



diplomatura universitaria en
inteligencia artificial



FACULTAD DE CIENCIAS
EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES

Procesamiento de Lenguaje Natural

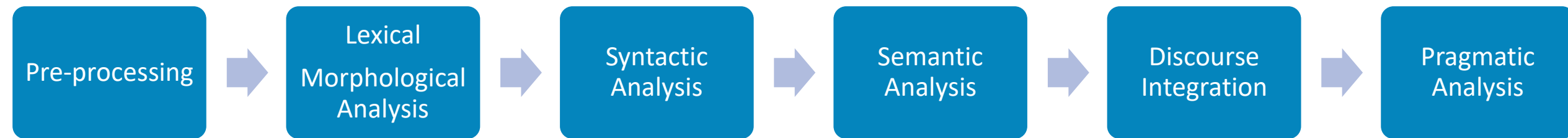
Análisis del significado

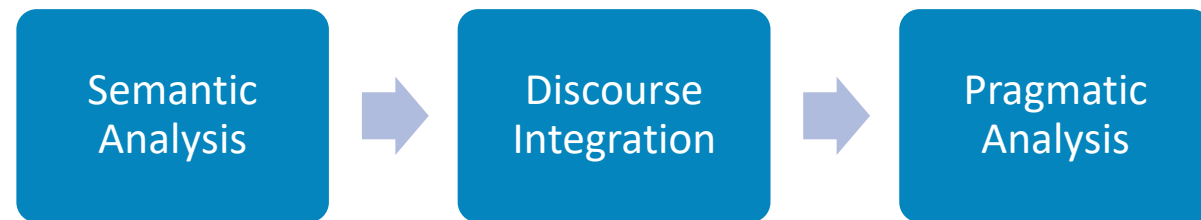
Qué vamos a ver hoy?

Análisis Semántico

- Análisis Semántico
- Semántica léxica
- Desambiguación
- Named Entity Recognition.
- Etiquetado de roles semánticos
- Discurso y pragmática.
- Resumen
- Extracción de Información
- Detección de Estados Afectivos.
 - Sentimiento, emociones, personalidad.
- Traducción
- Detección de Tópicos.







Analiza el significado del texto.

Mapea las estructuras sintácticas
y los objetos respecto al dominio
de la tarea.

Semantic
Analysis



Discourse
Integration



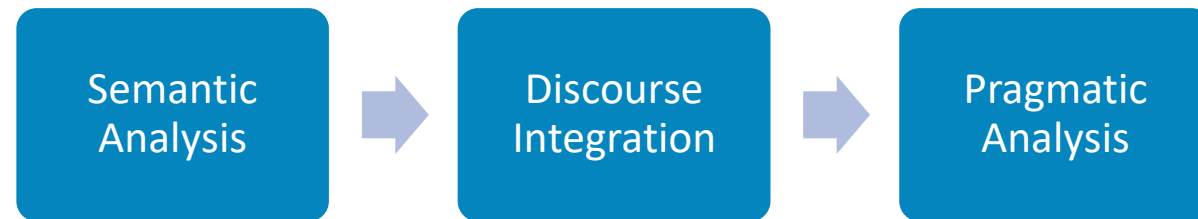
Pragmatic
Analysis

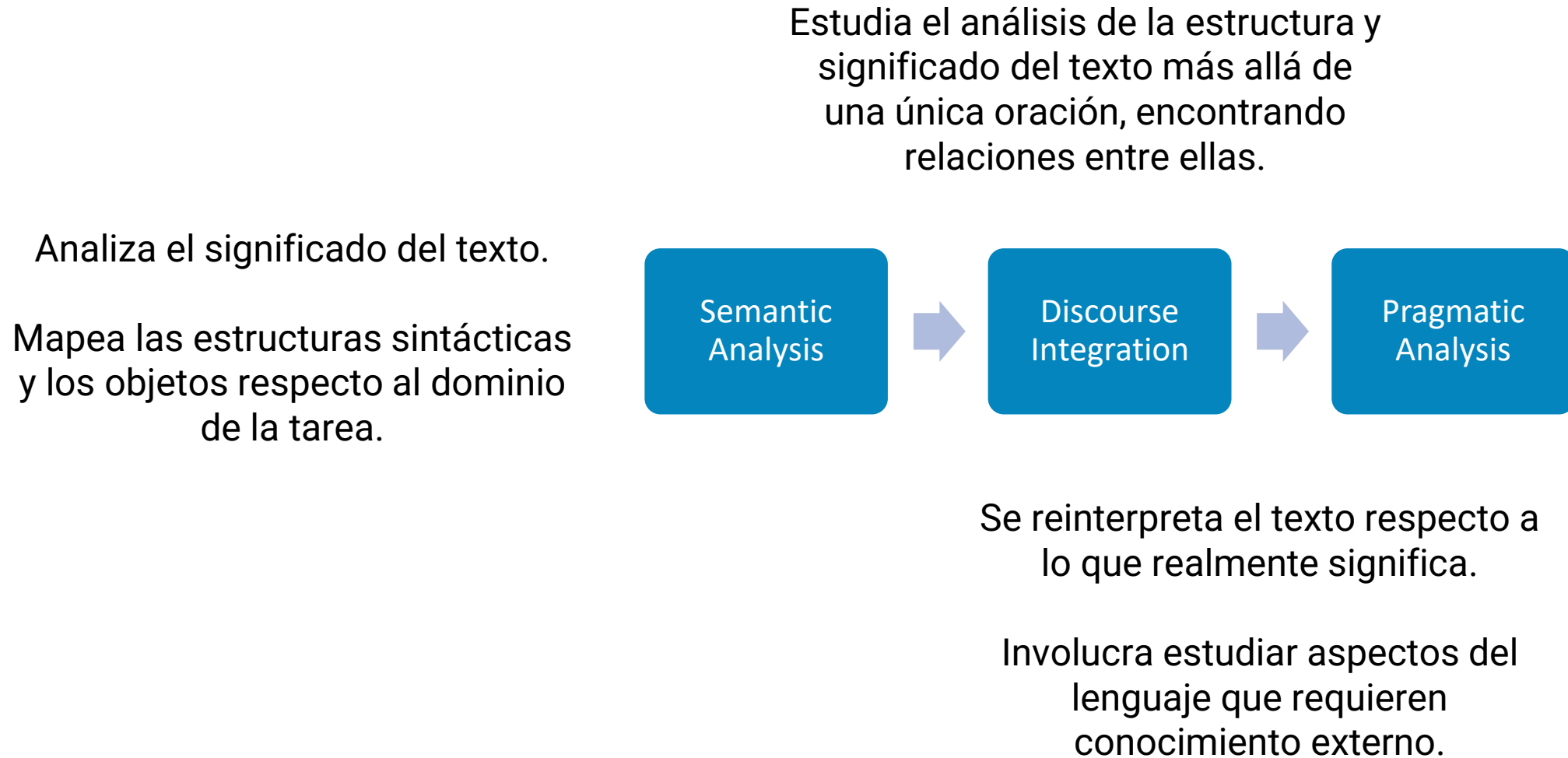


Estudia el análisis de la estructura y significado del texto más allá de una única oración, encontrando relaciones entre ellas.

Analiza el significado del texto.

Mapea las estructuras sintácticas y los objetos respecto al dominio de la tarea.





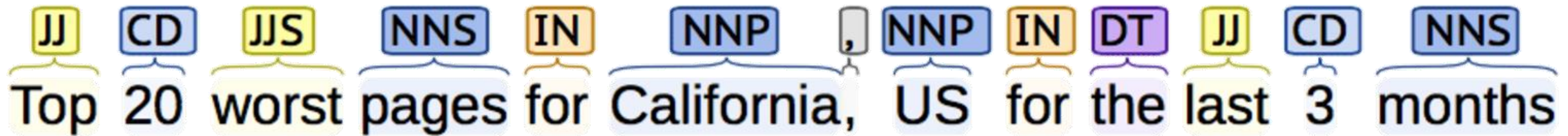
Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.



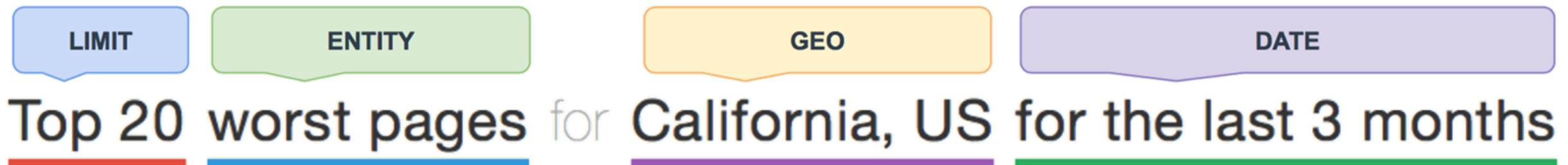
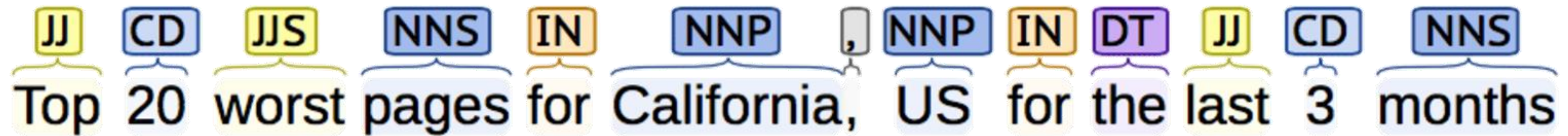
Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.



- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).
- Permite analizar texto en un nivel conceptual opuesto a términos simples.
 - Por ejemplo, durante el proceso de consulta y coincidencia de documentos para la recuperación de información.
- Permite enriquecer textos con el uso de fuentes léxicas.



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).

Entidades

Conceptos

Relaciones

Predicados



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).

Instancias
particulares.
(Tandil, Juan,
Pedro, María...)

Entidades

Conceptos

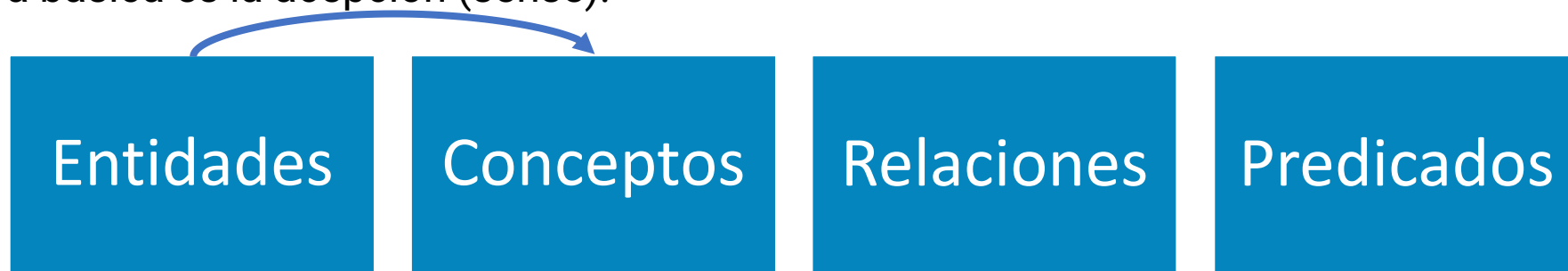
Relaciones

Predicados



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).

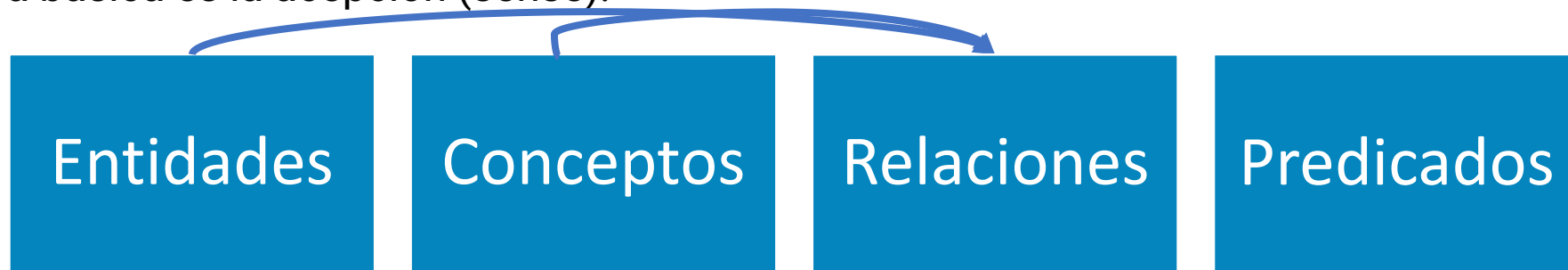


Representan las categorías de las entidades
(personas, ciudades...)



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).



Representan las relaciones entre las entidades y los conceptos
(Tandil es una ciudad)



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).

Entidades

Conceptos

Relaciones

Predicados

Representan las estructuras verbales.
(Ej. Roles semánticos)



Análisis Semántico

- Describe el proceso de comprender el lenguaje natural.
- La semántica se ocupa de la determinación de lo que realmente **significa una oración** al relacionar características sintácticas y desambiguar palabras con múltiples definiciones para el contexto dado.
 - El significado de las palabras y cómo se combinan para formar el significado de las oraciones.
- Este nivel implica la interpretación apropiada del significado de las oraciones, en lugar del análisis a nivel de palabras o frases individuales.
 - La estructura y el contexto ayudan en la comprensión.
 - La estructura básica es la acepción (sense).

Word Sense
Disambiguation

Semantic Role
Labeling

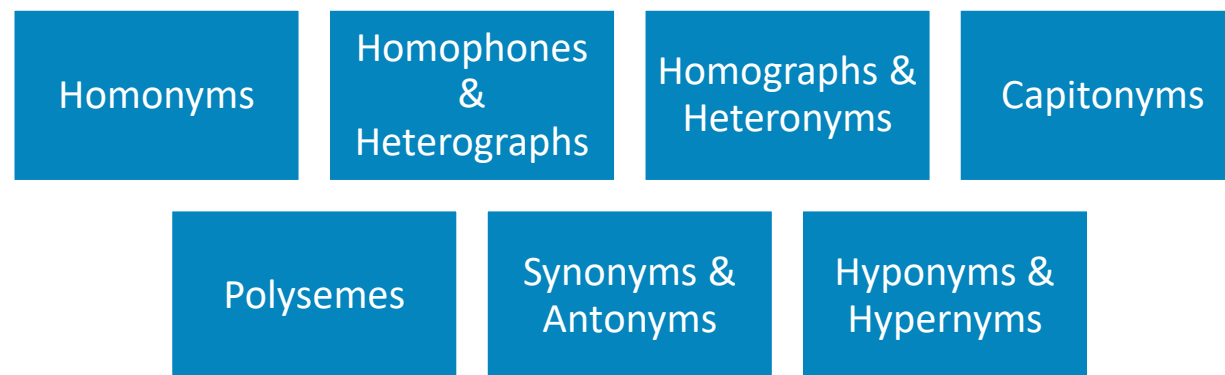
Named Entity
Recognition



- Semántica léxica se refiere al análisis del significado de cada una de los items léxicos.
 - Palabras.
 - Sub-palabras.
 - Afijos.
 - Palabras compuestas.
 - Frases.
- Se trata de la relación entre los items léxicos, el significado de las oraciones y su sintaxis.
 - Cada ítem léxico tiene su propia sintaxis, forma y significado.
 - También obtienen significado a partir de los ítems léxicos que los rodean → contexto, posición

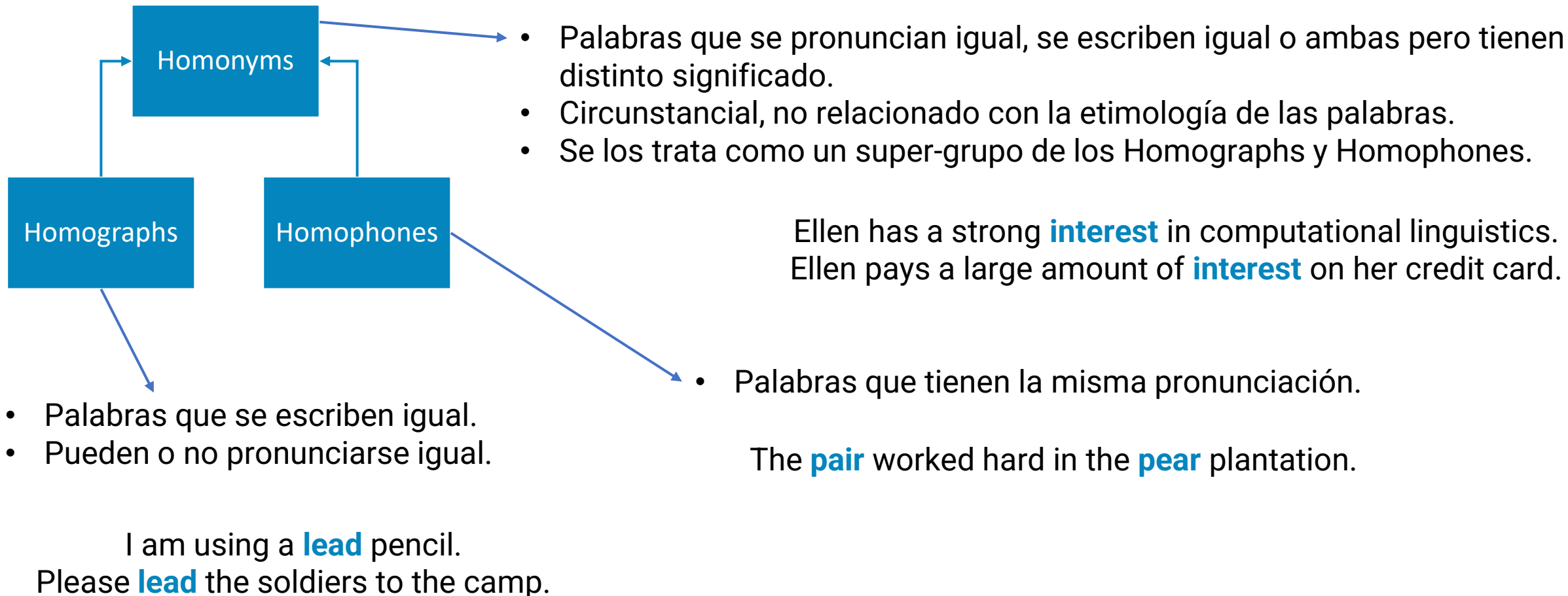
- Semántica léxica se refiere al análisis del significado de cada una de los items léxicos.
 - Palabras.
 - Sub-palabras.
 - Afijos.
 - Palabras compuestas.
 - Frases.
- Se trata de la relación entre los items léxicos, el significado de las oraciones y su sintaxis.
 - Cada ítem léxico tiene su propia sintaxis, forma y significado.
 - También obtienen significado a partir de las ítems léxicos que los rodean → contexto, posición
- Relaciones entre los ítems léxicos.

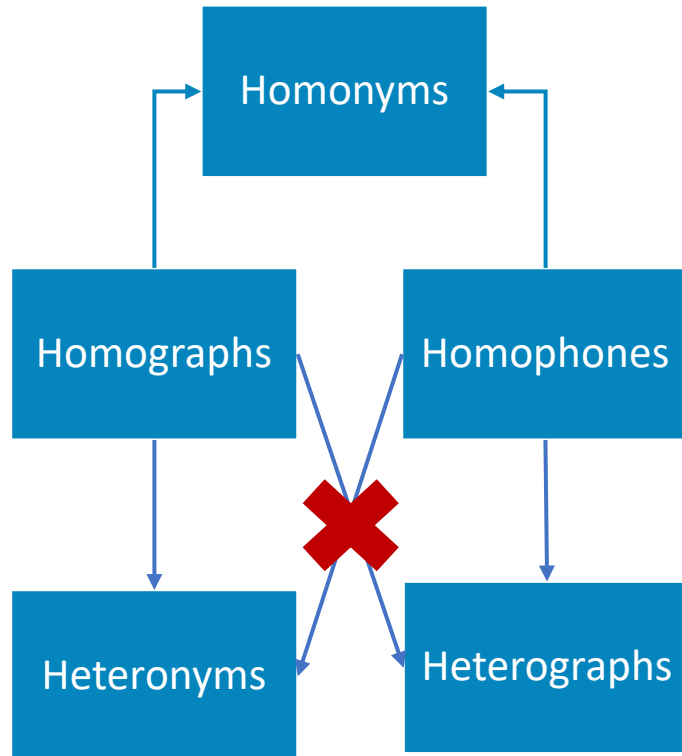
(hablamos de algunas cuando mencionamos WordNet)



Análisis Semántico

Semántica Léxica: Relaciones



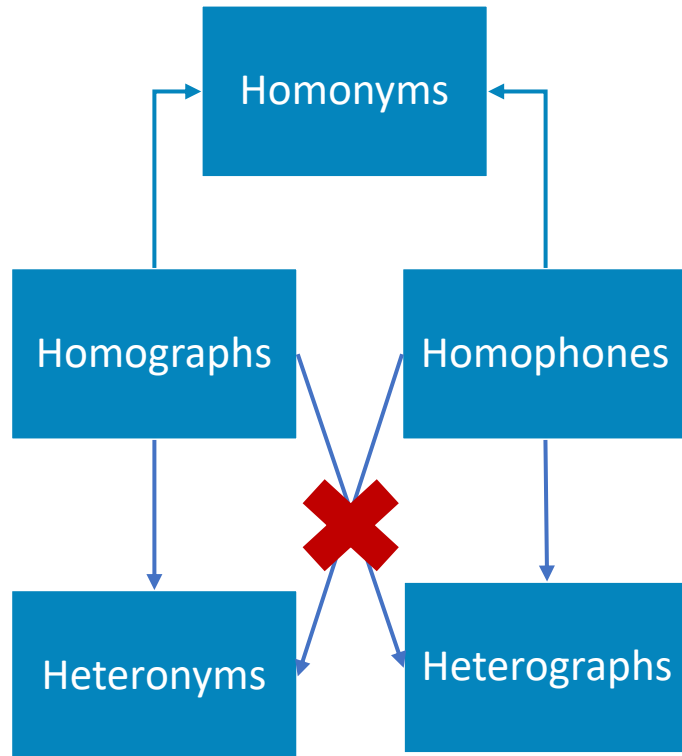


- Palabras que se pronuncian igual, pero se escriben diferente.
- La acepción correcta depende de la escritura.
- Son homophones pero no homographs.

- Palabras que se escriben igual pero tienen distinta pronunciación.
- Son homographs pero no homophones.
- La acepción correcta depende de la pronunciación.

Análisis Semántico

Semántica Léxica: Relaciones



		Escritura	
		<i>Igual</i>	<i>Diferente</i>
Sonido	<i>Igual</i>	Homonyms Homophones Homographs	Heterographs
	<i>Diferente</i>	Heteronyms	Heterophones Heterographs (todo el resto)

Polysemes

- Similar a homonyms.
- Palabras que se escriben igual y tienen diferentes significados.
 - Los **significados** están **relacionados**.

Bank:

1. Institución financiera.
2. El edificio que pertenece a la institución financiera.
3. Tener una cuenta bancaria en un determinado banco.
4. Ganar una suma de dinero.
5. Depositar dinero en una cuenta.
6. Un amontonamiento de tierra, arena o nieve.
7. Acumulación de nubes o niebla.
8. Una hilera de cosas.
9. La vera del río.



Polysemes

- Similar a homonyms.
- Palabras que se escriben igual y tienen diferentes significados.
 - Los **significados** están **relacionados**.

Bank:

1. Institución financiera.
2. El edificio que pertenece a la institución financiera.
3. Tener una cuenta bancaria en un determinado banco.
4. Ganar una suma de dinero.
5. Depositar dinero en una cuenta.
6. Un amontonamiento de tierra, arena o nieve.
7. Acumulación de nubes o niebla.
8. Una hilera de cosas.
9. La vera del río.

Bank en todos los casos representa homonyms, no todas las acepciones se encuentran relacionadas.



Polysemes

- Similar a homonyms.
- Palabras que se escriben igual y tienen diferentes significados.
 - Los **significados** están **relacionados**.

Bank:

1. Institución financiera.
2. El edificio que pertenece a la institución financiera.
3. Tener una cuenta bancaria en un determinado banco.
4. Ganar una suma de dinero.
5. Depositar dinero en una cuenta.
6. Un amontonamiento de tierra, arena o nieve.
7. Acumulación de nubes o niebla.
8. Una hilera de cosas.
9. La vera del río.

Bank en todos los casos representa homonyms, no todas las acepciones se encuentran relacionadas.



Polysemes

- Similar a homonyms.
- Palabras que se escriben igual y tienen diferentes significados.
 - Los **significados** están **relacionados**.

Bank:

1. Institución financiera.
2. El edificio que pertenece a la institución financiera.
3. Tener una cuenta bancaria en un determinado banco.
4. Ganar una suma de dinero.
5. Depositar dinero en una cuenta.
6. Un amontonamiento de tierra, arena o nieve.
7. Acumulación de nubes o niebla.
8. Una hilera de cosas.
9. La vera del río.

Capytonyms

- Palabras que tienen la misma escritura, pero el significado cambia de acuerdo a si están capitalizadas.
- Pueden tener o no la misma pronunciación.

March march

May may



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.

big, huge, large



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.

big, huge, large

The house is **big**.



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.

big, huge, large

The house is **big**.

The house is **huge**.

The house is **large**.



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.

big, huge, large

The house is big.

The house is huge.

The house is large.

He is my **big** brother.



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
- Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
- Es una relación entre las acepciones y no sobre la palabra en sí.
- Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.

big, huge, large

The house is big.

The house is huge.

The house is large.

He is my big brother.

He is my huge brother.



Synonyms & Antonyms

- Los **sinónimos** son palabras diferentes que tienen el mismo significado en alguno o todos los contextos.
 - Diferente en escritura y pronunciación.
 - Se pueden utilizar de forma intercambiable (de acuerdo al contexto).
 - Es una relación entre las acepciones y no sobre la palabra en sí.
 - Pueden existir para sustantivos, adjetivos, verbos, adverbios y preposiciones.
-
- Los **antónimos** son palabras que definen una relación binaria de oposición.
 - Tres tipos:
 - *Graduación*. Son antónimos en una cierta escala o medida.
 - *Complementarios*. No pueden ser medidos en ninguna escala.
 - Dividir, unir.
 - *Relacionales*. Las palabras tienen alguna relación entre ellas y la antonimia contextual es significada por esa relación.
 - Doctor, paciente.



Hypernyms



Hyponyms

- Los **hyponyms** son palabras que son una sub-clase de otras palabras.
- Son palabras con una acepción y contexto determinados en comparación con un hypernym.
- Los **hypernyms** son palabras que actúan como super-clases de hyponyms.
- Tienen una acepción más amplia que los hyponyms.



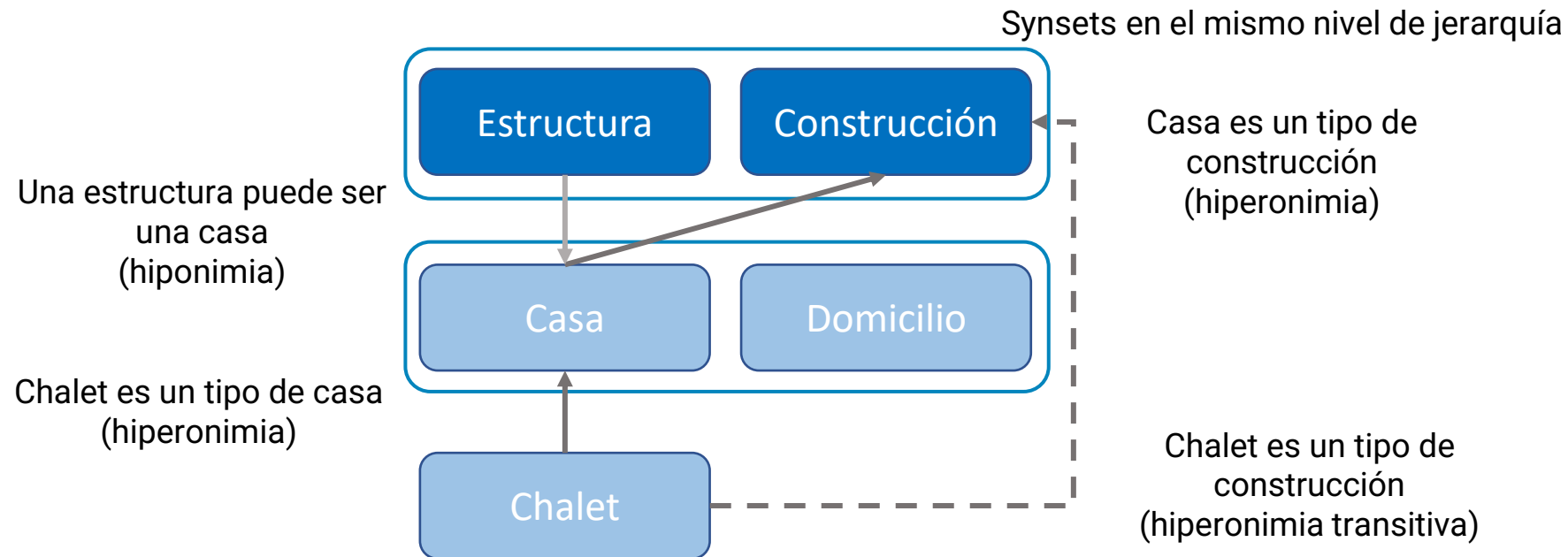
Análisis Semántico

Semántica Léxica: Relaciones

Hypernyms

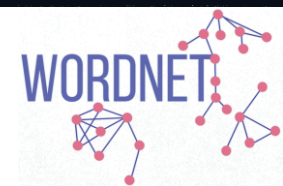
Hyponyms

- Los **hyponyms** son palabras que son una sub-clase de otras palabras.
- Son palabras con una acepción y contexto determinados en comparación con un hypernym.
- Los **hypernyms** son palabras que actúan como super-clases de hyponyms.
- Tienen una acepción más amplia que los hyponyms.



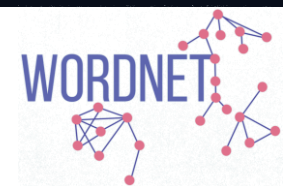
Análisis Semántico

Semántica Léxica: Relaciones en WordNet



Relación	Categoría a la que aplica	Ejemplos
Synonymy	Sustantivo Verbo Adjetivo Adverbio	Pipe, tube Rise, ascend Sad, unhappy Rapidly, speedily
Antonymy	Sustantivo Verbo Adjetivo Adverbio	Top, bottom Rise, fall Wet, dry Rapidly, slowly
Hyponymy	Sustantivo	Tree, plant
Meronymy	Sustantivo	Ship, fleet
Troponymy	Verbo	March, walk
Entailment	Verbo	Buy, pay
Derivation (pertainnyms)	Adjetivo	Magentic, magnetism





Relación de todo y partes

Jerarquía de maneras de realizar algo

Describen eventos que se suceden de forma unidireccional

Relación	Categoría a la que aplica	Ejemplos
Synonymy	Sustantivo Verbo Adjetivo Adverbio	Pipe, tube Rise, ascend Sad, unhappy Rapidly, speedily
Antonymy	Sustantivo Verbo Adjetivo Adverbio	Top, bottom Rise, fall Wet, dry Rapidly, slowly
Hyponymy	Sustantivo	Tree, plant
Meronymy	Sustantivo	Ship, fleet
Troponymy	Verbo	March, walk
Entailment	Verbo	Buy, pay
Derivation (pertainnyms)	Adjetivo	Magentic, magnetism

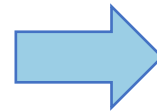


- Las palabras suelen tener más un significado posible.
 - La relación de polisemia, o heteronimia.



- Las palabras suelen tener más un significado posible.
 - La relación de polisemia, o heteronimia.

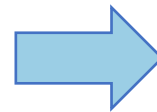
Una palabra en un contexto.
Un inventario fijo de potenciales
acepciones.



Decidir cuál de todas las
acepciones es la
adecuada

- Las palabras suelen tener más un significado posible.
 - La relación de polisemia, o heteronimia.

Una palabra en un contexto.
Un inventario fijo de potenciales
acepciones.

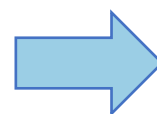


Decidir cuál de todas las
acepciones es la
adecuada

Ellen has a strong **interest** in computational linguistics.
Ellen pays a large amount of **interest** on her credit card.

- Las palabras suelen tener más un significado posible.
 - La relación de polisemia, o heteronimia.

Una palabra en un contexto.
Un inventario fijo de potenciales
acepciones.



Decidir cuál de todas las
acepciones es la
adecuada

Ellen has a strong **interest** in computational linguistics.
Ellen pays a large amount of **interest** on her credit card.

- Relevante para diversas tareas:
 - Machine Translation.
 - Information Retrieval
 - Text Mining & Information Extraction
 - ...

Ambigüedad
Léxica

Ambigüedad
Sintáctica

Ambigüedad
Semántica

Ambigüedad
Léxica

Ambigüedad
Sintáctica

Ambigüedad
Semántica

- Ambigüedad a nivel palabra.
- Una misma palabra puede tener un significado ambiguo de acuerdo a su contexto.

Ambigüedad
Léxica

Ambigüedad
Sintáctica

Ambigüedad
Semántica

- Ambigüedad a nivel palabra.
 - Una misma palabra puede tener un significado ambiguo de acuerdo a su contexto.
-
- Diferentes formas de interpretar una secuencia de palabras.
 - Estructural.
 - Frases pre-posicionales.
 - Coordinación.

Ambigüedad
Léxica

Ambigüedad
Sintáctica

Ambigüedad
Semántica

- La estructura sintáctica no es ambigua, pero el significado de las palabras puede ser mal interpretado.
 - Por ejemplo, por contexto.
- En contextos multi-lengua, se asocia también a los “falsos transparentes”.



Ambigüedad
Léxica

Ambigüedad
Sintáctica

Ambigüedad
Semántica

- La estructura sintáctica no es ambigua, pero el significado de las palabras puede ser mal interpretado.
 - Por ejemplo, por contexto.
- En contextos multi-lengua, se asocia también a los “falsos transparentes”.

X e Y están casados.

Juntos? O con otras personas?



- Tradicionalmente se utilizan métodos basados en diccionarios o fuentes de conocimiento externas.
 - No utilizan corpus.
- Existen otras técnicas basadas en modelos supervisados o no supervisados.
 - Requieren usar corpus.
- NLTK se basa en el método de Lesk que mide la superposición entre definiciones de sentido para todas las palabras en contexto.
 - Una palabra a la vez.
 - Contexto == conjunto de palabras en una oración o párrafo circundante.
- La desambiguación se incluye en la herramienta IBM Chatbot para determinar la intención de los usuarios.



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
- Requiere diccionarios (como WordNet).
- Dada una palabra W en un contexto C .
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C .
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C .



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
- Requiere diccionarios (como WordNet).
- Dada una palabra W en un contexto C .
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C .
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C .

Palabra: bank

Contexto: river bank

1. S: (n) sloping land (especially the slope beside a body of water)
2. S: (n) a financial institution that accepts deposits and channels the money into lending activities
3. S: (n) a long ridge or pile
4. S: (n) an arrangement of similar objects in a row or in tiers
5. ...

1. S: (n) a large natural stream of water (larger than a creek)



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
- Requiere diccionarios (como WordNet).
- Dada una palabra W en un contexto C.
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C.
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C.

Palabra: bank

Contexto: river bank

1. S: (n) sloping land (especially the slope beside a body of **water**)
2. S: (n) a financial institution that accepts deposits and channels the money into lending activities
3. S: (n) a long ridge or pile
4. S: (n) an arrangement of similar objects in a row or in tiers
5. ...

1. S: (n) a large natural stream of **water** (larger than a creek)



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
 - Requiere diccionarios (como WordNet).
 - Dada una palabra W en un contexto C.
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C.
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C.
 - Además de las definiciones se pueden agregar los ejemplos.
 1. S: (n) sloping land (especially the slope beside a body of water). *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
 2. S: (n) a financial institution that accepts deposits and channels the money into lending activities. *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
 3. S: (n) a long ridge or pile. *"a huge bank of earth"*
 4. S: (n) .an arrangement of similar objects in a row or in tiers. *"he operated a bank of switches"*
1. S: (n) **river** (a large natural stream of water (larger than a creek)) *"the river was navigable for 50 miles"*



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
- Requiere diccionarios (como WordNet).
- Dada una palabra W en un contexto C .
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C .
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C .
- Además de las definiciones se pueden agregar los ejemplos.
- En una version simplificada, no se consideran las definiciones de las palabras en el contexto, sino solo las palabras.
- Se le puede asignar peso a las palabras de las acepciones/contexto. Por ejemplo, IDF.



Word Sense Disambiguation: Lesk

- Implementación original de 1986.
- Requiere diccionarios (como WordNet).
- Dada una palabra W en un contexto C .
 - Obtener todas las acepciones S de W del diccionario.
 - Comparar las palabras en cada S con las palabras en cada una de las definiciones de las palabras en C .
 - Seleccionar el S con el mayor overlapping con las definiciones de las palabras en C .
- Además de las definiciones se pueden agregar los ejemplos.
- En una version simplificada, no se consideran las definiciones de las palabras en el contexto, sino solo las palabras.
- Se le puede asignar peso a las palabras.

Pros

- Simple.
- No require datos de entrenamiento etiquetados.

Cons

- Muy sensible a los términos en las definiciones.
- Las palabras de las definiciones podrían no tener overlap con el contexto (ni con sus definiciones).



Word Sense Disambiguation: Métodos Supervisados

- Requieren datos de training etiquetados.
- Se utilizan corpus con etiquetas de acepciones.
 - SemCor corpus derivado de WordNet y creado por Princeton University.
 - Sub-conjunto de Brown Corpus.
 - Aproximadamente 352 textos.
 - No cubre completamente WordNet.
 - Es uno de los corpus más grandes.
- La idea es tener la palabra a desambiguar, las acepciones correctas (lo que sería el target de nuestra predicción) y el contexto en el que aparece cada acepción.
- Qué características utilizar en el clasificador? (por ejemplo)
- Lista de palabras más frecuentes que suelen aparecer junto con la palabra para cada acepción.
 - Vector de co-ocurrencias. Las palabras pueden estar representadas por su lemma.
 - No hay indicaciones respecto a la posición que suele ocupar la palabra en ese contexto.
- Información de etiquetas POS y algún tipo de parsing sobre las palabras alrededor de la que nos interesa en ejemplos de uso de las acepciones.



Word Sense Disambiguation: Embeddings → sense2vec

- Se basa en word2vec.
- Trata de agregar contextualidad a los embeddings de las palabras, basadas en su desambiguación.
- En lugar de predecir un token dado el contexto, el objetivo es predecir el uso correcto dado el contexto.
- Requiere de corpus etiquetados.
 - Cuenta la cantidad de usos de una palabra, donde cada uso es presentado por una etiqueta.
 - Genera un embedding random para cada uno de los usos.
 - Entrena el modelo usando la arquitectura de word2vec.

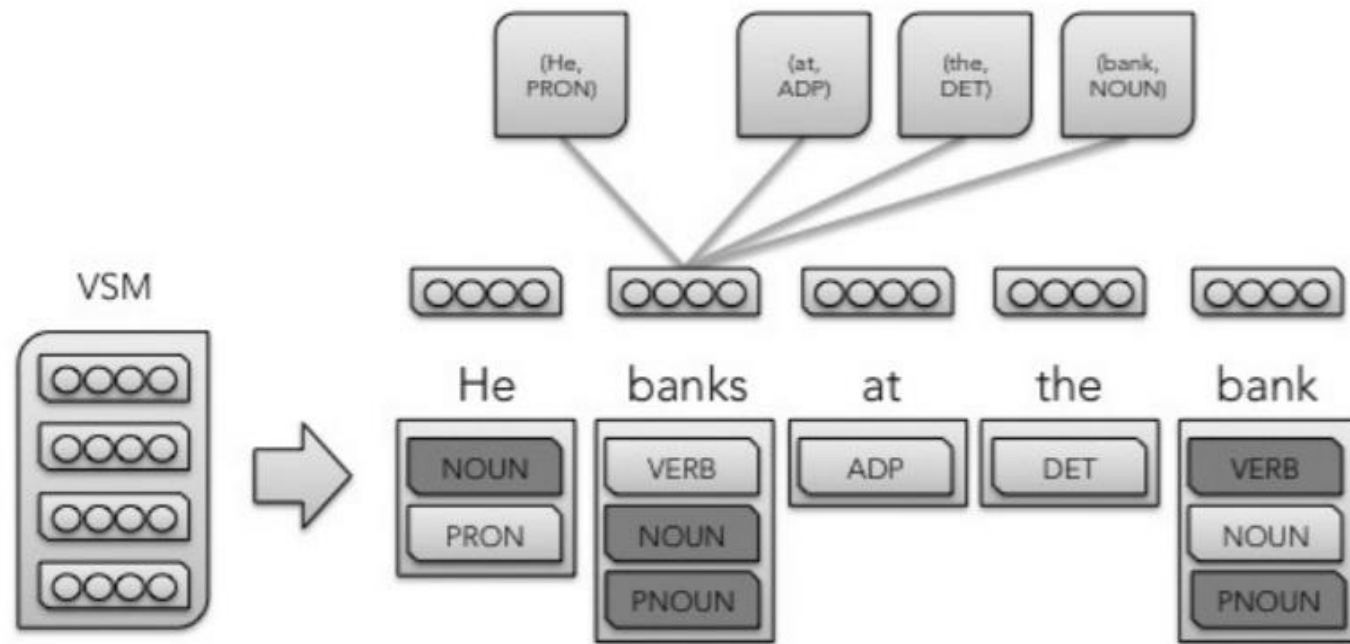
[Sense2vec - A Fast And Accurate Method For Word Sense Disambiguation In Neural Word Embeddings](#)



Análisis Semántico

Word Sense Disambiguation: Embeddings → sense2vec

- Se basa en word2vec.
- Trata de agregar contextualidad a los embeddings de las palabras, basadas en su desambiguación.
- En lugar de predecir un token dado el contexto, el objetivo es predecir el uso correcto dado el contexto.
- Requiere de corpus etiquetados.
 - Cuenta la cantidad de usos de una palabra, donde cada uso es presentado por una etiqueta.
 - Genera un embedding random para cada uno de los usos.
 - Entrena el modelo usando la arquitectura de word2vec.



[Method For Word Sense Disambiguation In Neural Word Embeddings](#)



Word Sense Disambiguation: Embeddings → GlossBERT

- Adaptación de BERT para la desambiguación basada en las representaciones contextualizadas que genera.
 - Problema de clasificación de oraciones.
- Trabajos previos mostraron que métodos simples basados en WordNet y glosarios han dado buenos resultados.
- Se basa en transformar la entrada para BERT.
- Para cada una de las oraciones que se tiene, construir pares de contexto-gloss (extraídos de WordNet).
 - Considerar todos los gloss correspondientes a cada una de las acepciones a desambiguar.
 - Se asignará clase “True” a la acepción correcta.

[GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge](#)



Word Sense Disambiguation: Embeddings → GlossBERT

- Adaptación de BERT para la desambiguación basada en las representaciones contextualizadas que genera.
 - Problema de clasificación de oraciones.
- Trabajos previos mostraron que métodos simples basados en WordNet y glosarios han dado buenos resultados.
- Se basa en transformar la entrada para BERT.
- Para cada una de las oraciones que se tiene, construir pares de contexto-gloss (extraídos de WordNet).
 - Considerar todos los gloss correspondientes a cada una de las acepciones a desambiguar.

Sentence with four targets:

Your research stopped when a convenient assertion could be made.

Context-Gloss Pairs of the target word [research]

	Label	Sense Key
[CLS] Your research ... [SEP] systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your research ... [SEP] a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your research ... [SEP] inquire into [SEP]	No	research%2:31:00::
[CLS] Your research ... [SEP] attempt to find out in a ... [SEP]	No	research%2:32:00::

edge



Word Sense Disambiguation: Embeddings → GlossBERT

- Adaptación de BERT para la desambiguación basada en las representaciones contextualizadas que genera.
 - Problema de clasificación de oraciones.
- Trabajos previos mostraron que métodos simples basados en WordNet y glosarios han dado buenos resultados.
- Se basa en transformar la entrada para BERT.
- Para cada una de las oraciones que se tiene, construir pares de contexto-gloss (extraídos de WordNet).
 - Considerar todos los gloss correspondientes a cada una de las acepciones a desambiguar.
 - Se asignará clase “True” a la acepción correcta.
- También utiliza SemCor.
- Diversas opciones de cómo combinar la salida de las diferentes capas.
 - Por cada token en la oración. Si hay más de uno se promedian.
 - Por oración completa.

[GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge](#)



Named Entity Recognition: Reconocimiento, Identificación o Extracción de Entidades

- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...

Michael Dell es el CEO de Dell Computer Corporation y vive en Austin Texas.



- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...

Michael Dell es el CEO de Dell Computer Corporation y vive en Austin Texas.

- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...

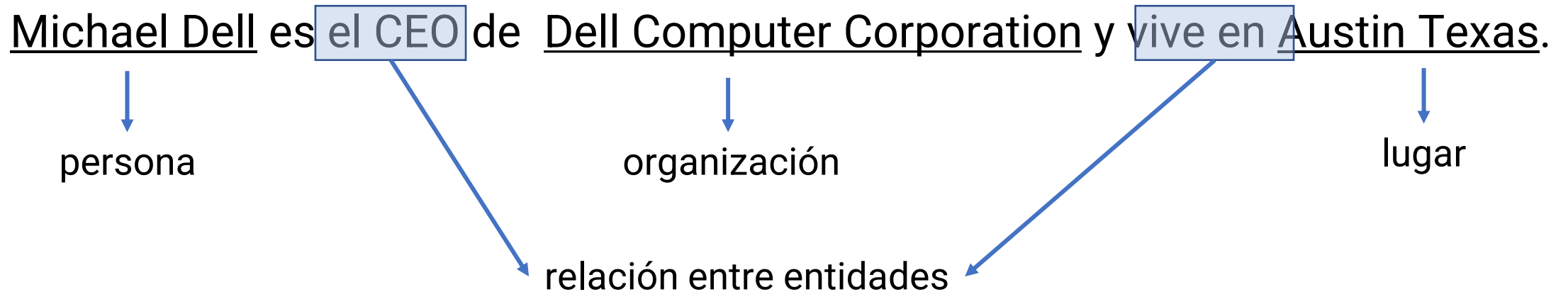
Michael Dell es el CEO de Dell Computer Corporation y vive en Austin Texas.

↓
persona

↓
organización

↓
lugar

- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...



- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...
- Puede sufrir de ambigüedad de tipo.
 - Sabemos que es una entidad, pero no sabemos de qué tipo.

Washington Nombre de persona, organización, entidad política, ciudad

IRA Persona, organización, entidad política

Louis Vuitton Persona, Organización, producto

- Primer paso para la extracción de información.
- Busca identificar y clasificar elementos en el texto en categorías predefinidas como los nombres de:
 - Personas.
 - Organizaciones.
 - Lugares.
 - Expresiones temporales.
 - Cantidades.
 - Valores monetarios, porcentajes.
 - ...
- NER puede dar respuesta a preguntas cómo:
 - Qué compañías fueron mencionadas en el artículo?
 - Qué productos estaban mencionados en la review?
 - Nombra a alguien el tweet? Incluye la ubicación de alguien?

Named Entity Recognition: Algunos usos

- **Categorizar tickets de Customer Support.**
 - Categorizar los issues y consultas puede ayudar a mejorar los tiempos de respuesta y la satisfacción al cliente.
 - Extraer datos como nombres de product, serial numbers, ... también puede ayudar a facilitar la asignación del issue al agente más adecuado.
- **Ganar insights de Customer Feedback.**
 - Las reviews pueden ser una buena fuente de customer feedback. Pueden proveer información acerca de aquellos que a los clients les gusta y no les gusta, y aspectos en los que se necesita mejorar.
 - Pueden permitir detectar problemas recurrentes.
- **Procesos de selección de personal.**
 - Los CVs no suelen tener un formato particular, lo que dificulta la detección manual de información.
 - NER permite extraer información relevante de las actividades profesionales y personales.
- **Salud.**
 - Extracción de información de los resultados de análisis y reports.
 - Roche lo hace con reportes de patologías y radiología.
- **Búsquedas.**
 - Mejorar la velocidad y relevancia de las búsquedas a partir del resumen de los textos descriptivos, resúmenes,



- **Basados en léxicos.**
 - Utilizan ontologías que contienen todas las palabras o términos relacionados a un tópico particular.
 - Por ejemplo, se pueden utilizar léxicos de ciudades o países para reconocer ubicaciones geográficas.
 - La desventaja es que requiere actualización a medida que nuevas palabras.
- **Basados en reglas.**
 - Utilizan una serie de reglas gramaticales.
 - Díficil de escalar y modificar.
 - Sirven para extraer nombres de calles, números de teléfono y otros tipos de datos que siguen patrones específicos.
 - No se adapta fácil a múltiples dominios.
 - Se puede alcanzar buena precisión, pero bajo recall.
 - Las que encontremos van a estar correctas, pero no vamos a encontrar todas.
- **Basados en técnicas de aprendizaje.**
 - Se requiere corpus de entrenamiento con ejemplos positivos (y negativos).
 - Hay corpus disponibles.
 - Medicina, películas, mails, “entidades emergentes”.

- **Basados en léxicos.**
 - Utilizan ontologías que contienen todas las palabras o términos relacionados a un tópico particular.
 - Por ejemplo, se pueden utilizar léxicos de ciudades o países para reconocer ubicaciones geográficas.
 - La desventaja es que requiere actualización a medida que nuevas palabras.
- **Basados en reglas.**
 - Utilizan una serie de reglas gramaticales.
 - Díficil de escalar y modificar.
 - Sirven para extraer nombres de calles, números de teléfono y otros tipos de datos que siguen patrones específicos.
 - No se adapta fácil a múltiples dominios.
 - Se puede alcanzar buena precisión, pero bajo recall.
 - Las que encontremos van a estar correctas, pero no vamos a encontrar todas.
- **Basados en técnicas de aprendizaje.**
 - Se requiere corpus de entrenamiento con ejemplos positivos (y negativos).
 - Hay corpus disponibles.
 - Medicina, películas, mails, “entidades emergentes”.

Only RB B-NP O
France NNP I-NP B-LOC
and CC I-NP O
Britain NNP I-NP B-LOC
backed VBD B-VP O
Fischler NNP B-NP B-PER
's POS B-NP O
proposal NN I-NP O
. . O O



- **Basados en léxicos.**
 - Utilizan ontologías que contienen todas las palabras o términos relacionados a un tópico particular.
 - Por ejemplo, se pueden utilizar léxicos de ciudades o países para reconocer ubicaciones geográficas.
 - La desventaja es que requiere actualización a medida que nuevas palabras.
- **Basados en reglas.**
 - Utilizan una serie de reglas gramaticales.
 - Díficil de escalar y modificar.
 - Sirven para extraer nombres de calles, números de teléfono y otros tipos de datos que siguen patrones específicos.
 - No se adapta fácil a múltiples dominios.
 - Se puede alcanzar buena precisión, pero bajo recall.
 - Las que encontremos van a estar correctas, pero no vamos a encontrar todas.
- **Basados en técnicas de aprendizaje.**
 - Se requiere corpus de entrenamiento con ejemplos positivos (y negativos).
 - Hay corpus disponibles.
 - Medicina, películas, mails, “entidades emergentes”.

Se pueden combinar!



- Booking utiliza NER para extraer de las reviews no estructuradas detalles como los lugares visitados, el tipo de propiedad y las facilidades/comodidades.
- Utilizaron datos sintéticos creados a partir de combinaciones y permutaciones de su léxico de lugares, propiedades y comodidades.
- Diversos modelos.
- En modelos de clasificación tradicional, agregaron características como:
 - Tiene números?
 - Tiene letras?
 - Tiene números y letras?
 - Contiene hyphes?
 - Alterna con mayúsculas en el medio del texto?

<https://booking.ai/named-entity-classification-d14d857cb0d5>



- Booking utiliza NER para extraer de las reviews no estructuradas detalles como los lugares visitados, el tipo de propiedad y las facilidades/comodidades.
- Utilizaron datos sintéticos creados a partir de combinaciones y permutaciones de su léxico de lugares, propiedades y comodidades.
- Diversos modelos.
- En modelos de clasificación tradicional, agregaron características como:
 - Tiene números?
 - Tiene letras?
 - Tiene números y letras?
 - Contiene hyphes?
 - Alterna con mayúsculas en el medio del texto?
- Diversas aplicaciones en salud.
- Extraer de reportes.
 - Diagnósticos.
 - Síntomas.
 - Medicación.

<https://booking.ai/named-entity-classification-d14d857cb0d5>



Análisis Semántico

Named Entity Recognition

- spaCy. Los modelos entrenados con OntoNote 5 incluyen los siguientes tipos de entidades:
- spaCy. Los modelos entrenados con Wikipedia tienen un esquema de anotado más general y reconocen:

TYPE	DESCRIPTION
PER	Named person or family.
LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
ORG	Named corporate, governmental, or other organizational entity.
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art.

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

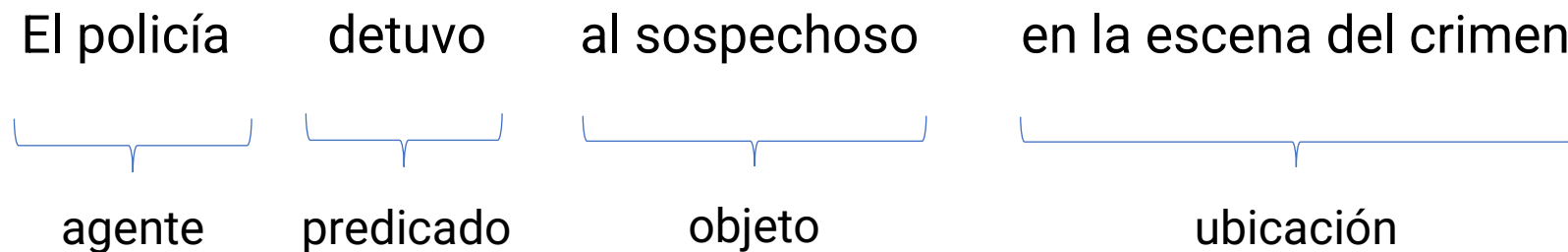


- Un análisis semántico superficial para representar eventos y sus participantes.
- Para cada clausula, determina el rol semántico por cada frase nominal que hace de argumento del verbo.
- Para qué sirve?
 - Question Answering.
 - Por ejemplo, permite responder las preguntas “W” (what, where, when, ...).
 - Hacer resúmenes y extraer las ideas principales de textos.
 - Hacer representaciones semánticas de los textos en forma de grafos.
 - Comparar traducciones respecto a su contenido y semántica.
 - Puede ayudar en el entrenamiento de chatbots.

- Un análisis semántico superficial para representar eventos y sus participantes.
- Para cada clausula, determina el rol semántico por cada frase nominal que hace de argumento del verbo.

El policía detuvo al sospechoso en la escena del crimen

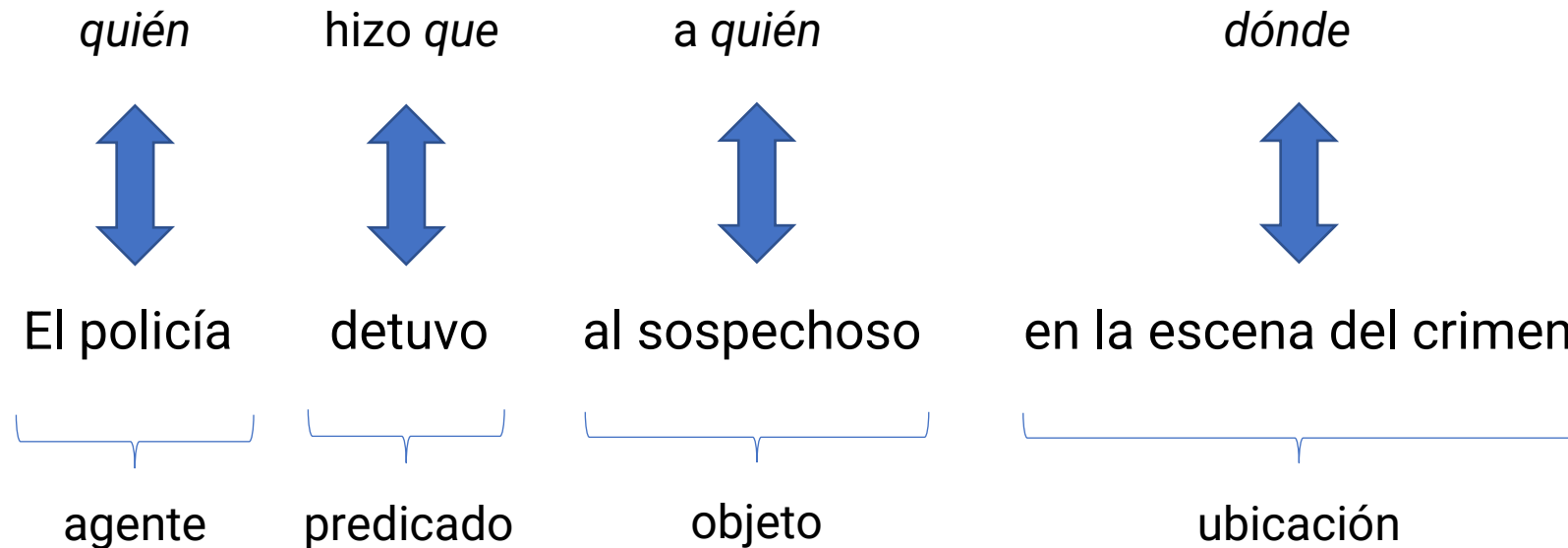
- Un análisis semántico superficial para representar eventos y sus participantes.
- Para cada clausula, determina el rol semántico por cada frase nominal que hace de argumento del verbo.



Análisis Semántico


Semantic Role Labeling

- Un análisis semántico superficial para representar eventos y sus participantes.
- Para cada clausula, determina el rol semántico por cada frase nominal que hace de argumento del verbo.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - *PropBank*: Proto-roles.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - **FrameNet: frame-specific roles.** 
 - *PropBank*: Proto-roles.
- Define tres elementos:


centrado en una noción abstracta que generaliza las descripciones de verbos similares (ej. “describir”, “caracterizar”) y sustantivos (ej. “descripción”)

Frame

Frame
elements

Lexical
units



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - **FrameNet: frame-specific roles.** 
 - *PropBank*: Proto-roles.
- Define tres elementos:

Frame

- Representaciones esquemáticas de situaciones que involucran distintos participantes, propiedades y otros roles conceptuales.
 - Por ejemplo: una situación de compra-venta.
- Los frames se encuentran conectados entre si.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - **FrameNet: frame-specific roles.**
 - *PropBank*: Proto-roles.

- Define tres elementos:

Frame


```
frame(TRANSPORTATION)
frame_elements(MOVER(S), MEANS, PATH)
scene(MOVER(S) move along PATH by MEANS)

frame(DRIVING)
inherit(TRANSPORTATION)
frame_elements(DRIVER (=MOVER), VEHICLE
(=MEANS), RIDER(S) (=MOVER(S)), CARGO
(=MOVER(S)))
scenes(DRIVER starts VEHICLE, DRIVER con-
trols VEHICLE, DRIVER stops VEHICLE)

frame(RIDING_1)
inherit(TRANSPORTATION)
frame_elements(RIDER(S) (=MOVER(S)), VE-
HICLE (=MEANS))
scenes(RIDER enters VEHICLE,
VEHICLE carries RIDER along PATH,
RIDER leaves VEHICLE )
```

centrado en una noción abstracta que
descripciones de verbos
describir", "caracterizar") y
(ej. "descripción")




- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - **FrameNet: frame-specific roles.** 
 - *PropBank*: Proto-roles.
- Define tres elementos:
 - Participantes, propiedades y roles en un frame.
 - Puede incluir agentes u objetos inanimados.
 - Definidos de forma relativa a un único frame.
 - Cualquier conexión entre Frame Elementos tiene que ser realizada de forma explícita.
- Los argumentos sintácticos de un verbo (o cualquier otra forma de predicado) se corresponden con los Frame Elements correspondientes al frame asociado a dicho predicado.

Frame
elements

centrado en una noción abstracta que generaliza las descripciones de verbos similares (ej. "describir", "caracterizar") y sustantivos (ej. "descripción")



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - **FrameNet: frame-specific roles.** 
 - *PropBank*: Proto-roles.
- Define tres elementos:
 - Representan el pairing de un lemma con un significado.
 - Puede incluir formas declinadas.
 - Por ejemplo: see, saw, seen.
 - Puede incluir expresiones de varias palabras.
 - Por ejemplo: pick up, New York.
 - Puede corresponderse con cualquier POS tag.
 - Las definiciones parecen de diccionario y cada entrada representa aspectos más finos que lo que el frame distingue.

Lexical
units



- Un intermedio entre el parsing y el análisis semántico completo.
 - La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**
- orientado a los verbos
- Agrega una capa semántica al Penn Treebank.
 - Intenta capturar la estructura predicado-argumento a partir de anotar los predicados y los roles semánticos de sus argumentos.
 - Diferencias con FrameNet.
 - Cada verbo tiene su propio predicado.
 - Más cercano al parsing sintáctico.
 - Anotaciones más simples.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**
- Argumentos standarizados:
 - Arg0: agent (giver).
 - Arg1: patient (thing given).
 - Arg2: instrument/attribute (entity given to).
 - Arg3: starting point/attribute.
 - Arg4: ending point.
 - ArgM: modifier.

orientado a los
verbos



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**
- Argumentos standarizados:
 - Arg0: agent (giver).
 - Arg1: patient (thing given).
 - Arg2: instrument/attribute (entity given to).
 - Arg3: starting point/attribute.
 - Arg4: ending point.
 - ArgM: modifier.

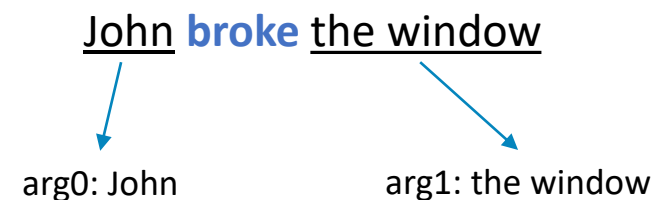
orientado a los
verbos

John broke the window



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**

orientado a los
verbos



- Argumentos standarizados:
 - Arg0: agent (giver).
 - Arg1: patient (thing given).
 - Arg2: instrument/attribute (entity given to).
 - Arg3: starting point/attribute.
 - Arg4: ending point.
 - ArgM: modifier.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**

orientado a los
verbos

- Argumentos standarizados:
 - Arg0: agent (giver).
 - Arg1: patient (thing given).
 - Arg2: instrument/attribute (entity given to).
 - Arg3: starting point/attribute.
 - Arg4: ending point.
 - ArgM: modifier.

Obama met him privately in the White House, on Thursday

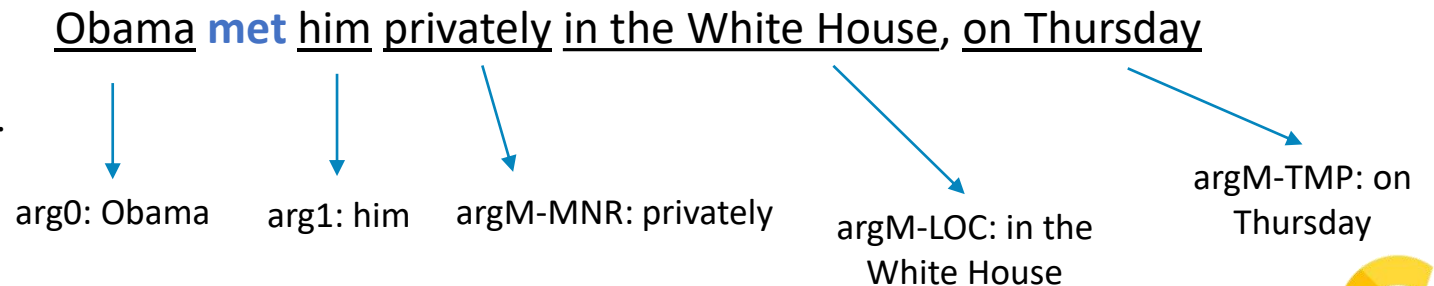


- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - ***PropBank*: Proto-roles.**

orientado a los
verbos

- Argumentos standarizados:

- Arg0: agent (giver).
- Arg1: patient (thing given).
- Arg2: instrument/attribute (entity given to).
- Arg3: starting point/attribute.
- Arg4: ending point.
- ArgM: modifier.



- Un intermedio entre el parsing y el análisis semántico completo.
- La mayoría de los enfoques actuales son supervisados, entrenados sobre:
 - *FrameNet*: frame-specific roles.
 - *PropBank*: Proto-roles.

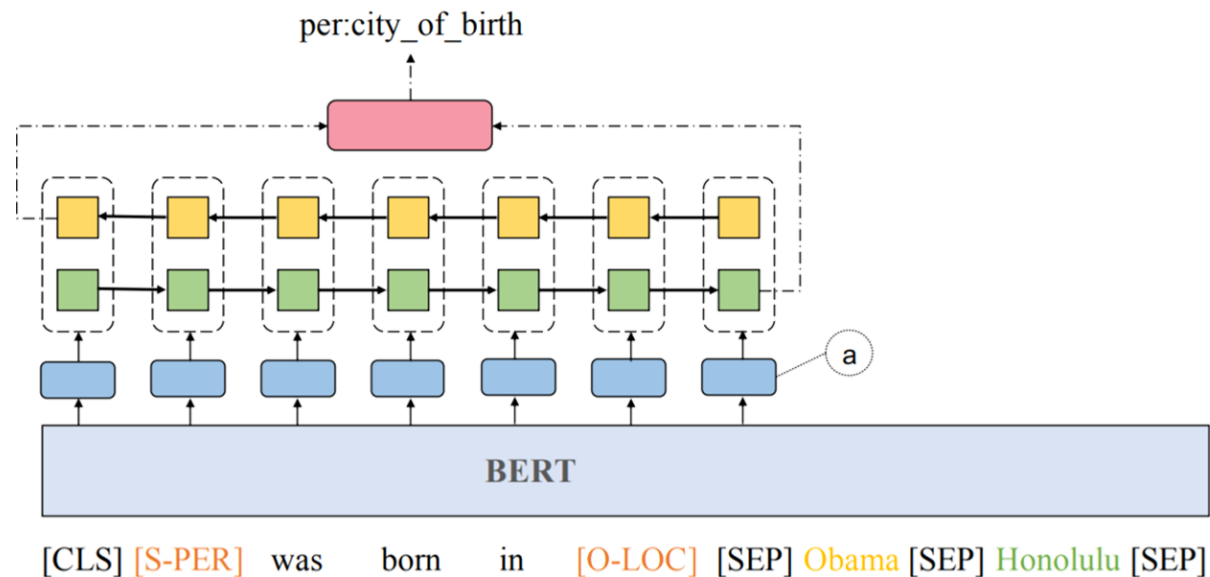
centrado en una noción abstracta que generaliza las descripciones de verbos similares (ej. "describir", "caracterizar") y sustantivos (ej. "descripción")

orientado a los verbos
- Pasos:
 1. Parsear la oración.
 2. Encontrar los predicados en la oración.
 3. Aplicar word sense disambuation sobre el predicado para determinar el sujeto.
 4. Identificar los argumentos semánticos.



Semantic Role Labeling: Embeddings → BERT

- Aprovechan BERT para la tarea de predecir la relación entre dos entidades, dada la oración y las entidades.
- Modifican la entrada a BERT.
 - En el BERT original la entrada era [CLS] oración [SEP] oración [SEP] ...
 - Ahora proponen: [[CLS] sentence [SEP] subject [SEP] object [SEP]]
 - En la oración, las entidades son enmascaradas.
 - También modifican la entrada para la predicción de predicados.



[Simple BERT Models for Relation Extraction and Semantic Role Labeling](#)

- El discurso trata del análisis de la estructura y el significado del texto más allá de una sola oración, haciendo conexiones entre palabras y oraciones.
- Involucra la resolución de referencias a elementos anteriores o posteriores en el discurso.
 - Por ejemplo, resolución de pronombres.
 - Uno de los problemas más difíciles. Enfoques recientes se basan en deep learning.

coreference information + the parse tree + named entity information → information extraction

- El discurso trata del análisis de la estructura y el significado del texto más allá de una sola oración, haciendo conexiones entre palabras y oraciones.
- Involucra la resolución de referencias a elementos anteriores o posteriores en el discurso.
 - Por ejemplo, resolución de pronombres.
 - Uno de los problemas más difíciles. Enfoques recientes se basan en deep learning.

coreference information + the parse tree + named entity information → information extraction

London is the capital and largest city of England and of the United Kingdom. Standing on the River Thames in the south-east of England, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

- El discurso trata del análisis de la estructura y el significado del texto más allá de una sola oración, haciendo conexiones entre palabras y oraciones.
- Involucra la resolución de referencias a elementos anteriores o posteriores en el discurso.
 - Por ejemplo, resolución de pronombres.
 - Uno de los problemas más difíciles. Enfoques recientes se basan en deep learning.

coreference information + the parse tree + named entity information → information extraction

`London` is the capital and largest city of England and of the United Kingdom. Standing on the River Thames in the south-east of England, `London` has been a major settlement for two millennia. `It` was founded by the Romans, who named `it` Londinium.

- El discurso trata del análisis de la estructura y el significado del texto más allá de una sola oración, haciendo conexiones entre palabras y oraciones.
- Involucra la resolución de referencias a elementos anteriores o posteriores en el discurso.
 - Por ejemplo, resolución de pronombres.
 - Uno de los problemas más difíciles. Enfoques recientes se basan en deep learning.

coreference information + the parse tree + named entity information → information extraction

London is the capital and largest city of England and of the United Kingdom. Standing on the River Thames in the south-east of England, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

- El nivel pragmático se ocupa del uso del conocimiento del mundo real y de comprender cómo eso impacta en el significado de lo que se está comunicando.
 - Cómo se usa el lenguaje para lograr los objetivos.
 - La influencia del contexto en el significado.
- Al analizar la dimensión contextual de los textos y consultas, se obtiene una representación más detallada.
- Este nivel involucra principalmente el procesamiento y la comprensión de las consultas de los usuarios integrando el historial y los objetivos del usuario, así como el contexto en el que se realiza la consulta.
- Los contextos pueden incluir hora y lugar.
- Facilita la conversación entre el sistema IR y los usuarios.
- Permite obtener el propósito sobre el cual se planea utilizar la información que se busca.



- Realizar un resumen de un texto más largo.
- Por qué hacer resúmenes automáticos?
 - Los resúmenes reducen el tiempo de lectura.
 - Los resúmenes facilitan el proceso de selección de textos.
 - El resumen automático mejora la efectividad de la indexación.
 - Los algoritmos de resumen automático son menos parciales que los humanos.
 - Los resúmenes personalizados son útiles en los sistemas de preguntas y respuestas, ya que proporcionan información personalizada.

Extractive Summary

- Identifica las oraciones o extractos importantes del texto y los reproduce textualmente como parte del resumen.
- Solo se usa texto existente en el proceso de resumen.
- El resumen puede resultar gramaticalmente extraño.

Abstractive Summary

- Emplea técnicas de NLP más potentes para interpretar texto y generar el resumen.
- Implica parafrasear y acortar partes del documento fuente.
- Crea nuevas frases y oraciones.
- Problema para deep learning, puede superar las inconsistencias gramaticales del método extractivo.
- Se desempeña mejor que la extracción, pero es más complicado.

- La extracción de información es la tarea de extraer automáticamente información estructurada de documentos no estructurados y/o semiestructurados.
- Ampliamente utilizado en:
 - Question Answering Systems.
 - Machine Translation.
 - Entity Extraction.
 - Event Extraction.
 - Verificación de información.



- La extracción de información es la tarea de extraer automáticamente información estructurada de documentos no estructurados y/o semiestructurados.
- Ampliamente utilizado en:
 - Question Answering Systems.
 - Machine Translation.
 - Entity Extraction.
 - Event Extraction.
 - Verificación de información.

London is the capital and largest city of England and of the United Kingdom. Standing on the River Thames in the south-east of England, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

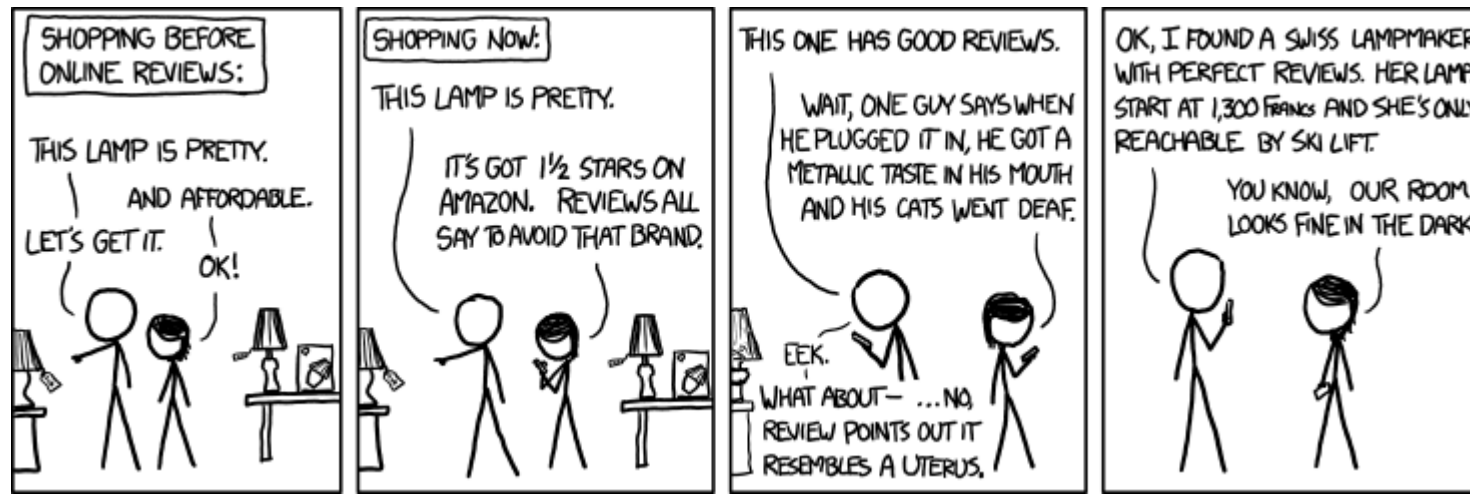


- The capital and most populous city of England and the United Kingdom.
- A major settlement for two millennia.



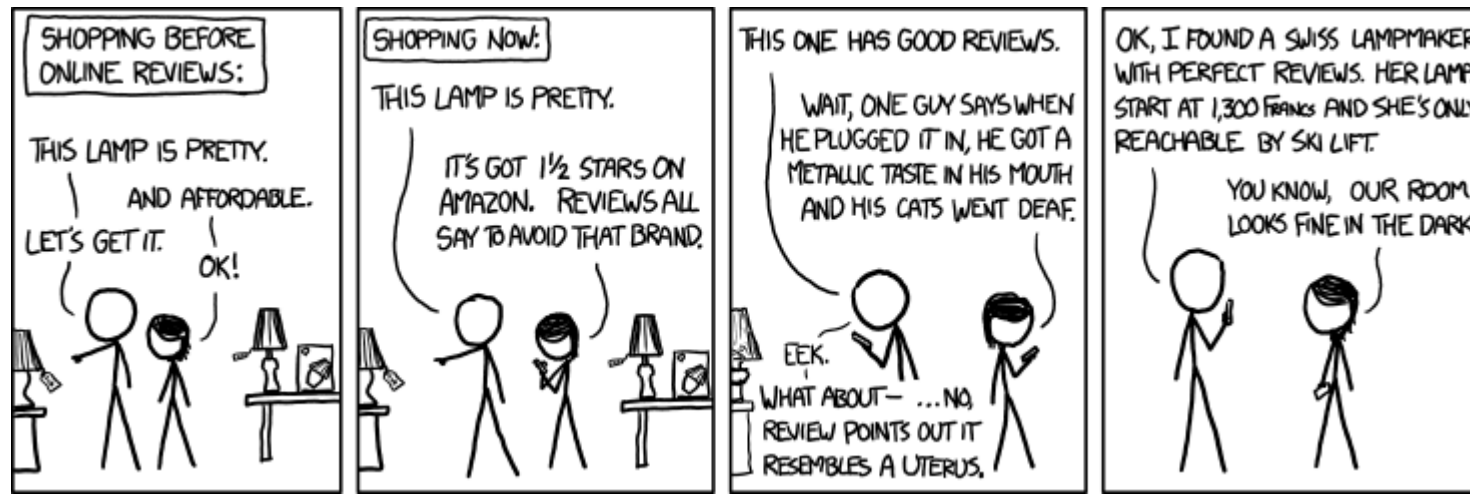
- Además de su significado, las palabras también tienen asociadas connotaciones o significados afectivos.
- Se relaciona con las opiniones, sentimientos o emociones.
- Tres dimensiones:
 - *Valencia*. Cuán placentero es un estímulo.
 - +: happy, pleased, satisfied, contented, hopeful
 - -: unhappy, annoyed, unsatisfied, melancholic, despaired, or bored
 - *Arousal*. La intensidad de la emoción provocada por el estímulo.
 - +: stimulated, excited, frenzied, wide-awake, or aroused
 - -: relaxed, calm, sluggish, dull, sleepy, or unaroused
 - *Dominance*. El grado de control ejercido por el estímulo.
 - +: in control, influential, important, dominant, autonomous, or controlling
 - -: controlled, influenced, cared-for, awed, submissive, or guided





- El análisis de sentimientos (también conocido como minería de opinión) intenta identificar y extraer sentimientos dentro de un texto determinado en blogs, reseñas, redes sociales, foros, noticias, etc.
- Los modelos de análisis de sentimientos detectan la polaridad dentro de un texto (por ejemplo, una opinión positiva o negativa), ya sea un texto completo, un párrafo, una oración o una cláusula.
- Comprender los sentimientos de las personas es esencial para las empresas, ya que los clientes pueden expresar abiertamente sus pensamientos y sentimientos.





- Tradicionalmente detección con léxicos o modelos supervisados. (algunos léxicos ya vienen integrados en NLTK)
 - VADER lexicon (incluye emojis, dice que fue creado especialmente para social media)
 - SentiWordNet (basado en WordNet, asigna un valor de polaridad negativa y positiva a cada synset).
 - TextBlob lexicon (también asociado a WordNet).
 - AFFIN lexicon (incluye emoticons).
- Una de las desventajas del uso de léxicos es que las personas pueden expresar sus emociones de diferentes maneras.
 - Algunas palabras que típicamente expresan enojo, como “bad” or “kill”, también pueden expresar felicidad.

Discurso & Pragmática

Emotion Analysis

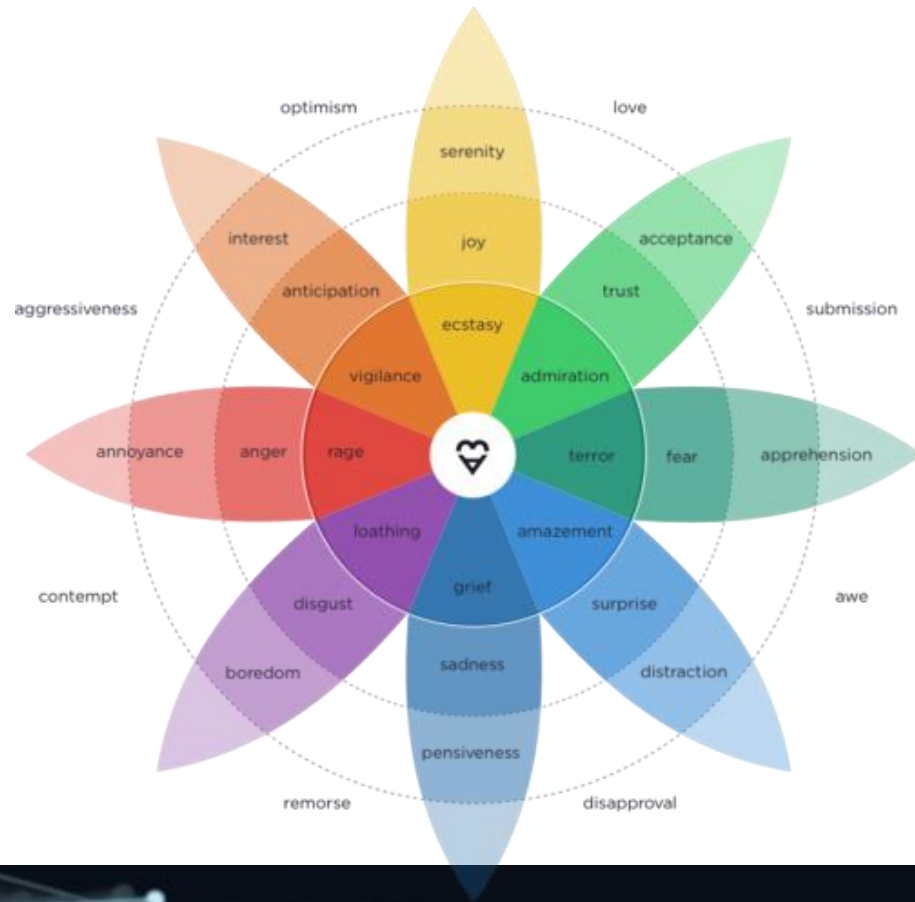
- Tiene como objetivo detectar emociones, como la felicidad, la frustración, la ira, la tristeza, etc.
 - Se suele decir que puede ser la causa de los sentimientos.



Discurso & Pragmática

Emotion Analysis

- Tiene como objetivo detectar emociones, como la felicidad, la frustración, la ira, la tristeza, etc.
 - Se suele decir que puede ser la causa de los sentimientos.



Discurso & Pragmática

Emotion Analysis

- Tiene como objetivo detectar emociones, como la felicidad, la frustración, la ira, la tristeza, etc.
 - Se suele decir que puede ser la causa de los sentimientos.
- También hay léxicos específicos para esta tarea:
 - EmoLex (comprende varios léxicos creados tanto manualmente como de forma automática, hashtags)
 - Para cada término le asignan un sentimiento, y dentro de ese sentimiento el nivel de la emoción.
 - SentiSense
 - Asocia sentimiento y emoción a synsets de WordNet.
 - Disponible también en Español.



Discurso & Pragmática

Emotion Analysis

- Tiene como objetivo detectar emociones, como la felicidad, la frustración, la ira, la tristeza, etc.
 - Se suele decir que puede ser la causa de los sentimientos.
- También hay léxicos específicos para esta tarea:
 - EmoLex (comprende varios léxicos creados tanto manualmente como de forma automática, hashtags)
 - Para cada término le asignan un sentimiento, y dentro de ese sentimiento el nivel de la emoción.
 - SentiSense
 - Asocia sentimiento y emoción a synsets de WordNet.
 - Disponible también en Español.
- Se puede extender a otros estados afectivos:
 - *Humor*: cheerful, gloomy, irritable, listless, depressed.
 - *Actitud* o predisposición a otros humanos u objetos: distant, cold, warm, supportive, contemptuous, friendly, liking, loving, hating, valuing, desiring
 - *Personalidad*: nervous, anxious, reckless, morose, hostile, jealous.

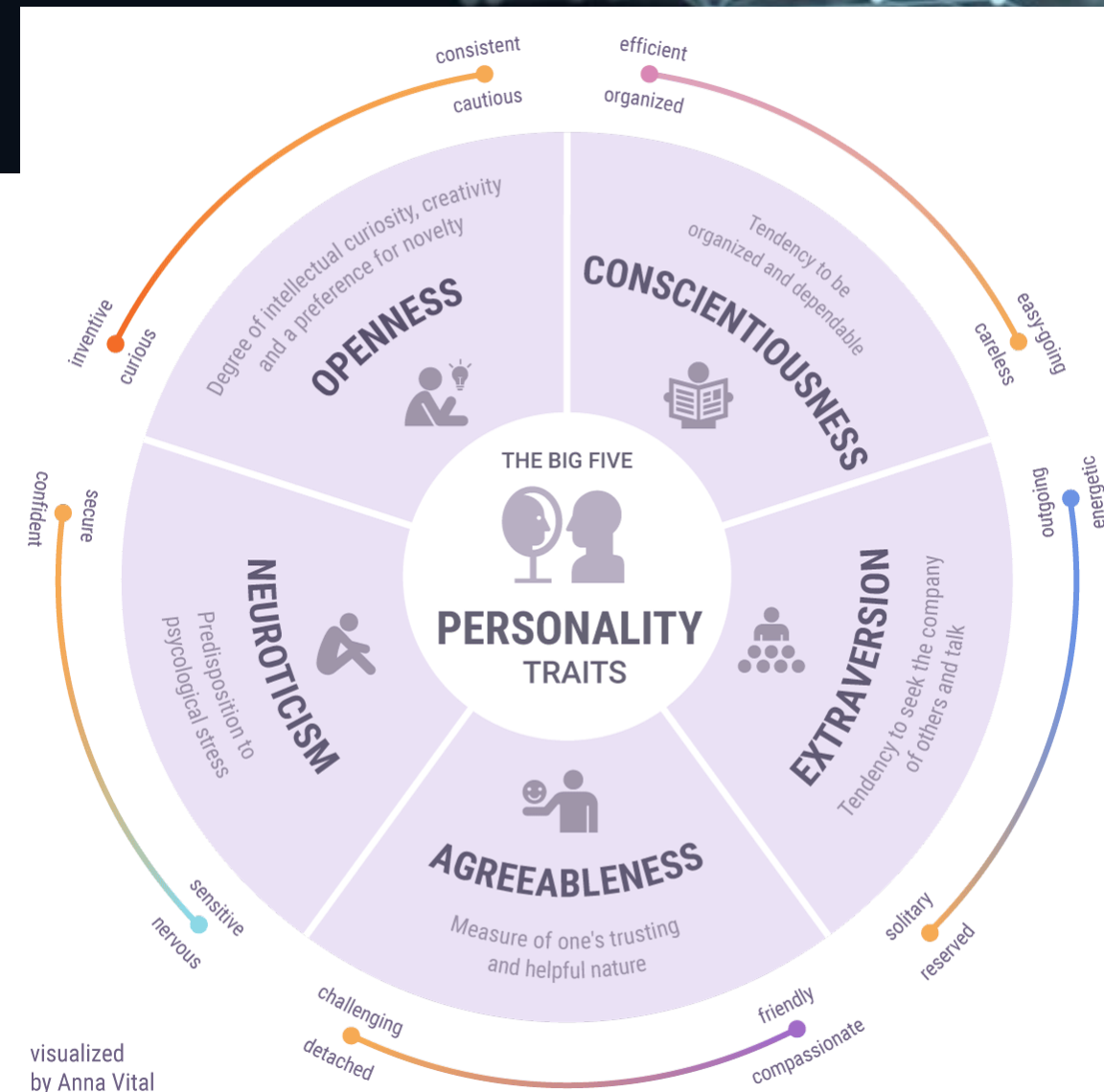


Discurso & Pragmática

Personality Analysis

- Personalidad es uno de los más estudiados.
 - Big 5 Personality traits

Extraversion vs Introversion	
Sociable, asertivo	Distante, reservado, tímido
Emotional stability vs Neuroticism	
Calmo, sin emociones	Inseguro, ansioso
Agreeableness vs Disagreeable	
Amistoso, cooperativo	Antagonista, “busca pelea”
Conscientiousness vs Unconscientious	
Auto-disciplinado, organizado	Ineficiente, descuidado
Openness to experience	
intelectual, perspicaz	Superficial, sin imaginación



[Personality Structure Emergence Of The Five-factor Model](#)



- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
- Por qué es difícil?
 - Orden de las palabras.
 - Aceptaciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.



- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
- Por qué es difícil?
 - Orden de las palabras.
 - Aceptaciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.

Quiero ir a la playa más bonita.

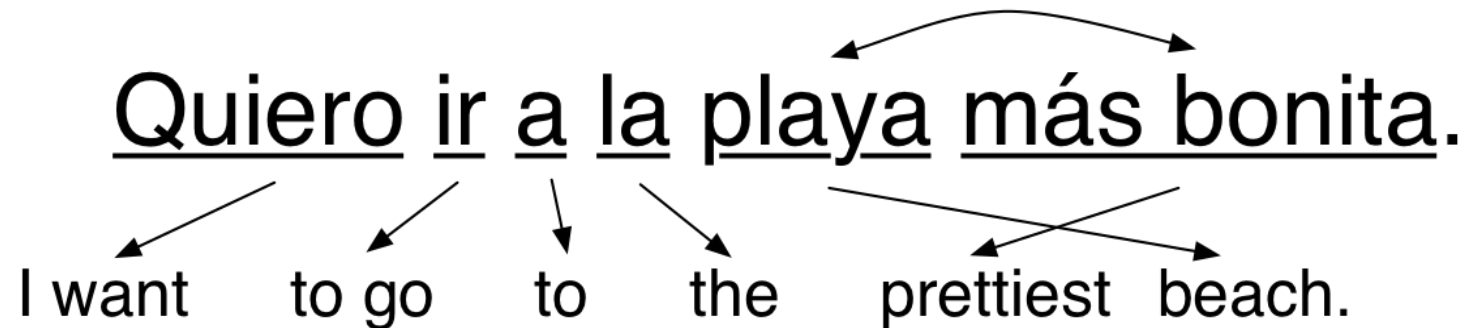


- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
- Por qué es difícil?
 - Orden de las palabras.
 - Aceptaciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.

Quiero ir a la playa más bonita.

I want to go to the beach more pretty.

- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
- Por qué es difícil?
 - Orden de las palabras.
 - Aceptaciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.



- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
- Por qué es difícil?
 - Orden de las palabras.
 - Acepciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.
- Diversos enfoques:
 - Traducción palabra por palabra.
 - Transferencia sintáctica.
 - Enfoques interlinguales.
 - Traducción basada en ejemplos.
 - Traducción basada en estadística.
 - Redes neuronales.



- La traducción automática es la tarea de convertir automáticamente de un idioma natural a otro, preservando el significado del texto de entrada y produciendo texto fluido en el idioma de salida.
 - Por qué es difícil?
 - Orden de las palabras.
 - Acepciones de las palabras.
 - Pronombres.
 - Tiempos verbales.
 - Idioms.
 - Diversos enfoques:
 - Traducción palabra por palabra.
 - Transferencia sintáctica.
 - Enfoques interlinguales.
 - Traducción basada en ejemplos.
 - Traducción basada en estadística.
 - Redes neuronales.
- Google Translate (from 2016).
 - Microsoft Translate (from 2016).
 - Translation on Facebook.
 - OpenNMT: An open-source neural machine translation system.



Detección de tópicos

- La detección de tópicos provee modelos para organizar, comprender y resumir (extraer palabras relevantes) grandes cantidades de textos.
- Son métodos para encontrar grupos de palabras (los tópicos) dentro de una colección de textos que mejor los representan.

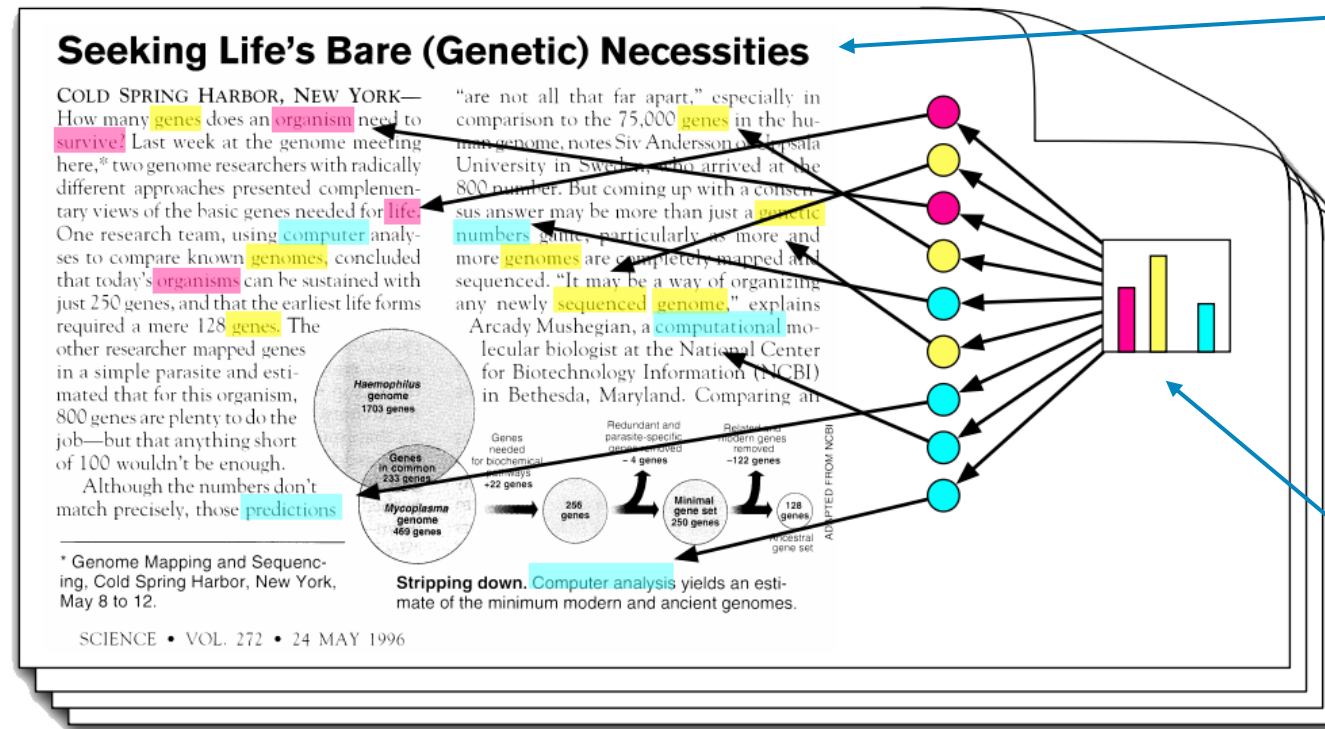
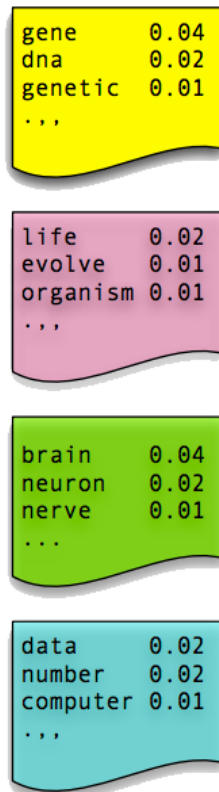


Semántica, Pragmática, Discurso

Detección de tópicos

- La detección de tópicos provee modelos para organizar, comprender y resumir (extraer palabras relevantes) grandes cantidades de textos.
- Son métodos para encontrar grupos de palabras (los tópicos) dentro de una colección de textos que mejor los representan.

Tópicos
(identificados
por grupos de
palabras,
importancia de
cada palabra)



Texto

Proporciones
en las que
aparece cada
tópico

Dynamic topic model D.M. Blei 2012



Detección de tópicos

- La detección de tópicos provee modelos para organizar, comprender y resumir (extraer palabras relevantes) grandes cantidades de textos.
- Son métodos para encontrar grupos de palabras (los tópicos) dentro de una colección de textos que mejor los representan.
- Se basan en que cada texto se puede describir mediante una distribución de tópicos, y cada tópico puede descripto por una distribución de palabras.
- Métodos tradicionales basados en distribuciones probabilísticas y factorización de matrices.
 - Hay más que estos.

LSA

LDA



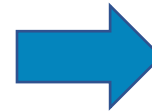
LDA

Latent Dirichlet Allocation

- Método basado en distribución de probabilidades.

Las palabras que
pertenecen a un
documento

(lo que ya sabemos)



Las palabras que
pertenecen a un
tópico

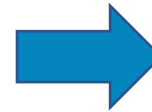
(lo que queremos calcular)

LDA

Latent Dirichlet Allocation

- Método basado en distribución de probabilidades.

Las palabras que
pertenecen a un
documento



Las palabras que
pertenecen a un
tópico

- No es relevante el orden de las palabras en el texto.
- No es relevante el rol grammatical de las palabras.
- Palabras que aparecen en la mayoría de los documentos pueden eliminarse.
 - ~ 80-90%
- El número de tópicos a encontrar debe conocerse.
- Una misma palabra puede aparecer en varios tópicos.
 - Representa a los textos como una mezcla de tópicos.

LDA

Las palabras que
pertenecen a un
documento



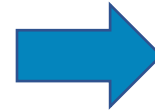
Las palabras que
pertenecen a un
tópico

1. Me gusta comer broccoli y bananas.
2. Tomé un smoothie de banana con espinaca en el desayuno.
3. Los erizos y los perros son bonitos.
4. La vecina adoptó un perro ayer.
5. Mirá este erizo bonito comiendo una rodaja de banana.



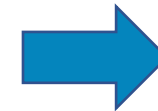
LDA

Las palabras que
pertenecen a un
documento



Las palabras que
pertenecen a un
tópico

1. Me gusta comer broccoli y bananas.
2. Tomé un smoothie de banana con espinaca en el desayuno.
3. Los erizos y los perros son bonitos.
4. La vecina adoptó un perro ayer.
5. Mirá este erizo bonito comiendo una rodaja de banana.



Oraciones 1 y 2: 100% tópico A.
Oraciones 3 y 4: 100% tópico B.
Oración 5: 60% tópico A, 40% tópico B.

Tópico A: 30% broccoli, 15% bananas, 10% desayuno, 10% comiendo ...

Tópico B: 35% erizo, 20% perro ...



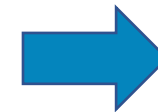
LDA

Las palabras que
pertenecen a un
documento



Las palabras que
pertenecen a un
tópico

1. Me gusta comer broccoli y bananas.
2. Tomé un smoothie de banana con espinaca en el desayuno.
3. Los erizos y los perros son bonitos.
4. La vecina adoptó un perro ayer.
5. Mirá este erizo bonito comiendo una rodaja de banana.



Oraciones 1 y 2: 100% tópico A.
Oraciones 3 y 4: 100% tópico B.
Oración 5: 60% tópico A, 40% tópico B.

Comida

Tópico A: 30% broccoli, 15% bananas, 10% desayuno, 10% comiendo ...

Animales bonitos

Tópico B: 35% erizo, 20% perro ...



LDA

- Asume que los documentos se crean de la siguiente forma:
- Cuando se está escribiendo un document:
- Se decide la cantidad de palabras que va a tener.
 - Por ejemplo, mediante una distribución Poisson.
- Elegir una mezcla de los tópicos para el documento.
 - De acuerdo a una distribución de Dirichlet sobre un conjunto fijo de K tópicos.
 - Por ejemplo, un documento va a estar asociado en 2/3 a animales bonitos y 1/3 a comida.
- Generar cada palabra en el document como:
 - Elegir un tópico (de acuerdo a las distribuciones seleccionadas).
 - Usar el tópico para generar la palabra (de acuerdo a la distribución del tópico).
 - Por ejemplo, si el tópico es comida, se podría generar la palabra “brócoli” con una probabilidad de 0.3, la palabra “banana” con una probabilidad de 0.15 ...
- Asumiendo este modelo, LDA intenta hacer backtracking desde los documentos para encontrar un set de tópicos que es probable que haya generado la colección.

LDA

Latent Dirichlet Allocation

1. Crear una representación inicial de los K tópicos asignando las palabras w en el/los texto/s de forma random a los tópicos.
2. Para cada palabra w calcular:
 - a) $P(\text{tópico } t \mid \text{texto } d) = \text{proporción de palabras en el texto } d \text{ que actualmente se encuentran asignadas a } t.$
 - b) $P(\text{palabra } w \mid \text{tópico } t) = \text{proporción de palabras asignadas al tópico } t \text{ sobre el total.}$
3. Actualizar la probabilidad de una palabra w de pertenecer a un tópico como
$$P(\text{palabra } w \mid \text{tópico } t, \text{texto } d) = P(\text{tópico } t \mid \text{texto } d) * P(\text{palabra } w \mid \text{tópico } t)$$
4. Repetir iterativamente hasta convergencia.
5. Elegir las N palabras con mayor probabilidad de pertenecer a cada tópico.

LDA

Latent Dirichlet Allocation

1. Crear una representación inicial de los K tópicos asignando las palabras w en el/los texto/s de forma random a los tópicos.
2. Para cada palabra w calcular:
 - a) $P(\text{tópico } t \mid \text{texto } d) = \text{proporción de palabras en el texto } d \text{ que actualmente se encuentran asignadas a } t$. Trata de capturar para un texto d cuantas de sus palabras pertenecen al tópico, a mayor cantidad de palabras de d que pertenezcan a t , mayor probabilidad de que w efectivamente pertenezca a t .
 - b) $P(\text{palabra } w \mid \text{tópico } t) = \text{proporción de palabras asignadas al tópico } t \text{ sobre el total}$. Trata de capturar cuántos textos pertenecen a t debido a la aparición de w . Si una palabra tiene una alta probabilidad de pertenecer a un tópico, todos los textos conteniendo a w se encontrarán relacionados con t . De forma similar, si no es muy probable que w esté en t , los textos que contengan a w tendrán una baja probabilidad de pertenecer a t , dado que el resto de las palabras pertenecerán a otros tópicos.
3. Actualizar la probabilidad de una palabra w de pertenecer a un tópico como
$$P(\text{palabra } w \mid \text{tópico } t, \text{texto } d) = P(\text{tópico } t \mid \text{texto } d) * P(\text{palabra } w \mid \text{tópico } t)$$
4. Repetir iterativamente hasta convergencia.
5. Elegir las N palabras con mayor probabilidad de pertenecer a cada tópico.

LDA

Latent Dirichlet Allocation

Podemos hacer una selección de las palabras, no incluir todas. Aplicar pre-procesamiento. Eliminar stopwords!

No existe el K mágico

1. Crear una representación inicial de los K tópicos asignando las palabras w en el/los texto/s de forma random a los tópicos.
2. Para cada palabra w calcular:
 - a) $P(\text{tópico } t \mid \text{texto } d) = \text{proporción de palabras en el texto } d \text{ que actualmente se encuentran asignadas a } t$. Trata de capturar para un texto d cuantas de sus palabras pertenecen al tópico, a mayor cantidad de palabras de d que pertenezcan a t , mayor probabilidad de que w efectivamente pertenezca a t .
 - b) $P(\text{palabra } w \mid \text{tópico } t) = \text{proporción de palabras asignadas al tópico } t \text{ sobre el total}$. Trata de capturar cuántos textos pertenecen a t debido a la aparición de w . Si una palabra tiene una alta probabilidad de pertenecer a un tópico, todos los textos conteniendo a w se encontrarán relacionados con t . De forma similar, si no es muy probable que w esté en t , los textos que contengan a w tendrán una baja probabilidad de pertenecer a t , dado que el resto de las palabras pertenecerán a otros tópicos.
3. Actualizar la probabilidad de una palabra w de pertenecer a un tópico como
$$P(\text{palabra } w \mid \text{tópico } t, \text{texto } d) = P(\text{tópico } t \mid \text{texto } d) * P(\text{palabra } w \mid \text{tópico } t)$$
4. Repetir iterativamente hasta convergencia.
5. Elegir las N palabras con mayor probabilidad de pertenecer a cada tópico.



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.



<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.



1. Crear una representación inicial de los K tópicos asignando las palabras w en el/los texto/s de forma random a los tópicos.
 - Siguiendo algún tipo de distribución, por ejemplo, uniforme.

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.



1. Crear una representación inicial de los K tópicos asignando las palabras w en el/los texto/s de forma random a los tópicos.
 - Siguiendo algún tipo de distribución, por ejemplo, uniforme.

Asignación de palabras a los tópicos

	lechuza	gato	Alice	Harry
Tópico 1		X	X	
Tópico 2	X			X

Frecuencia de palabras en documento por tópico

	Tópico 0	Tópico 1
Documento 1	2	2
...
Documento n

Frecuencia de palabras por tópico

	Tópico 0	Tópico 1
lechuza	5	30
Gato	20	15
Alice	8	3
Harry	0	25

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

2. Cálculo de probabilidad por palabra.

- Arrancamos con el documento 1, lechuza.
- Asumimos que todas las otras asignaciones son correctas, menos la de esta lechuza.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

	Tópico 1	Tópico 2
Documento 1	2	2 1
...
Documento n

	Tópico 1	Tópico 2
lechuza	5	30 29
Gato	20	15
Alice	8	3
Harry	0	25

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

2. Cálculo de probabilidad por palabra.

- Arrancamos con el documento 1, lechuza.
- Asumimos que todas las otras asignaciones son correctas, menos la de esta lechuza.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

	Tópico 1	Tópico 2
lechuza	5	30 29
Gato	20	15
Alice	8	3
Harry	0	25

	Tópico 1	Tópico 2
Documento 1	2	2 1
...
Documento n

	Tópico 1	Tópico 2
Documento 1	2	2 1
...
Documento n

$$P(\text{tópico } t \mid \text{documento } d) = \frac{\text{cantidad palabras en documento } d \text{ en el tópico } t + \alpha}{\text{cantidad de palabras en documento } d - 1 + k * \alpha}$$

$$P(\text{tópico } 2 \mid \text{documento } 1) = \frac{1 + 0.1}{4 - 1 + 2 * 0.1} = \frac{1.1}{3.2} = 0.34$$

<https://lettier.com/projects/lda-topic-modeling/>

LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

2. Cálculo de probabilidad por palabra.

- Arrancamos con el documento 1, lechuza.
- Asumimos que todas las otras asignaciones son correctas, menos la de esta lechuza.

	Tópico 1	Tópico 2
lechuza	5	30 29
Gato	20	15
Alice	8	3
Harry	0	25

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

	Tópico 1	Tópico 2
Documento 1	2	1 1
...
Documento n

	Tópico 1	Tópico 2
lechuza	5	30 29
Gato	20	15
Alice	8	3
Harry	0	25

$$P(\text{palabra } w \mid \text{tópico } t) = \frac{\text{frecuencia palabra } w \text{ en el tópico } t + \beta}{\text{frecuencia de palabras en tópico } t + |\text{vocabulario}| * \beta}$$

$$P(\text{palabra lechuza} \mid \text{tópico } 2) = \frac{29 + 0.5}{72 + 4 * 0.5} = \frac{29.5}{74} = 0.398$$

<https://lettier.com/projects/lda-topic-modeling/>

LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

3. Actualizar la probabilidad de una palabra w de pertenecer a un tópico.

$$P(\text{palabra } w \mid \text{tópico } t, \text{documento } d) = P(\text{tópico } t \mid \text{documento } d) * P(\text{palabra } w \mid \text{tópico } t)$$

$$P(\text{palabra lechuza} \mid \text{tópico 2, documento 1}) = 0.34 * 0.398 = 0.135$$

- Esto se calcula para los distintos tópicos.
- Se asigna la palabra al tópico para el cual la probabilidad sea máxima.

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

	Tópico 2	Tópico 1	Tópico 2	Tópico 1
Documento 1	lechuza	gato	Alice	Harry

3. Actualizar la probabilidad de una palabra w de pertenecer a un tópico.

$$P(\text{palabra } w \mid \text{tópico } t, \text{documento } d) = P(\text{tópico } t \mid \text{documento } d) * P(\text{palabra } w \mid \text{tópico } t)$$

$$P(\text{palabra lechuza} \mid \text{tópico 2, documento 1}) = 0.34 * 0.398 = 0.135$$

- Esto se calcula para los distintos tópicos.
- Se asigna la palabra al tópico para el cual la probabilidad sea máxima.
- Repetimos esto para todas las palabras.
- Repetimos todo hasta convergencia.
 - No se detectan más cambios.
 - Se cumple número de iteraciones.

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

- Cuando terminemos, vamos a tener las matrices finales de tópicos por palabra y documento por tópico.
 - Podemos calcular la relevancia de cada palabra por tópico y de cada tópico por documento.

	Tópico 1	Tópico 2
Documento 1	3	1
Documento 2	15	26
Documento 3	19	43

$$\theta = P(\text{tópico } t \mid \text{documento } d) = \frac{\text{frecuencia palabras en tópico } t \text{ para documento } d + \alpha}{\text{frecuencia palabras en documento } d + k * \alpha}$$

	Tópico 1	Tópico 2
lechuza	6	29
Gato	21	14
Alice	9	2
Harry	0	25

$$\phi = P(\text{palabra } w \mid \text{tópico } t) = \frac{\text{frecuencia palabra } w \text{ en tópico } t + \beta}{\text{frecuencia total para tópico } t + |\text{vocabulario}| * \beta}$$

<https://lettier.com/projects/lda-topic-modeling/>



LDA

- $\alpha = 0.1$
- $\beta = 0.5$
- $k = 2$
- Vamos a calcular:
 - Relación entre words y tópicos.
 - Relación entre documentos y tópicos.

- Cuando terminemos, vamos a tener las matrices finales de tópicos por palabra y documento por tópico.
 - Podemos calcular la relevancia de cada palabra por tópico y de cada tópico por documento.

	Tópico 1	Tópico 2
Documento 1	3	1
Documento 2	14	26
Documento 3	19	43

	Tópico 1	Tópico 2
Documento 1	0.738	0.261
Documento 2	0.366	0.633
Documento 3	0.307	0.692

	Tópico 1	Tópico 2
lechuza	6	29
Gato	21	14
Alice	9	2
Harry	0	25

	Tópico 1	Tópico 2
lechuza	0.171	0.409
Gato	0.565	0.201
Alice	0.25	0.034
Harry	0.013	0.354

[om/projects/lda-topic-modeling/](https://github.com/projects/lda-topic-modeling/)



LDA

Parámetros

- α
 - Controla la “mezcla” de tópicos en los documentos.
 - Valores bajos, los documentos tendrán una menor mezcla de tópicos.
 - Valores altos, los documentos parecerán más mezclados.
- β
 - Controla la distribución de palabras por tópico.
 - Valores bajos, los tópicos tendrán menos palabras.
 - Valores altos, los tópicos tendrán más palabras.
 - Los tópicos tendrán una mayor cantidad de palabras communes.
- Idealmente, queremos que los documentos tengan una cantidad relativamente baja de tópicos y que las palabras pertenezcan a pocos tópicos.
 - Valores menores a 1



LDA

- Cómo evaluar los tópicos que encontramos?
 - Son interpretables?
 - Son únicos/disjuntos?
 - Son exhaustivos? Están todos los textos representados?
- “A ojo”.
 - Cuáles son las palabras más relevantes por tópico?
 - Cómo es la relación tópicos/documentos?
 - Qué es un tópico?
- **Métricas de evaluación intrínsecas.**
 - Capturar la semántica del modelo.
 - Evaluar la interpretabilidad.
- **Evaluación en el contexto de una tarea.**
 - Sirve el modelo para otra tarea como clasificación?



Detección de tópicos: Cómo evaluar?

- Comúnmente se utilizan dos métricas.
- **Perplexity.**
 - Cuán “sorprendido” (perplejo) está un modelo por datos que no ha visto antes.
 - Se mide como el log-likelihood normalizado sobre una partición de test.
 - Cuán probable son los datos nuevos en función del modelo?
 - No suele correlacionarse con interpretaciones humanas.
- **Topic coherence.**
 - Determina el nivel de semejanza semántica entre las palabras con mayor score en el tópico.
 - Varias formas de medirlo.



Detección de tópicos: Cómo evaluar?

- Comúnmente se utilizan dos métricas.
- **Perplexity.**
 - Cuán “sorprendido” (perplejo) está un modelo por datos que no ha visto antes.
 - Se mide como el log-likelihood normalizado sobre una partición de test.
 - Cuán probable son los datos nuevos en función del modelo?
 - No suele correlacionarse con interpretaciones humanas.
- **Topic coherence.**
 - Determina el nivel de semejanza semántica entre las palabras con mayor score en el tópico.
 - Varias formas de medirlo.

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{PMI}(w_i, w_j) \quad (1)$$

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (2)$$

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (4)$$

$$C_S = \frac{P(w_1, \dots, w_N)}{\prod_{i=1}^N P(w_i)}$$

$$C_O = \frac{P(w_1, \dots, w_N)}{P(w_1 \vee \dots \vee w_N)}$$

$$C_F = \frac{\sum_{i=1}^N \sum_{j=1}^{2^{N-1}-1} m_f(w_i, S(i)_j)}{N \cdot (2^{N-1} - 1)} \quad (15)$$

$$m_f(w_i, S(i)_j) = \frac{P(W_i|S(i)_j) - P(W_i|\neg S(i)_j)}{P(W_i|S(i)_j) + P(W_i|\neg S(i)_j)} \quad (16)$$

<https://palmetto.demos.dice-research.org/>

Exploring the Space of Topic Coherence Measures



LDA

Latent Dirichlet Allocation

- Cómo evaluar los tópicos que encontramos?
 - Son interpretables?
 - Son únicos/disjuntos?
 - Son exhaustivos? Están todos los textos representados?

Pros

- Es rápido.
 - Depende de cantidad de textos, tamaño del vocabulario/textos.
- Es intuitivo.
- Puede determinar los tópicos para documentos nuevos.
 - Considerar dominio.

Cons

- Puede no ser fácil la parametrización.
 - Tiene algunos parámetros que no consideramos que pueden afectar la calidad de los tópicos.
- Si bien los tópicos son intuitivos, puede requerir un extra de interpretación humana.

LSA

Latent Semantic Analysis

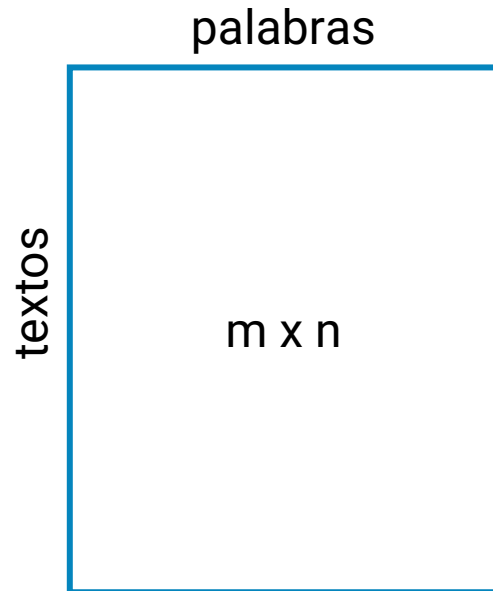
- La idea es representar los textos en una menor dimensión, obtenidas mediante la factorización de matrices.
- Se basa en la idea de que las palabras aparecerán en textos que muestran algún rasgo de semejanza, si tienen un significado parecido.
- Requiere definir el número K de tópicos (dimension de la matriz).



LSA

Latent Semantic Analysis

- La idea es representar los textos en una menor dimensión, obtenidas mediante la factorización de matrices.
- Se basa en la idea de que las palabras aparecerán en textos que muestran algún rasgo de semejanza, si tienen un significado parecido.
- Requiere definir el número K de tópicos (dimension de la matriz).



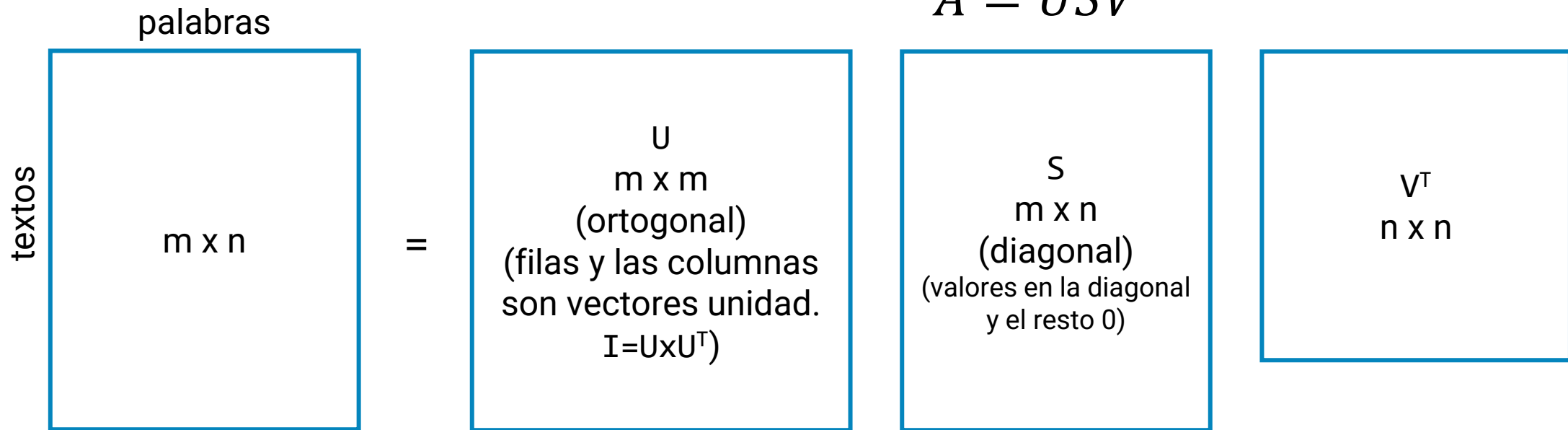
- Construir la matriz de textos por palabras.
- Cada fila representa un texto y cada columna una palabra.
- Se pueden utilizar diferentes ponderaciones.
 - Por lo general, se utiliza TF-IDF.
- Se pueden incluir todas las palabras, o solo algunas.
 - Aplicar pre-procesamiento. Eliminar stopwords.
- Matriz rara, ruidosa y redundante.

LSA

Latent Semantic Analysis

- Vamos a utilizar SVD (Singular Value Decomposition)
- Cada fila de U tiene la representación vectorial de los textos.
- V tiene la representación vectorial de las palabras.

$$A = USV^T$$



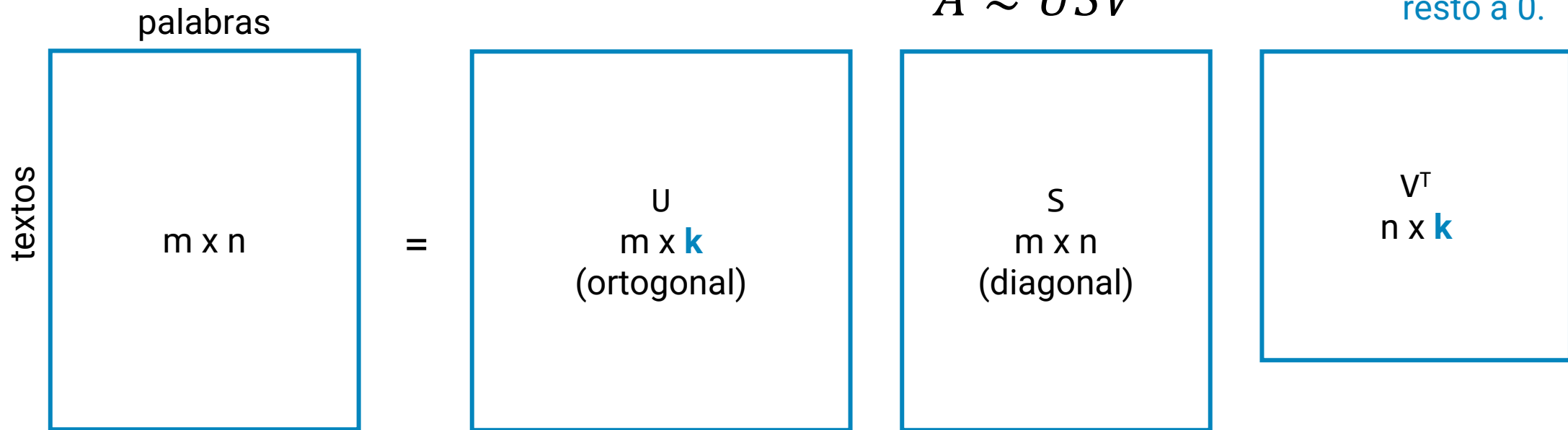
LSA

Latent Semantic Analysis

- Vamos a utilizar SVD (Singular Value Decomposition)
- Cada fila de U tiene la representación vectorial de los textos.
 - Los vectores se truncarán a longitud K (la cantidad de tópicos).
- V tiene la representación vectorial de las palabras.
 - También se trunca a longitud K.

Vamos a quedarnos con los K valores más grandes de U, el resto a 0.

$$A \approx USV^T$$



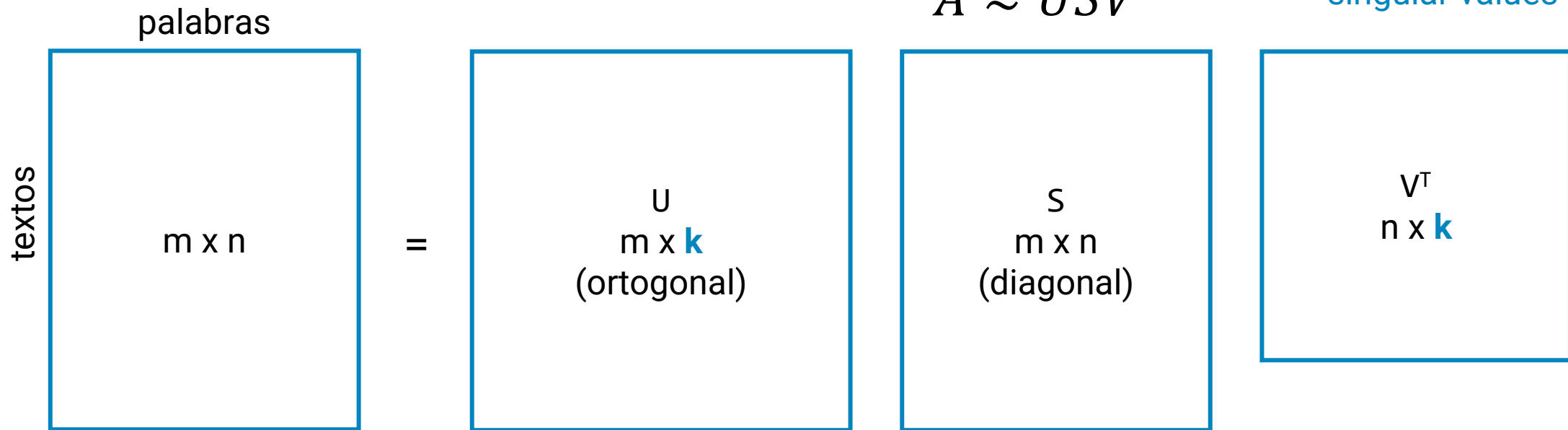
LSA

Latent Semantic Analysis

- Vamos a utilizar SVD (Singular Value Decomposition)
- Cada fila de U tiene la representación vectorial de los textos.
 - Los vectores se truncarán a longitud K (la cantidad de tópicos).
- V tiene la representación vectorial de las palabras.
 - También se trunca a longitud K.

Vamos a calcular
para los K mayores
singular values

$$A \approx USV^T$$



LSA

Latent Semantic Analysis

- Puede no ser fácil elegir K.
- Las representaciones no son intuitivas como en LDA.
 - No sabemos cuáles son los tópicos.
 - Los valores en los vectores son arbitrarios.
- Se requiere de una gran cantidad de textos.
- Una representación no del todo eficiente.
- No tan rápido como LDA.
- Puede no capturar completamente las relaciones de sinonimia.



Resumen!

- **Análisis semántico.** Analiza el significado del texto. Mapea las estructuras sintácticas y los objetos respecto al dominio de la tarea.
- **Discurso.** Estudia el análisis de la estructura y significado del texto más allá de una única oración, encontrando relaciones entre ellas.
- **Pragmática.** Se reinterpreta el texto respecto a lo que realmente significa. Involucra estudiar aspectos del lenguaje que requieren conocimiento externo.

Semantic
Analysis



Discourse
Integration



Pragmatic
Analysis

Word Sense
Disambiguation

Dada una palabra con múltiples acepciones, determinar cuál es la adecuada.

Named Entity
Recognition

Busca identificar y clasificar elementos en el texto en categorías predefinidas de: personas, organizaciones, lugares...

Semantic Role
Labeling

Para cada clausula, determina el rol semántico por cada frase nominal que hace de argumento del verbo.

Extracción de
Información

Extraer automáticamente información estructurada de documentos no estructurados y/o semiestructurados.

Detección de
Estados
Afectivos

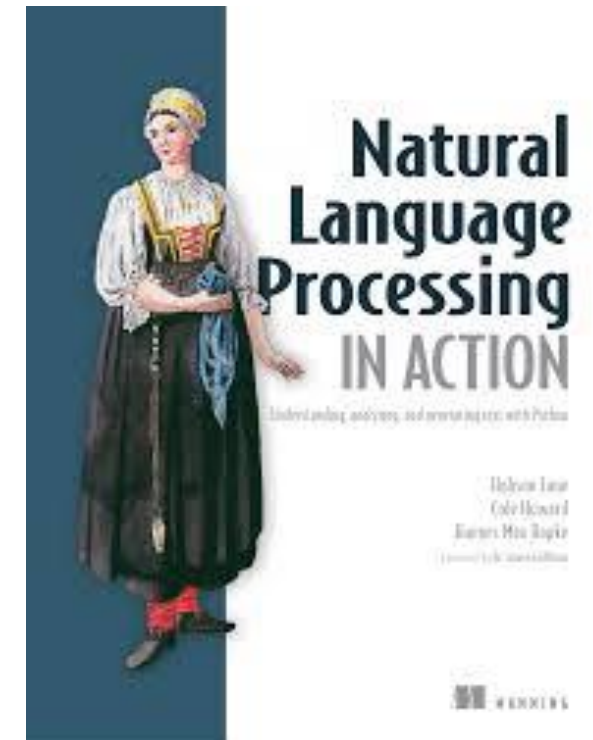
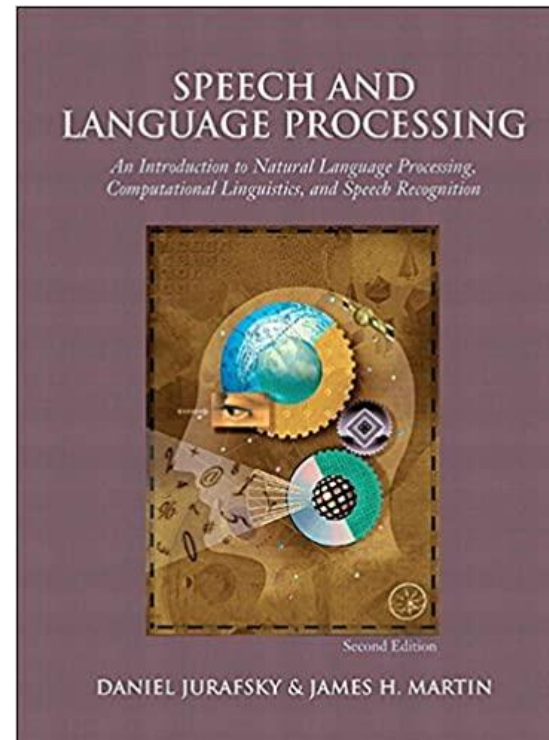
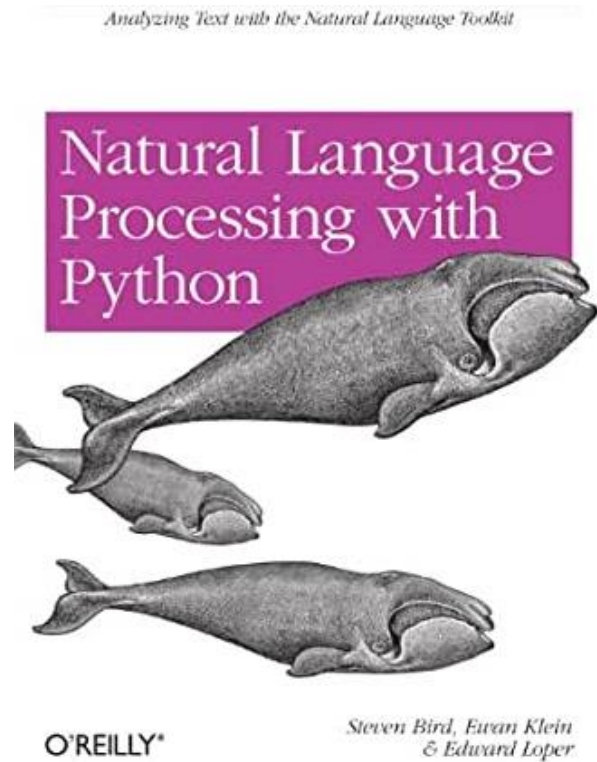
Intenta identificar y extraer sentimientos dentro de un texto determinado en blogs, reseñas, redes sociales, foros, noticias, etc. También se pueden analizar las emociones y personalidad.

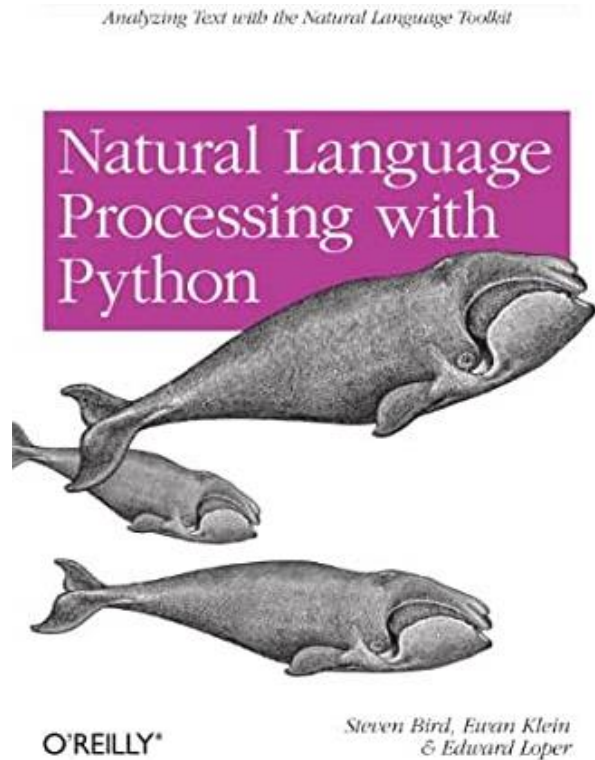
Detección de
Tópicos

Modelos para organizar, comprender y resumir (extraer palabras relevantes) grandes cantidades de textos.

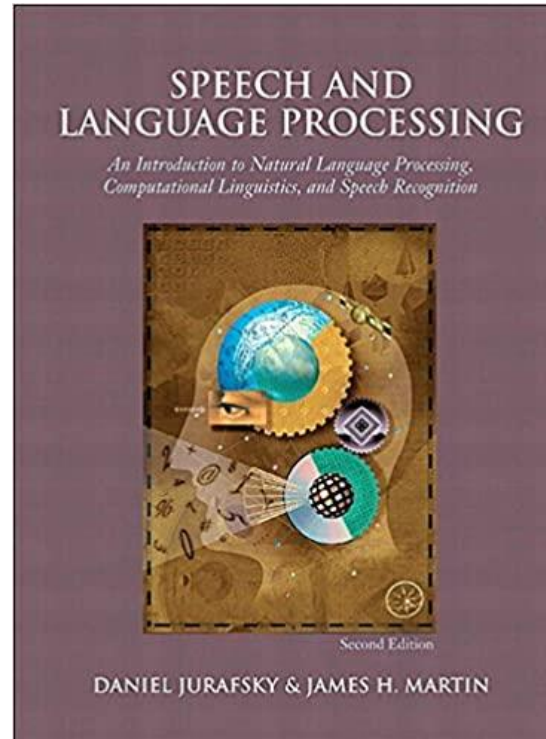


Lecturas sugeridas

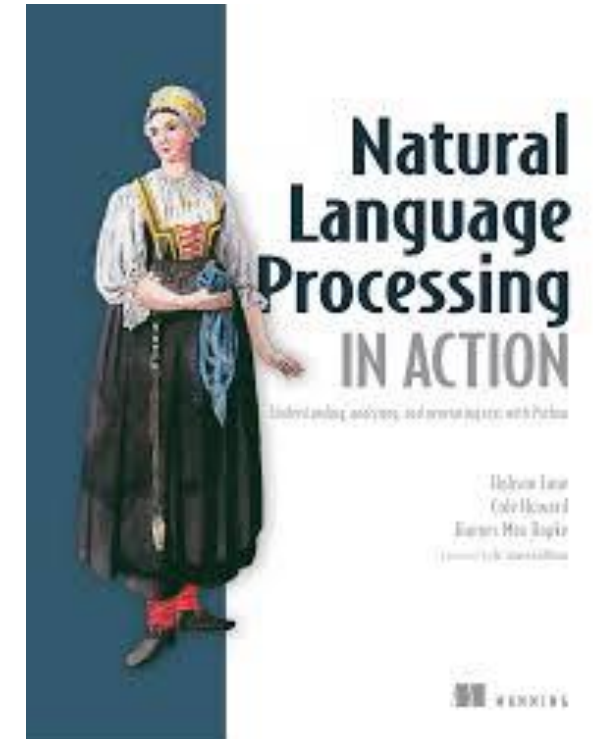




Capítulo 7. "Extracting Information from Text"
Capítulo 10. "Analyzing the Meaning of Sentences"



Capítulo 18. "Information Extraction"
Capítulo 19. "Word Senses and WordNet"
Capítulo 20. "Semantic Role Labeling"
Capítulo 21. "Lexicons for Sentiment, Affect, and Connotation"



Capítulo 4. "Finding meaning in word counts (semantic analysis)"



diplomatura universitaria en
inteligencia artificial



FACULTAD DE CIENCIAS
EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES

Procesamiento de Lenguaje Natural

Análisis del significado