeutico value Anestesia) and (tieneViaDeAdministracion value Oral)	SELECT ?nombre WHERE{ ?medi med:tieneViaDeAdministracion med:Oral. ?medi med:tieneGrupoTerapeutico med:Anestesia. ?medi med:tieneNombre ?nombre}
3.-¿Cuales son los medicamentos que tienen una cantidad de principio activo por porción mayor a 50mg?	tienePrincipioActivoPorPorcion some ((tieneCantidadDeActivo some xsd:float[>50.0f]) and (tieneMedida value "MG."))	SELECT ?nombre ?cant ?medida WHERE{ ?medi med:tienePrincipioActivoPorPorcion ?nodo. ?medi med:tieneNombre ?nombre. ?nodo med:tieneCantidadDePrincipioActivo ?cant. ?nodo med:tieneMedidaDePrincipioActivo ?medida. Filter(?cant>50.0). Filter(STR(?medida)="MG")}

DL-QUERY

- *Description Logics Query* (DL Query): es un lenguaje de consulta de ontologías que obedece a la sintaxis Manchester y requiere de un agente razonador para conseguir respuestas. Si bien los resultados de las inferencias son fácilmente accesibles desde este lenguaje, es difícil realizar consultas que requieran filtrar datos por medio de valores específicos.

DL-QUERY

¿Cuáles son las dosis recomendadas para niños de los medicamentos que tienen como principio activo el ibuprofeno?

```
(inverse tieneDosisIndicada some (tienePrincipioActivoPorPorcion some (tieneActivo value "IBUPROFENO"))) and Dosis_Indicada_Para_Ninos
```

Execute Add to ontology

Query results

Instances (4 of 4)

- d10
- d11
- d6
- d8

¿Cuáles son las formas farmacéuticas de los medicamentos que tienen como principio activo el Paracetamol?

```
inverse tieneFormulaFarmaceutica some (tienePrincipioActivoPorPorcion some ( tieneActivo value "PARACETAMOL"))
```

Execute Add to ontology

Query results

Instances (4 of 4)

- Solucion_Inyectable
- Solucion_Oral
- Supositorio
- Tableta



EXTRACCIÓN DE RELACIONES

10/27/2022

32

EXTRACCIÓN DE RELACIONES

- Es la tarea de predecir atributos y relaciones para entidades en una oración.
- Por ejemplo, dada una oración "Barack Obama nació en Honolulu, Hawái", un clasificador de relaciones tiene como objetivo predecir la relación "nacioEnLaCiudad".

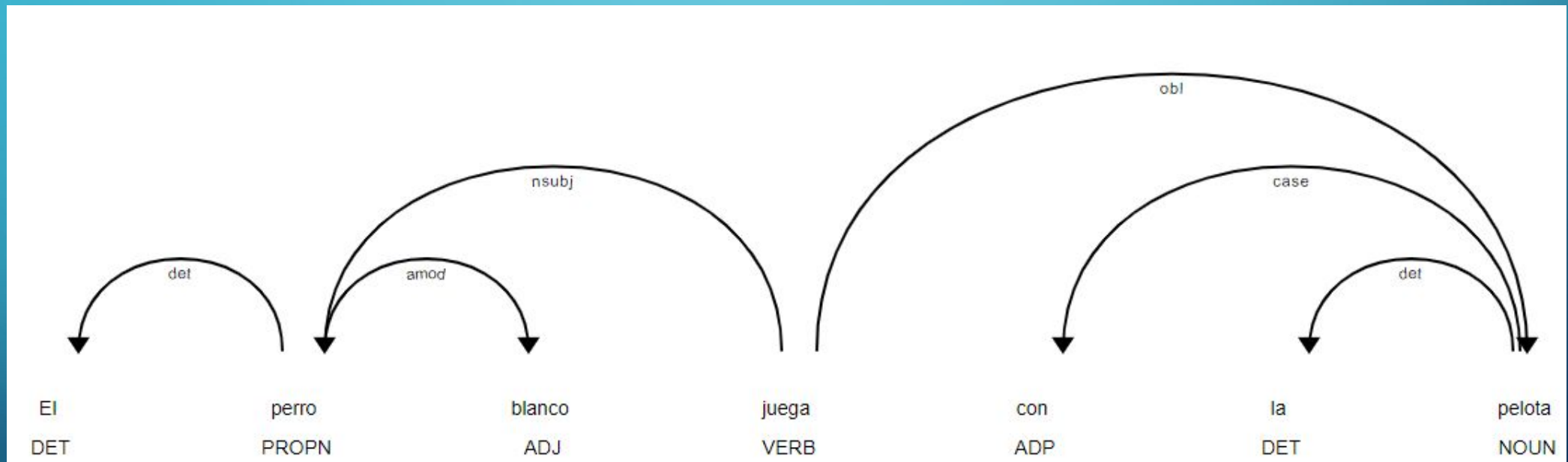
EXTRACCIÓN DE RELACIONES

- La extracción de relaciones es un componente clave para crear gráficos de conocimiento y ontologías.
- Tiene una importancia crucial para las aplicaciones de PLN como:
 - Búsqueda estructurada
 - Análisis de sentimientos
 - Question-Answering
 - Resúmenes automáticos.

EXTRACCIÓN DE RELACIONES

- Es un proceso en el que se utilizan **patrones** para:
- Identificar las entidades y los valores
- Identificar la relación semántica entre ellos
- Clasificar el tipo de relación
 - Propiedad de objeto
 - Propiedad de dato
 - IS-A (subclase de)
 - Instancia de una clase

DEPENDENCIAS EN SPACY



<https://colab.research.google.com/drive/1gedzAEntFPP4F7FTh5FRE1FaVtB8UUQC?usp=sharing>

10/27/2022

A decorative graphic on the left side of the slide, consisting of a network of light blue lines and small circles, resembling a circuit board or a neural network structure.

VECTORES DE PALABRAS

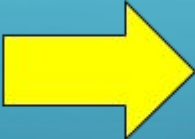
ONE-HOT ENCODING

- Representación vectorial con 0's y 1's, donde el 1 indica la existencia de una palabra en una oración y 0 la ausencia de esta.
- T1: Me gustó el carro rojo de María Luisa
- T2: Me encantó el carro rojo de Luisa
- T3: Me gustó el carro de María Luisa y de Luisa

	Me	gustó	encantó	el	carr o	rojo	de	María	Luisa	y
T1	1	1	0	1	1	1	1	1	1	0
T2	1	0	1	1	1	1	1	0	1	0
T3	1	0	1	1	1	0	1	1	1	1

ONE HOT- ENCODING

- Es una opción rápida para convertir texto a números por medio de vectores.
- Construye matrices muy grandes donde la mayoría de los valores son 0.



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

ONE HOT ENCODING

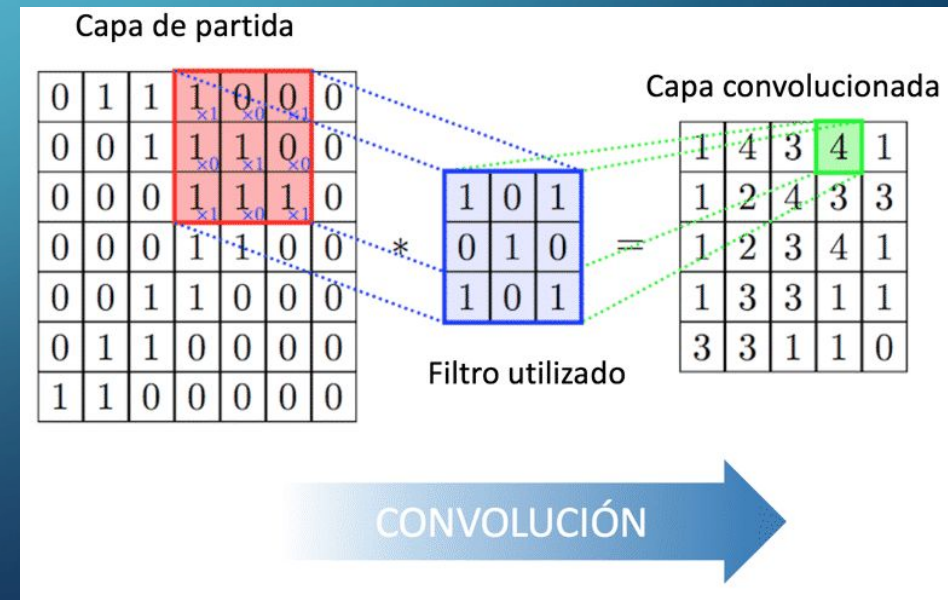
- María juega con Patricia en el jardín de la escuela.“
 - "El perro mordió a Jazmín en la pierna.“
 - "Patricia juega con el perro con una falsa pierna de pollo."
-
- {'maría': 1, 'juega': 2, 'con': 3, 'patricia': 4, 'en': 5, 'el': 6, 'jardín': 7, 'de': 8, 'la': 9, 'escuela': 10, 'perro': 11, 'mordió': 12, 'a': 13, 'jazmín': 14, 'pierna': 15, 'una': 16, 'falsa': 17, 'pollo': 18}

ONE HOT ENCODING

- el perro mordió a jazmín en la pierna
- {'maría': 1, 'juega': 2, 'con': 3, 'patricia': 4, 'en': 5, 'el': 6, 'jardín': 7, 'de': 8, 'la': 9, 'escuela': 10, 'perro': 11, 'mordió': 12, 'a': 13, 'jazmín': 14, 'pierna': 15, 'una': 16, 'falsa': 17, 'pollo': 18}
- [[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], el
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0], perro
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0], mordió
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0], a
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0], jazmín
- [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], en
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0], la
- [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]] pierna

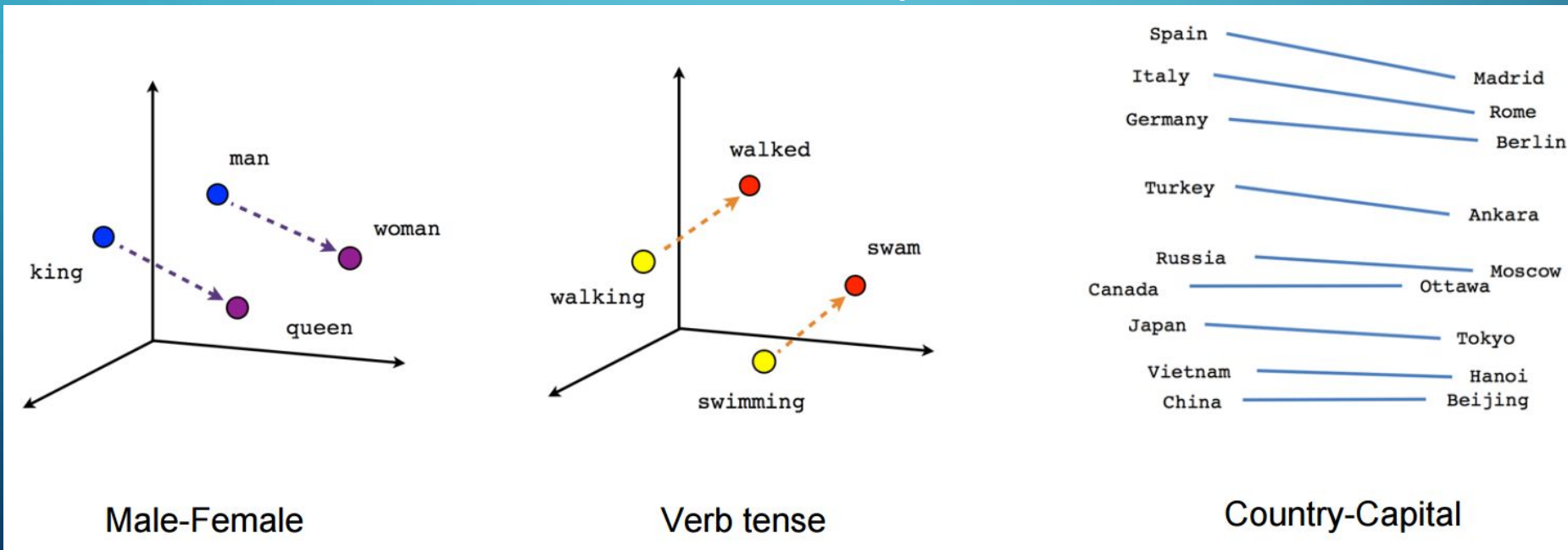
EMBEDDINGS

- Son vectores de palabras con dimensiones reducidas.
- Genera una representación distribuida acorde a un contexto
- Fácilmente interpretados por las computadoras
- No sirven para hacer predicciones de continuidad



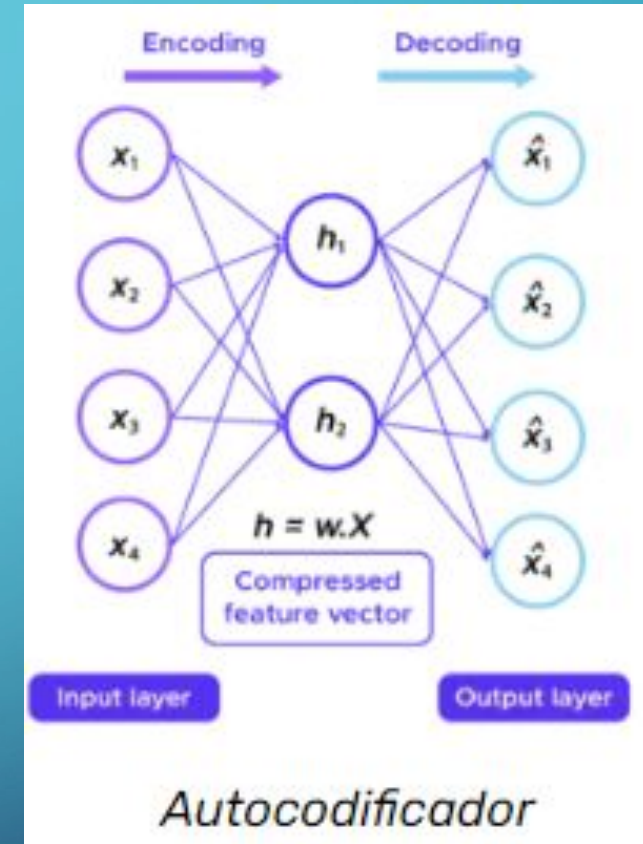
EMBEDDINGS

- Buscan posicionar palabras acorde a la similitud semántica entre ellas dentro de un espacio vectorial.
- Se basa en medir similitud entre vectores por medio de la similitud coseno

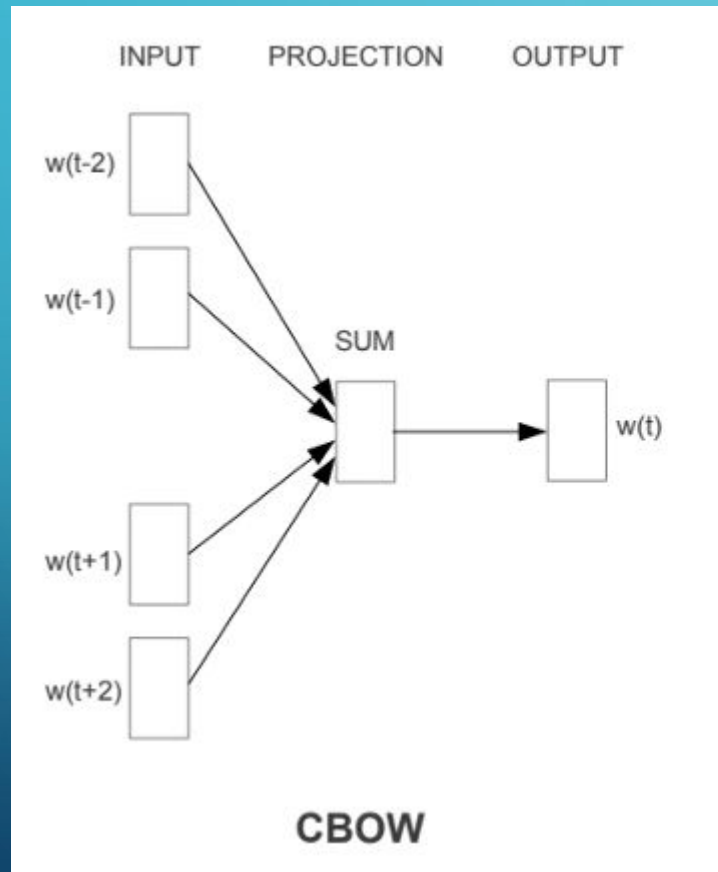


WORD2VEC

- ❖ Es un método de construcción de embeddings por medio de *una red neuronal de 3 capas (entrada, oculta y salida)*.
- ❖ La capa oculta no tiene función de activación y en cambio la de salida aplica una función softmax.
- ❖ Aunque en el entrenamiento, tanto la entrada y como el valor esperado son one hot vectors, la salida es en realidad una distribución de probabilidades.



CBOW MODEL

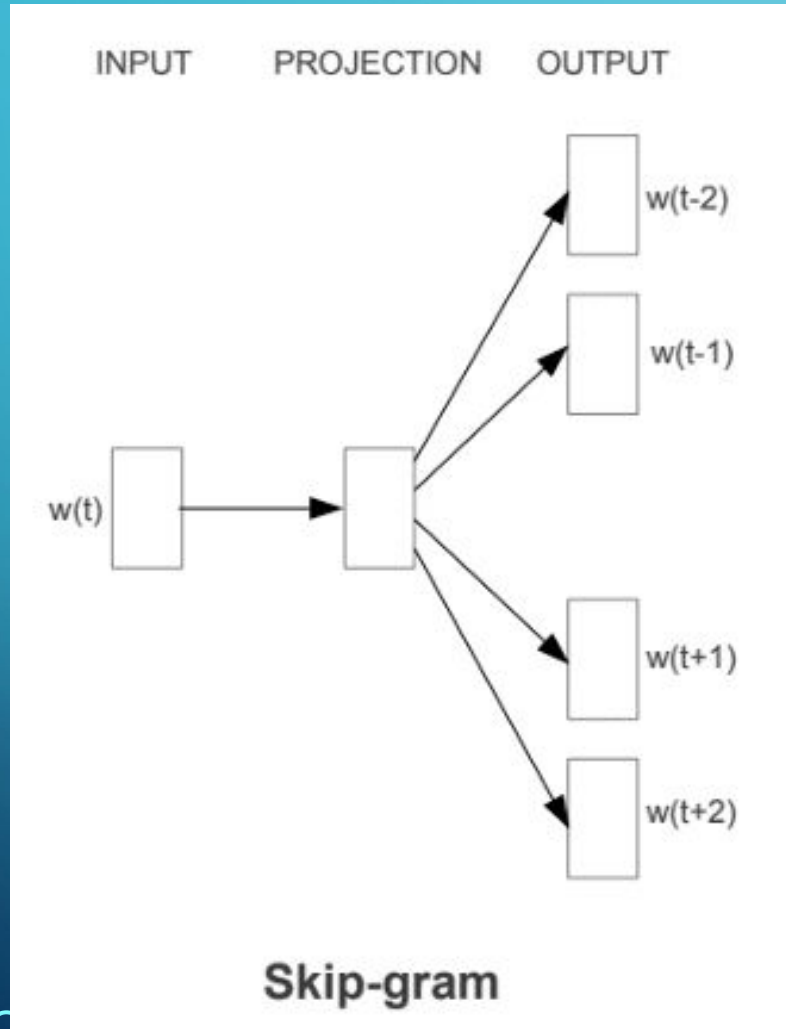


- El modelo se alimenta de las representaciones distribuidas del contexto, las cuales combina para predecir la palabra objetivo.

CBOW

- La capa Embedding **transformará cada palabra del contexto en un vector de embedding**. La matriz W de embedding se aprenderá a medida que se entrena el modelo. Las dimensiones resultantes son: (lot, context_size, embedding).
- A continuación, la capa GlobalAveragePooling1D permite **sumar los diferentes embedding** y obtener una dimensión de salida (batch_size, embedding).
- Por último, la capa densa de tamaño «voc_size» **permite predecir la palabra objetivo**.

SKIP GRAM



- La representación distribuida de la palabra de entrada se utiliza para **predecir el contexto**.
- El modelo se alimenta de la palabra **objetivo** y predice las palabras del contexto. El resultado de la capa oculta es la nueva representación de la palabra (h_1, \dots, h_N).

SKIP GRAM

- La red recibe como entrada la palabra objetivo.
- Para cada posición de contexto, obtenemos C distribuciones de probabilidad de V probabilidades, una para cada palabra.
- Skip Gram funciona bien con una pequeña cantidad de datos y se encuentra que representa bien las palabras raras.
- Por otro lado, CBOW es más rápido y tiene mejores representaciones para palabras más frecuentes.

el aprendizaje profundo permite aprender de forma eficiente representaciones a partir de datos utilizando una abstracción denominada grafo computacional y técnicas de optimización numérica

