

Procesamiento de Lenguaje Natural

Tópicos Avanzados en Analítica
Maestría en Analítica para la Inteligencia de Negocios

Sergio Alberto Mora Pardo - H2 2023

Procesamiento de Lenguaje Natural (NLP)

Clase 1 - Introducción y aplicaciones

Introducción:

1. ¿Qué es procesamiento de lenguaje Natural?
2. ¿Por qué el entendimiento del lenguaje natural es una tarea compleja?

Tareas del procesamiento de lenguaje natural.

Evolución: de reglas a aprendizaje de maquina.

Pipeline de NLP:

- Procesamiento
- Representación
- Modelamiento

Procesamiento de Texto: limpieza, tokenización y normalización.

Eliminación de stopwords y técnicas de reducción de palabras.

Introducción

**“A language is not just words. It’s a culture, a tradition,
a unification of a community,
a whole history that creates what a community is.
It’s all embodied in a language.”**

Noam Chomsky

“The art of creating machines that perform functions that require intelligence when performed by people”

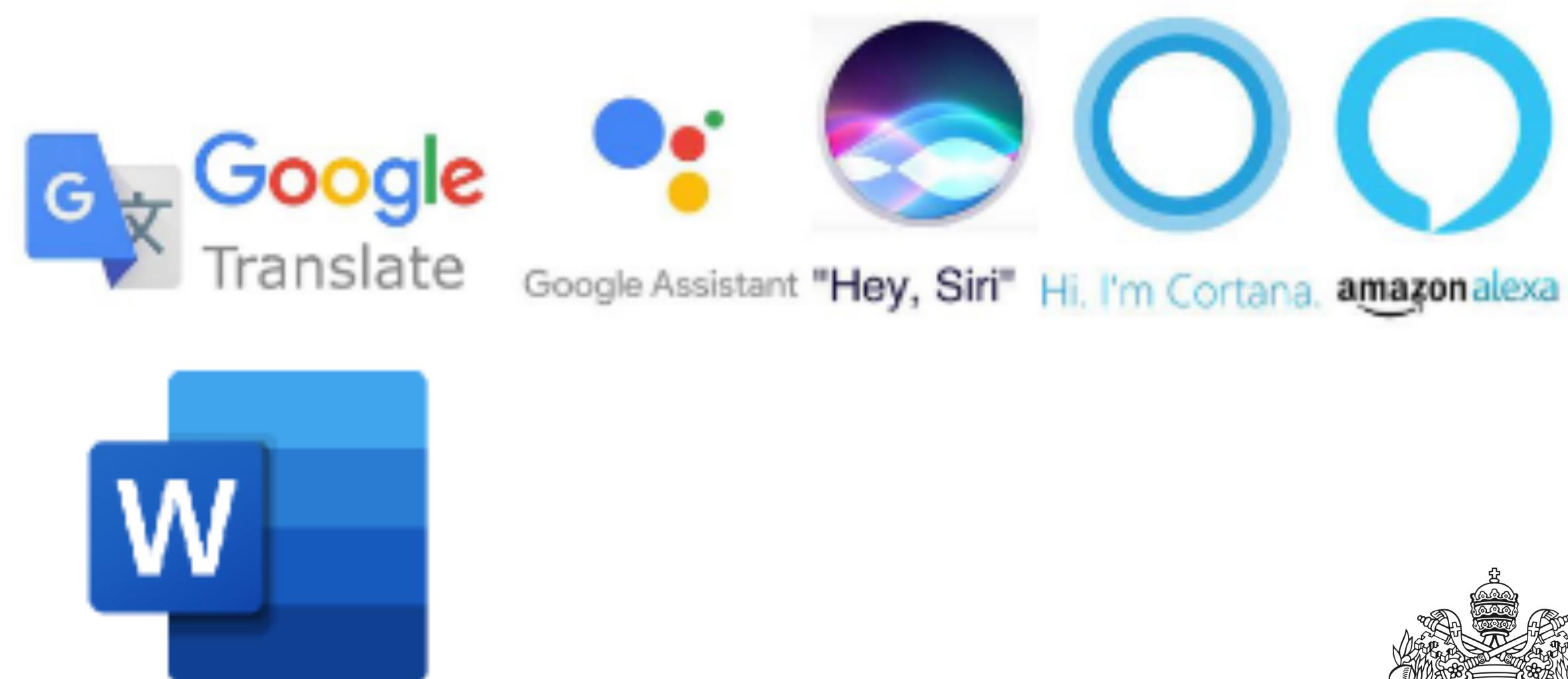
Raymond Kurzweil - 1990

Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural es el área que busca dotar a las computadoras con la capacidad de comprender el lenguaje natural del ser humano (escrito y hablado)



By:Seobility





2011 - IBM Watson vence a Ken Jennings y Brad Rutter en el show de TV Jeopardy.

Fuente: IBM research

<https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/>



Buscadores

Recuperación de información de corpus gigantescos...

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > news feature > article

NEWS FEATURE | 03 March 2021

Robo-writers: the rise and risks of language-generating AI

A remarkable AI can write like humans – but with no understanding of what it's saying.

Fuente: <https://www.nature.com/articles/d41586-021-00530-0>

Generación de texto usando modelos de lenguaje...

¿Del conjunto de palabras de mi vocabulario cuál es la palabra más probable?

Buenos Aires es la capital de _____

Bogotá

Universidad

Argentina

Generación de texto usando modelos de lenguaje...

¿Del conjunto de palabras de mi vocabulario cuál es la palabra más probable?

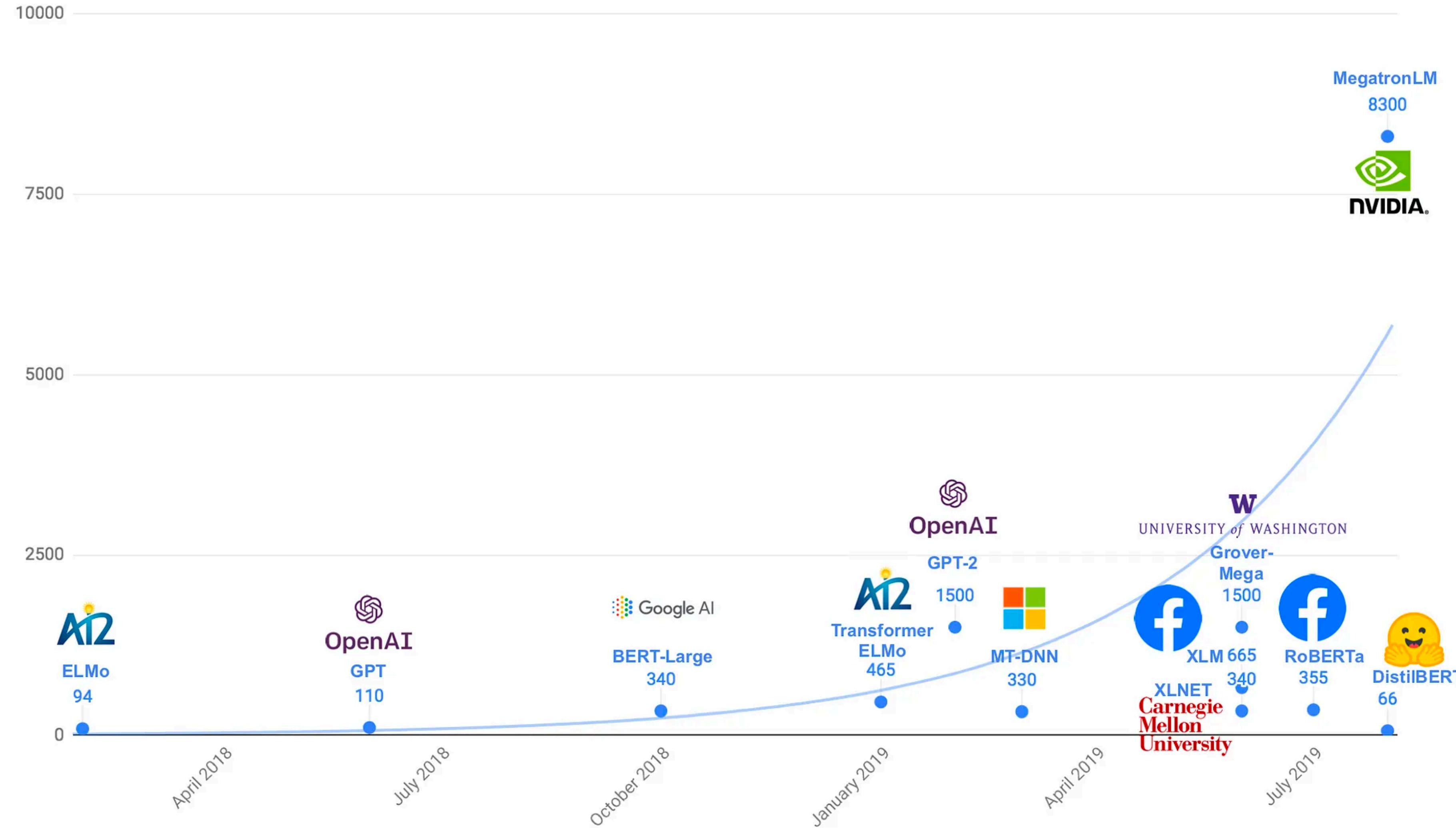


Chat GPT

Chat Generative Pre-Trained Transformer

Generación de texto usando modelos de lenguaje...

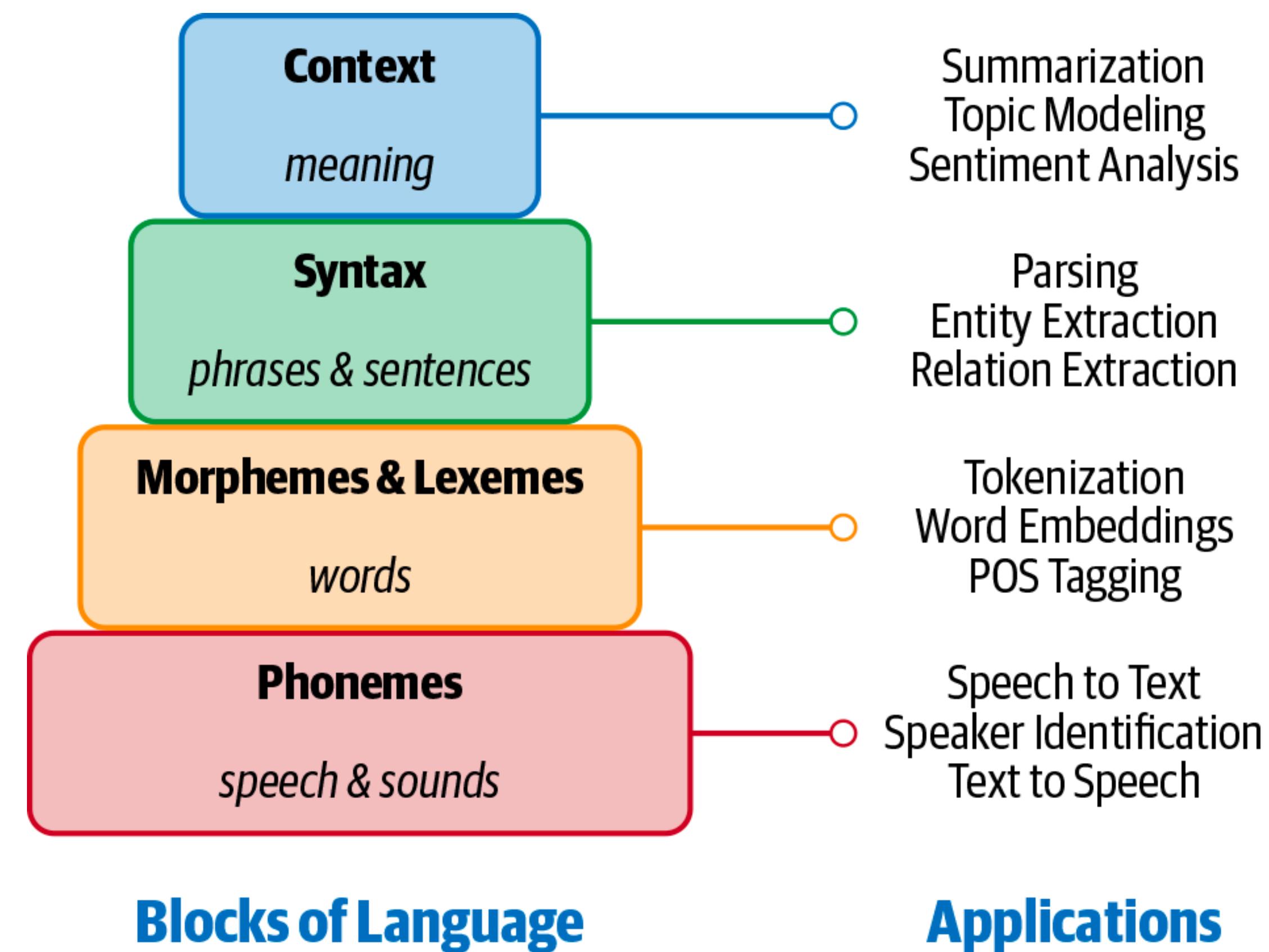
Generación de texto usando modelos de lenguaje...



Here's are some of the latest large models and their **size in millions of parameters**.

Linguistics

Fonemas, morfemas y lexemas, sintaxis y contexto.



Some concepts of linguistics

Sonido del lenguaje (fonemas)

Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/ ⁴ - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /th/ - then	10. /oo/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/ ³ - vision	12. /ow/ - cow	

Some concepts of linguistics

Lexemas y morfemas

unbreakable
un + break + able

tumbling
tumble + ing

cats
cat + s

unreliability
un + rely + able + ity

Lexema

CULP-

culpa
culpar
culpable

Morfema

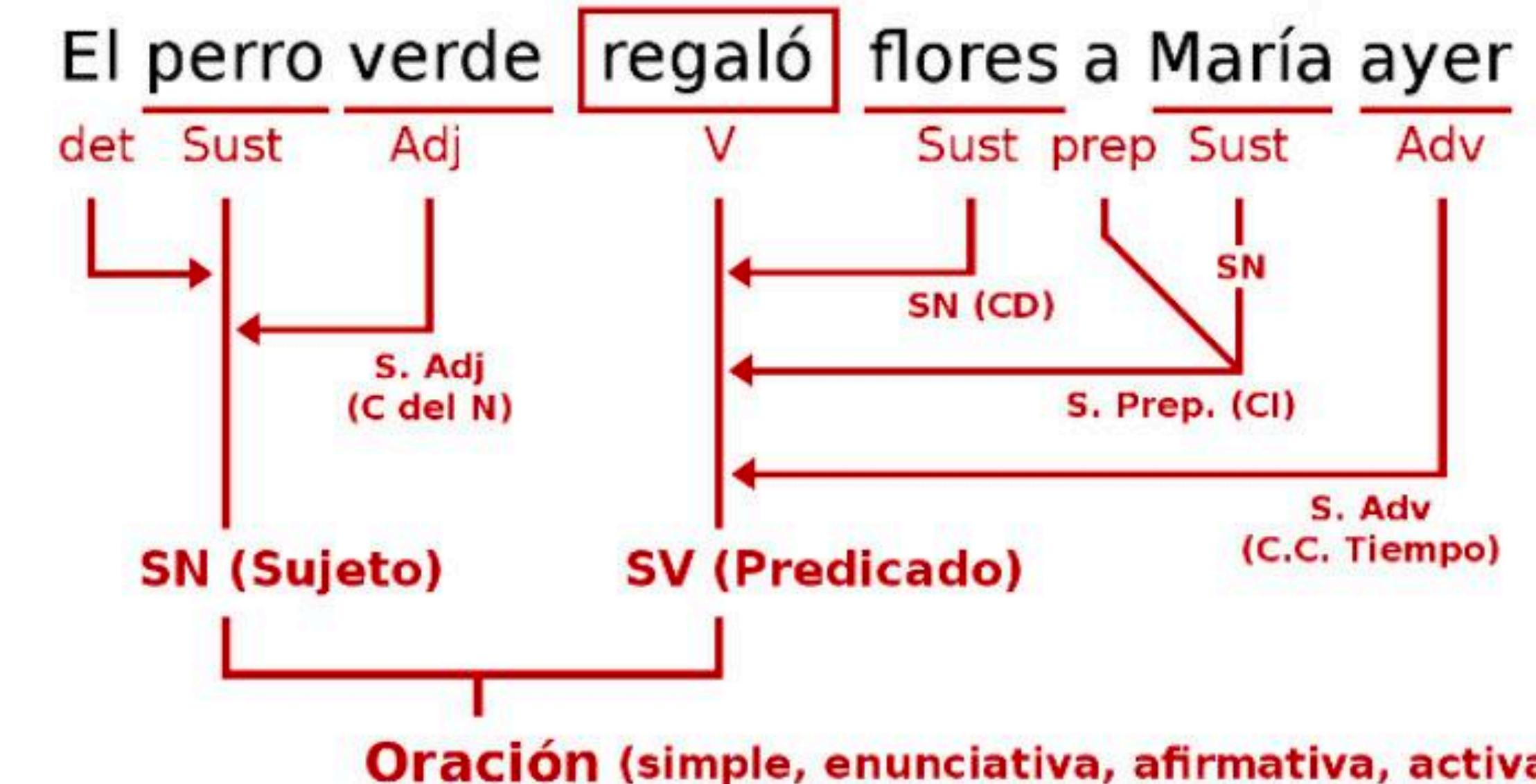
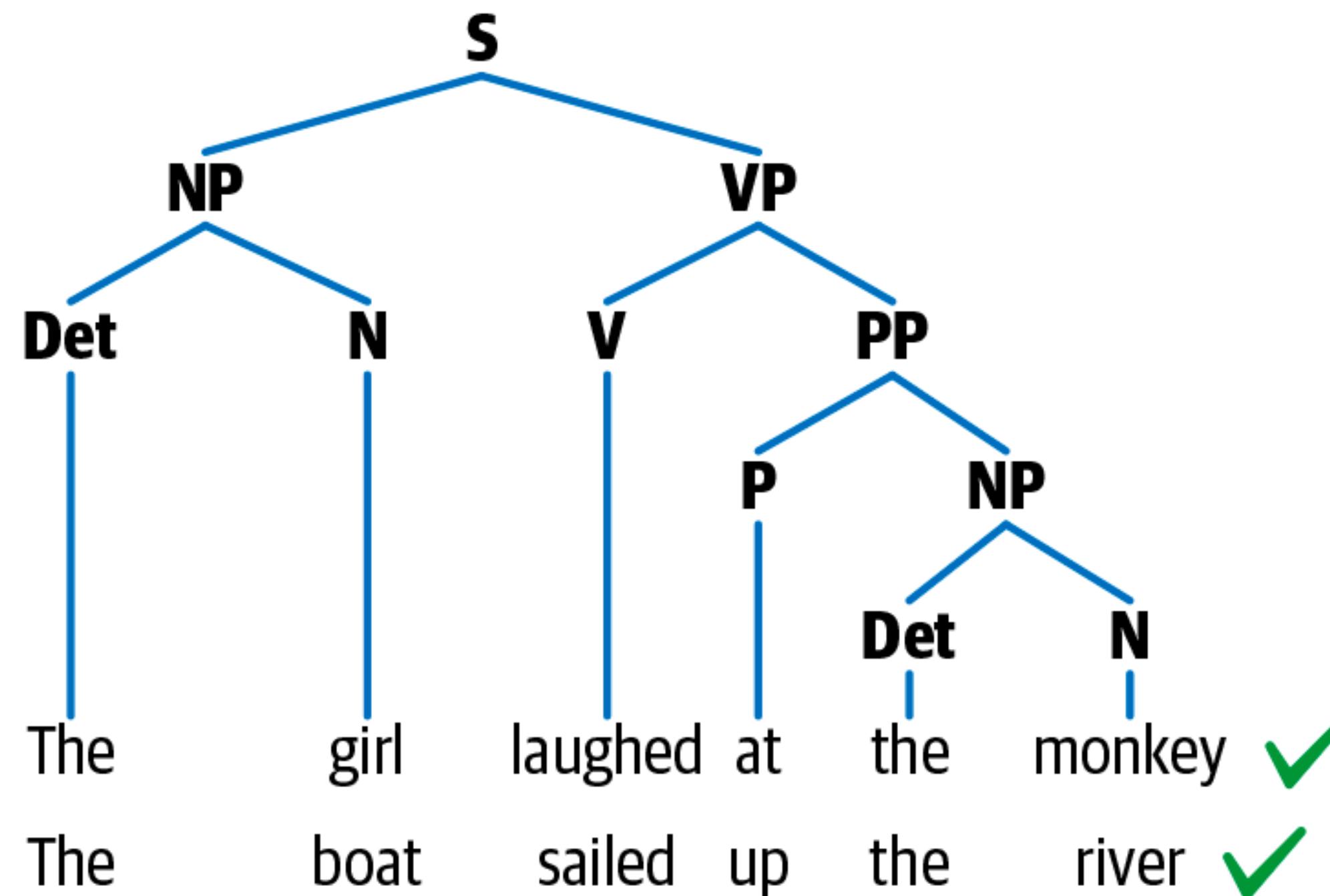
Inexplicablemente

-in
-mente

El lexema es la raíz de una palabra que lleva su significado principal, mientras que el morfema es la unidad más pequeña de significado en una palabra y puede ser un lexema o un afijo que modifica ese significado.

Some concepts of linguistics

Sintaxis



Some concepts of linguistics

Contexto

Semántica

Significado de las palabras y oraciones sin contexto externo

Significado directo

Pragmática

Agrega el conocimiento del mundo y contexto externo

Significado隐式 (Significado implícito)

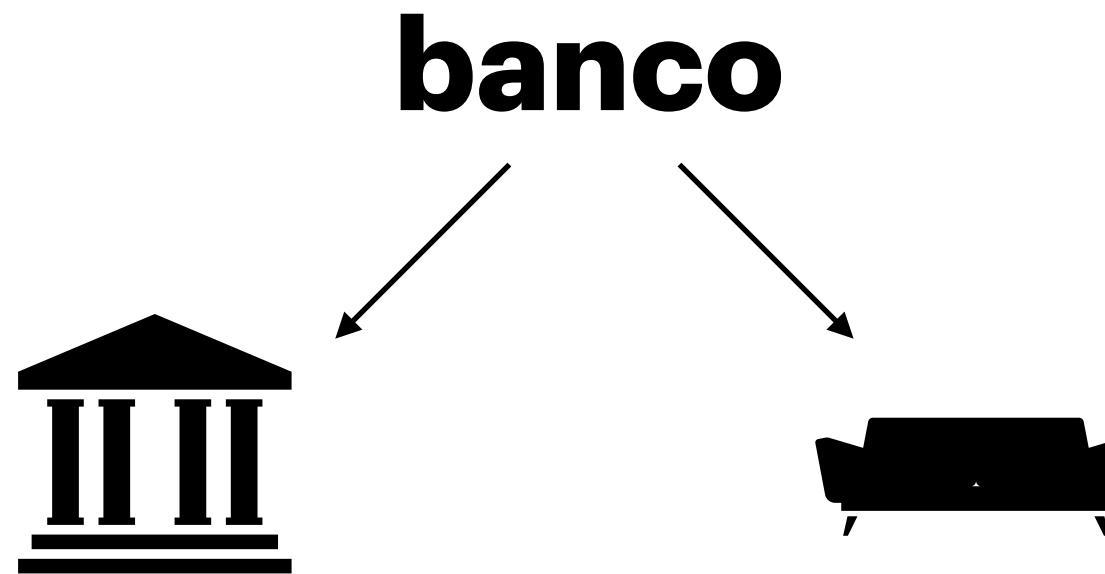
Ejemplos en NLP: detección del sarcasmo, Resumen, Modelación de temas

¿Porque el entendimiento del lenguaje es una tarea compleja?

Ambigüedad

Léxico-semántica (Polisemia)

La cita es en el **banco** donde les tomaron la fotografía



Sintáctica (Anfibología)

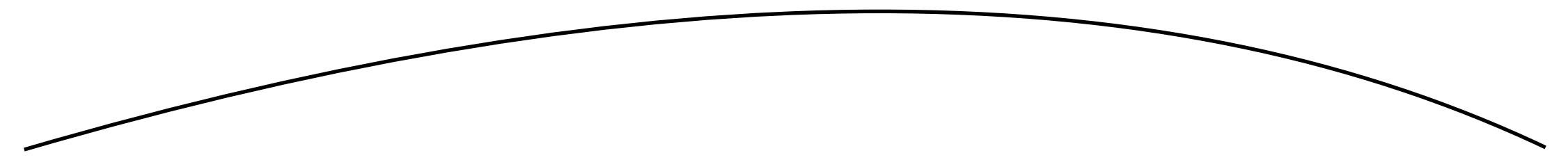
Vimos a Pedro tomando clase.
Sig. 1. Cuando tomábamos la clase, vimos a Pedro.
Sig. 2. Vimos a Pedro, que estaba tomando la clase.

Fonética

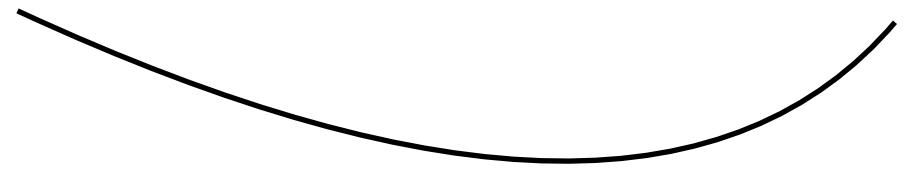
¿**Me diste** la ventana?
Sig. 1. Mediste- de medir
Sig. 2. Me diste- de dar

¿Porque el entendimiento del lenguaje es una tarea compleja?

Correferencia



La lámpara no cabe por la puerta porque esta es muy angosta.

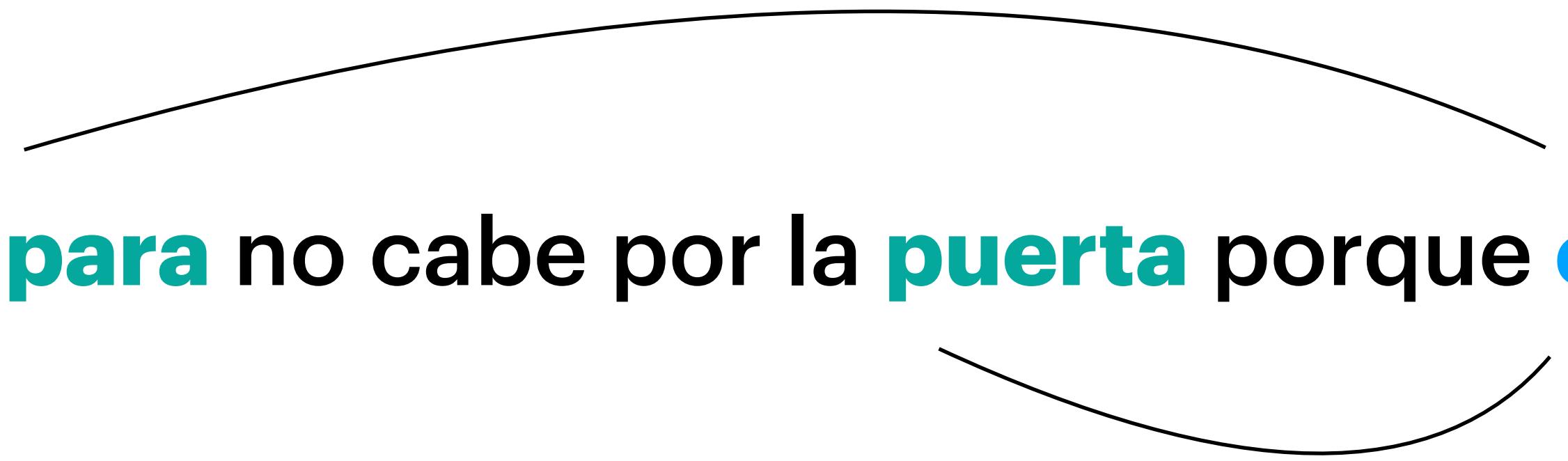


¿A que hace referencia?

- A. La lámpara
- B. La puerta

¿Porque el entendimiento del lenguaje es una tarea compleja?

Correferencia



La **lámpara** no cabe por la **puerta** porque **esta** es muy *alta*.

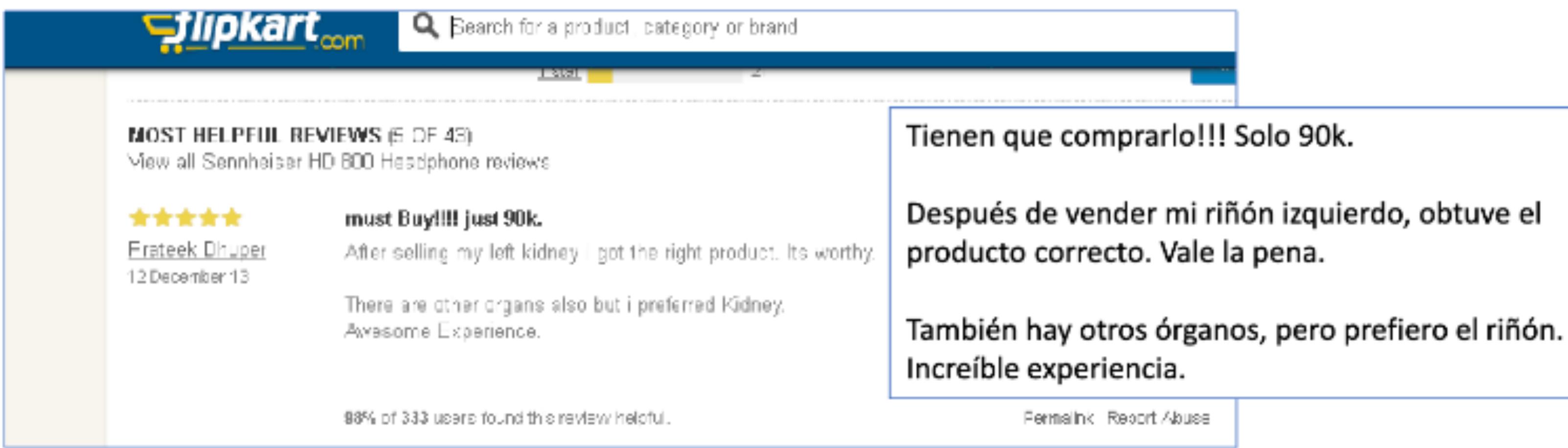
¿A que hace referencia?

- A. La lámpara
- B. La puerta

¿Porque el entendimiento del lenguaje es una tarea compleja?

Sarcasmo/Ironía

- No estarás muy cansado, ¿no?
- Me gusta su perfume, ¿cuánto tiempo ha pasado sumergido en él?



The screenshot shows a product review page from flipkart.com. At the top, there's a search bar and a navigation menu. Below it, a section titled "MOST HELPFUL REVIEWS (5 OF 43)" shows a review by "Frateek Dhuper" dated "12 December 13". The review has a 5-star rating and the text: "must Buy!!! just 90k. After selling my left kidney... got the right product. Its worthy. There are other organs also but i preferred Kidney. Awesome Experience." A note below says "88% of 333 users found this review helpful." To the right, a box highlights a portion of the review: "Tienen que comprarlo!!! Solo 90k. Después de vender mi riñón izquierdo, obtuve el producto correcto. Vale la pena." Below this, another box contains the translation: "También hay otros órganos, pero prefiero el riñón. Increíble experiencia."

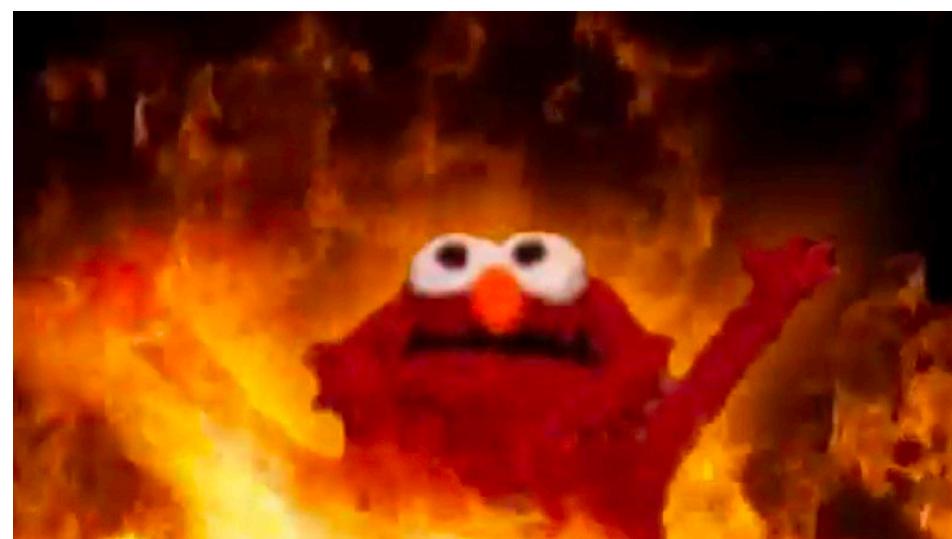
Fuente: <https://flipkart.com/>

¿Porque el entendimiento del lenguaje es una tarea compleja?

Otros

Lenguaje no estándar (redes sociales)

RT @Pere4L: Guy PLZ d/l the letter
I've written 2 Jeff Gaspin, he's THE man
who can giv us #Heroes S5 <http://tinyurl.com/y9pcaj7> #Heroes100 :)



Modismos

Chanda, seba, guayabo,
camello

Neologismos

Textear, Hipertexto, clickear,
escanear

Nombre de entidades

Musica ligera fue grabado en...
Luisito comunica no ha dado declaraciones...

Problemas abordados

Clasificación de texto

Ej. 1: Spam/No spam

La reunión quedó programa para el ... (**NO SPAM**)

Gana dinero fácil, invierte en P1pOpl3 ... (**SPAM**)

Ej. 2: Análisis de polaridad

Mejor PC gamer de la historia (+)

La pantalla se rompió a la semana, no la
compren. (-)

Problemas abordados

Identificación de partes de la oración (POS)

El niño juega con la pelota.

SUSTANTIVO
VERBO
ARTICULO
PREPOSICIÓN

El niño juega con la pelota.

**Comprensión sintáctica
(Parsing)**

Corrección de texto

Problemas abordados

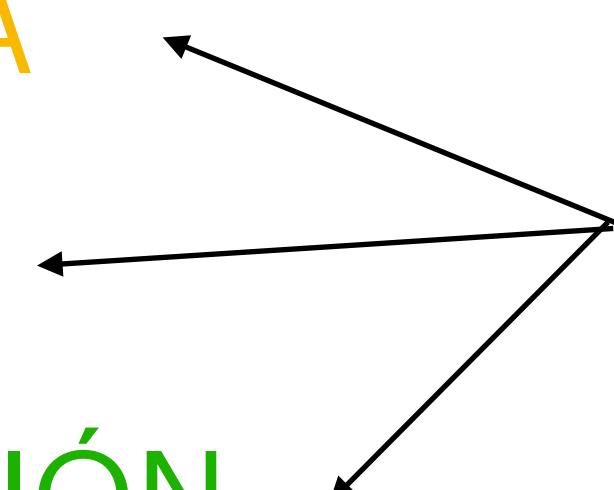
Identificación de entidades nombradas (NER)

Luis Fernando Múnera es el rector de la universidad Javeriana en Bogotá

NER:

clasificar en el texto entidades, personas, organizaciones, lugares, y expresiones de texto.

PERSONA
LUGAR
ORGANIZACIÓN



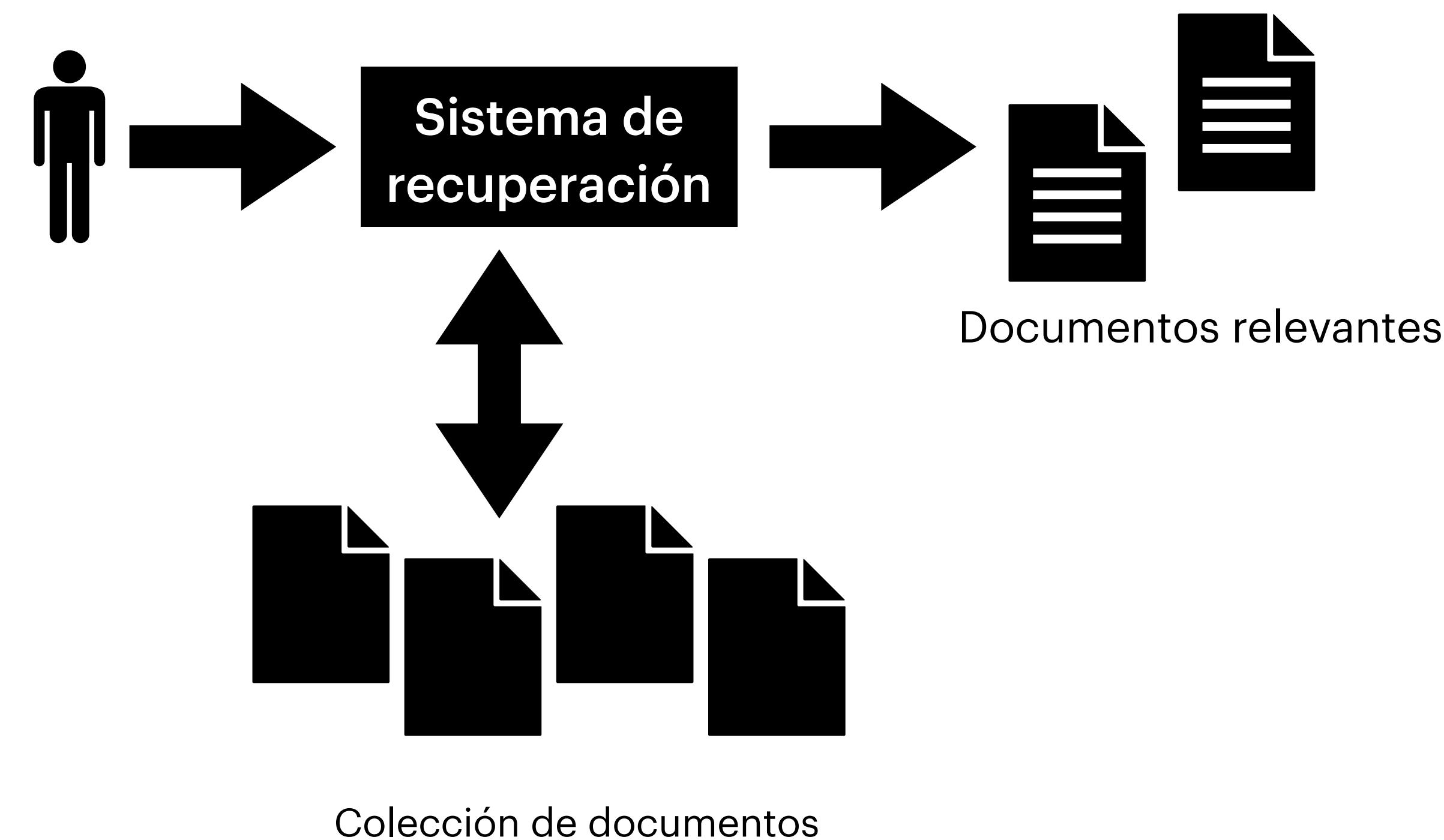
Luis Fernando Múnera es el rector de la Universidad Javeriana en Bogotá

Problemas abordados

Recuperación de Información

Sistema de recuperación

Acceder y recuperar la información textual (documentos) más apropiada aen base a una consulta (query).



Problemas abordados

Correferencia

Def.

Relación entre dos o más elementos lingüísticos cuyo referente es común.

Ej.

*El profesor le dice al estudiante que **él** debe renunciar a sus malos hábitos de estudio. A pesar de esto sus malos hábitos **éste** ha logrado superar la moría de los exámenes.*

Comentario:

En el texto anterior, ¿las palabras **él** y **éste** hacen referencia al profesor o al estudiante?

En textos largos los pronombres sigue siendo un problema desafiante.

Problemas abordados

Desafíos

Preguntas a Respuestas.

Los sistemas de Q&A requieren un conocimiento “no genérico” que debe establecerse mediante análisis de intenciones o modelos de conocimiento.

Sistema de dialogo (chatbot)

Human: What is your favorite animal?

GPT-3: My favorite animal is a dog.

Human: Why?

GPT-3: Because dogs are loyal and friendly

Resúmenes Automáticos.

Sintetizar o reducir el texto preservando la idea o ideas más importantes.

NLP Tasks

Principales tareas que aparecen con frecuencia en proyectos de NLP

- Modelación de Lenguaje:
 - A. Esta es la tarea de predecir cuál será la siguiente palabra en una oración en función de la historia de las palabras anteriores.
 - B. El objetivo de esta tarea es aprender la probabilidad de que una secuencia de palabras aparezca en un idioma determinado.

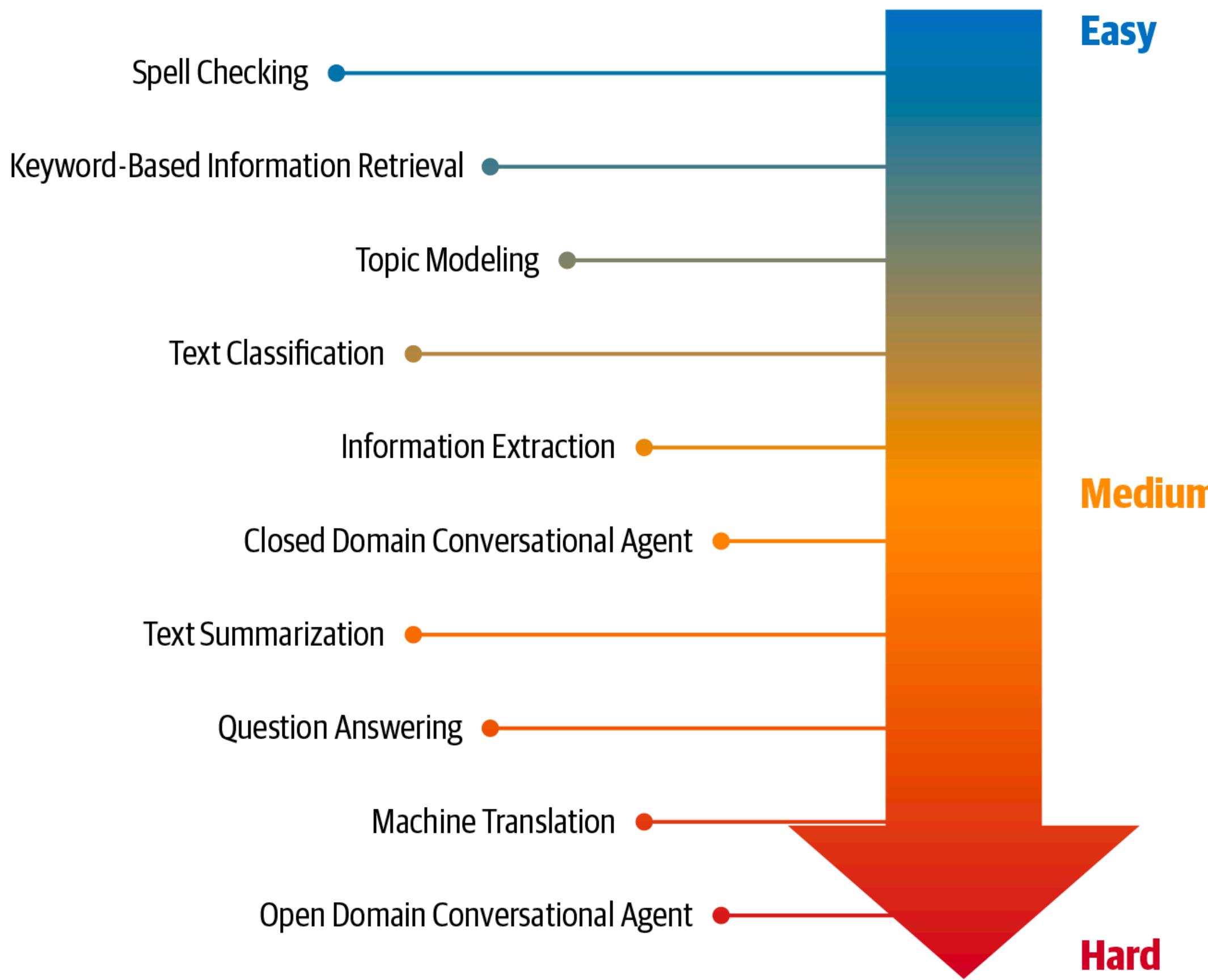
NLP Tasks

Principales tareas que aparecen con frecuencia en proyectos de NLP

- Modelación de Lenguaje
 - Resumen de Texto
- Clasificación de Texto
 - Question answering
- Extracción de Información
 - Machine translation
- Recuperación de Información
 - Topic modeling
- Agente Conversacional

NLP Tasks

Principales tareas que aparecen con frecuencia en proyectos de NLP



Pipeline NLP

NLP pipeline

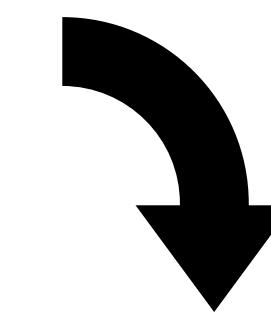
Componentes en la construcción de un modelo de NLP

Procesamiento
de texto.



Stopwords
Tokenización
Lemmatización

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy....



Representación

BOW

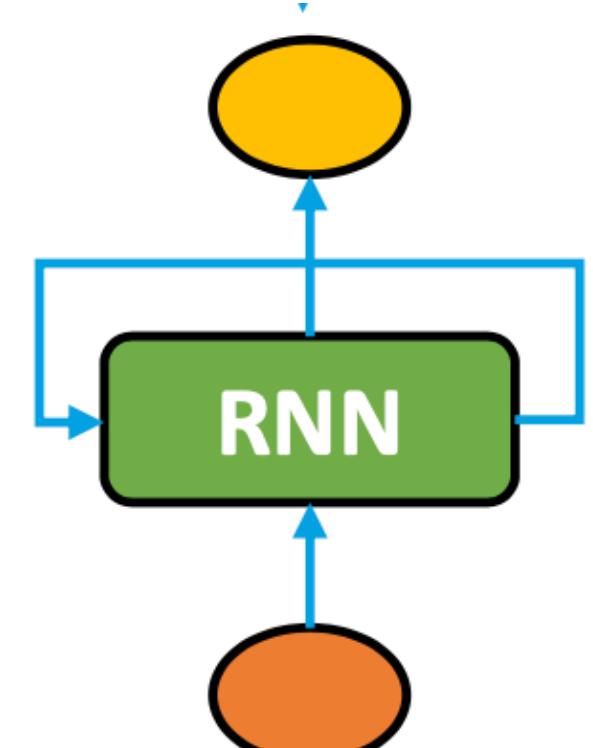


\vec{AB}

Word2Vec

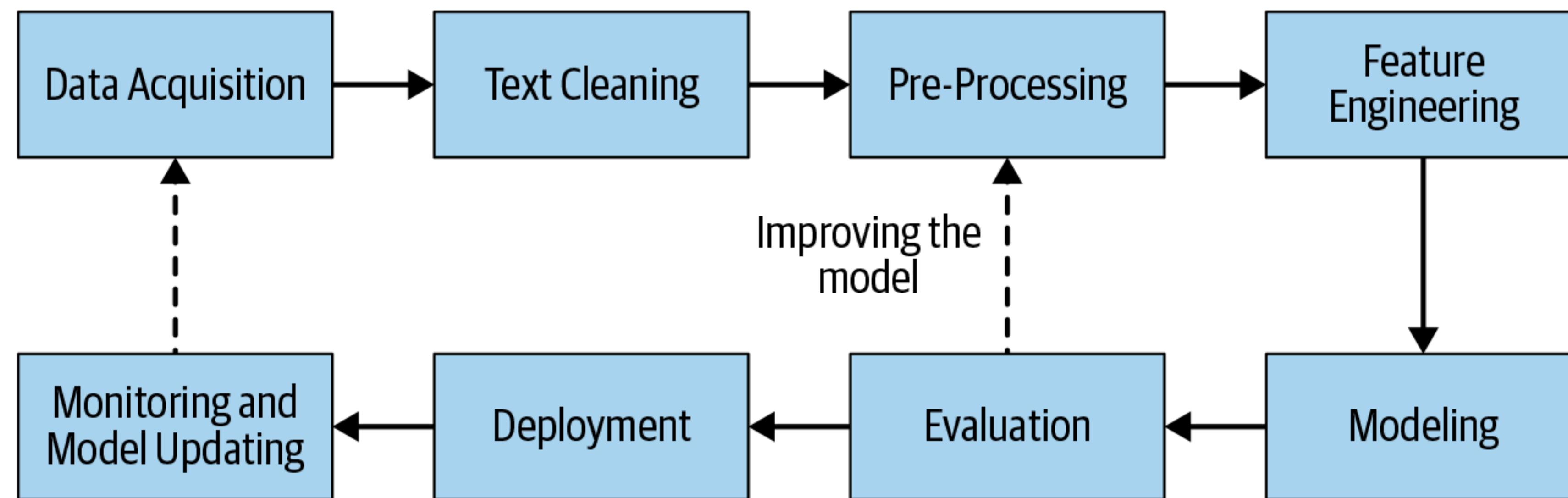
Modelamiento

- Traducción
- Generación de texto
- Clasificación de texto



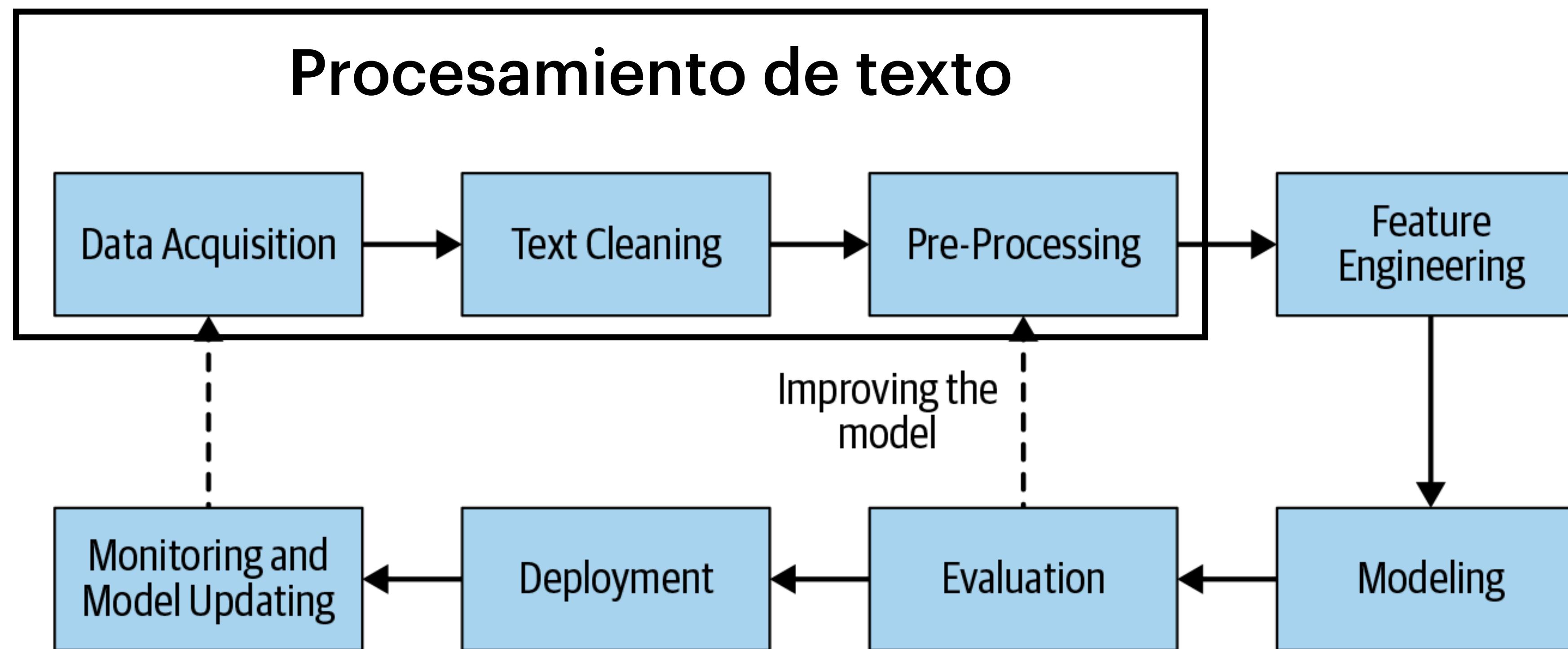
NLP pipeline

Componentes en la construcción de un modelo de NLP



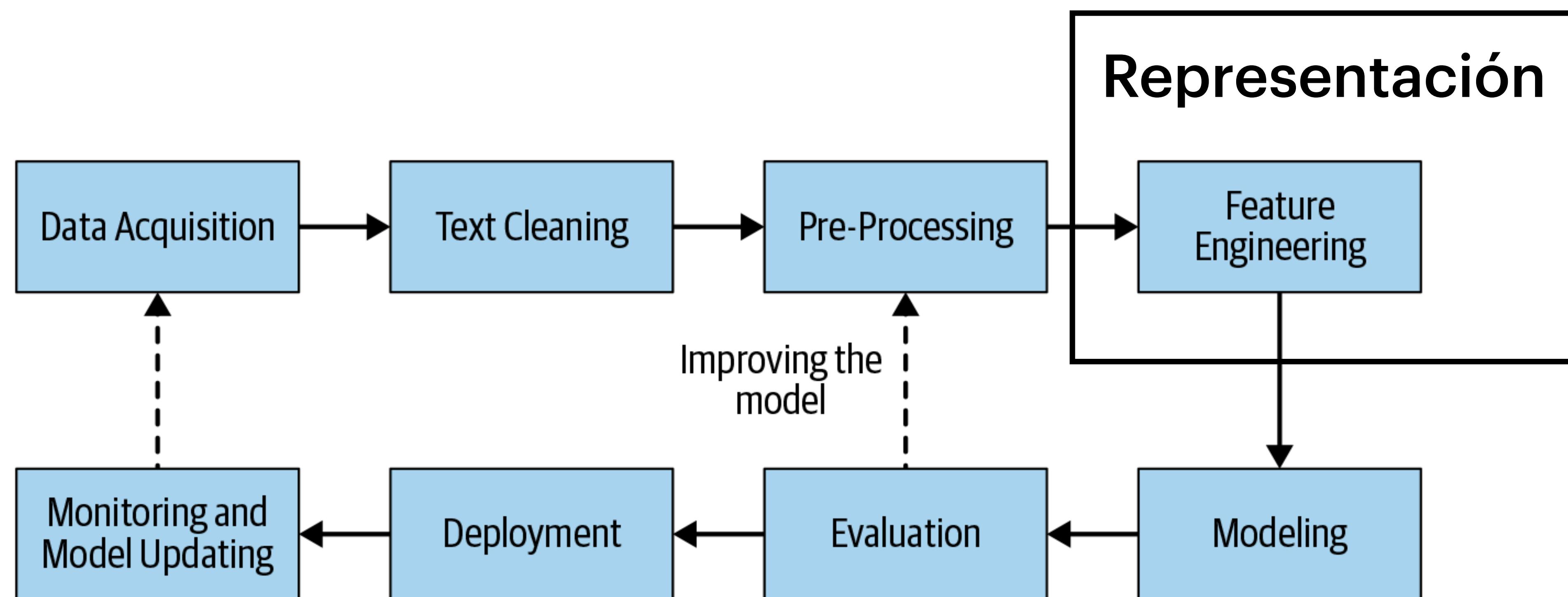
NLP pipeline

Componentes en la construcción de un modelo de NLP



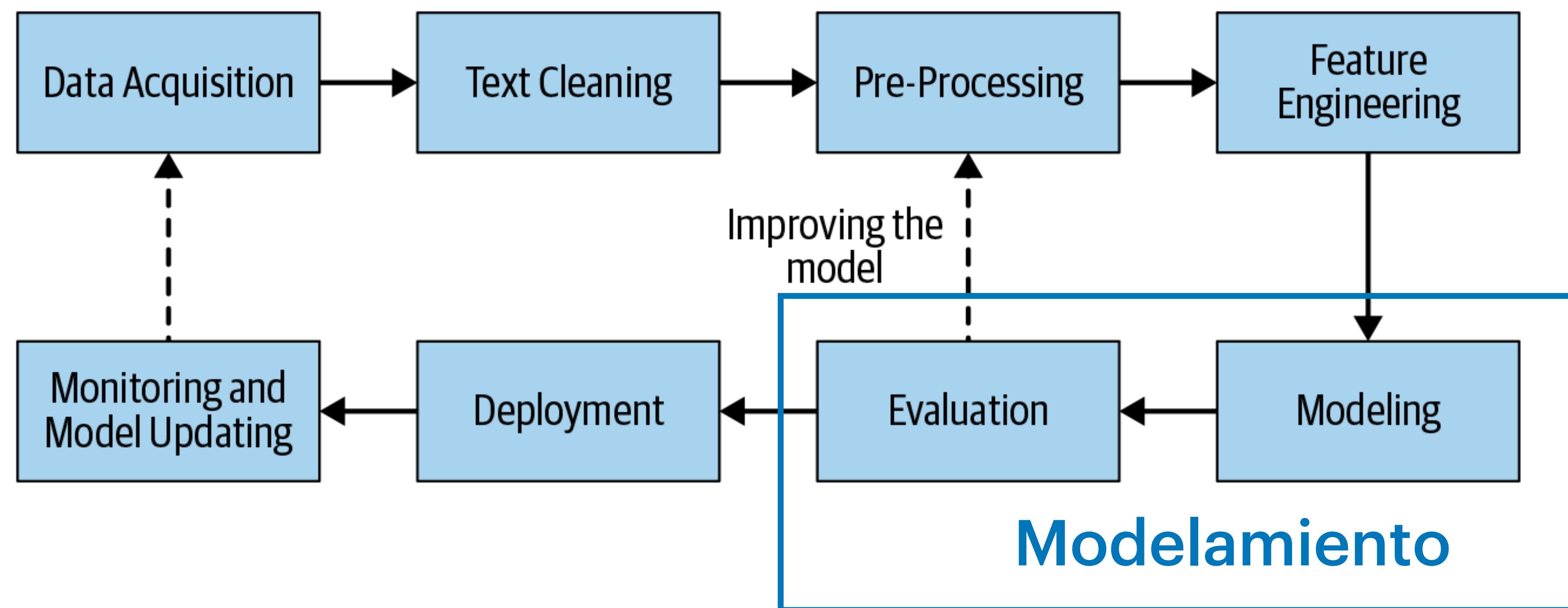
NLP pipeline

Componentes en la construcción de un modelo de NLP



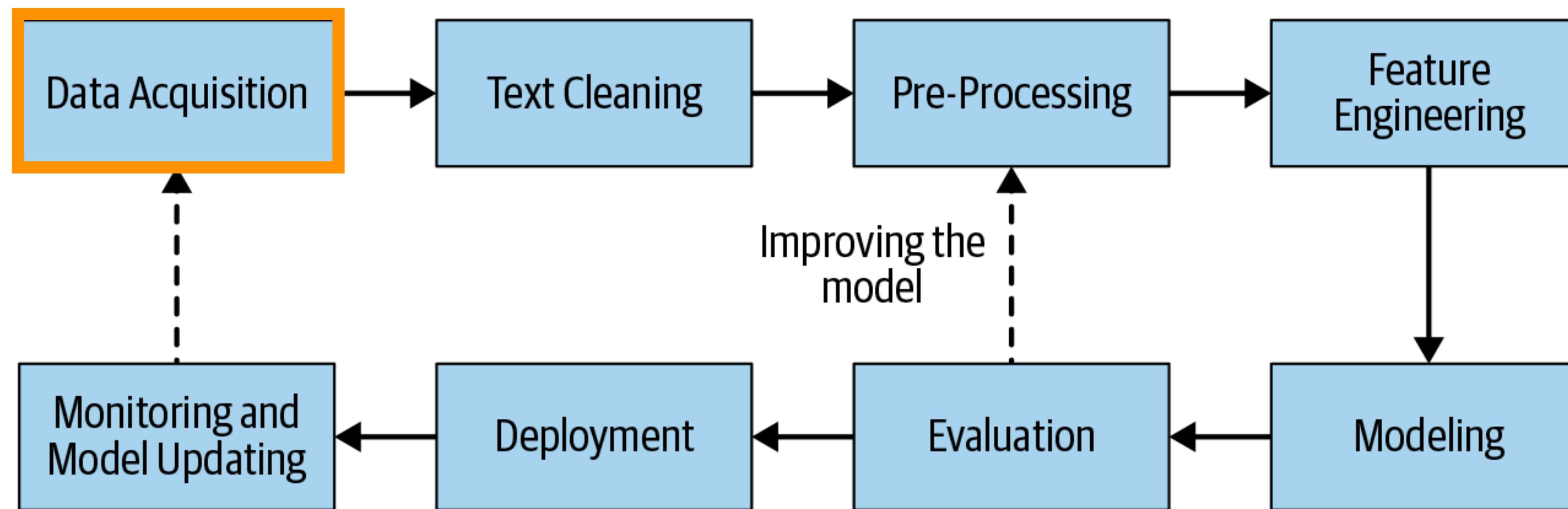
NLP pipeline

Componentes en la construcción de un modelo de NLP



NLP pipeline

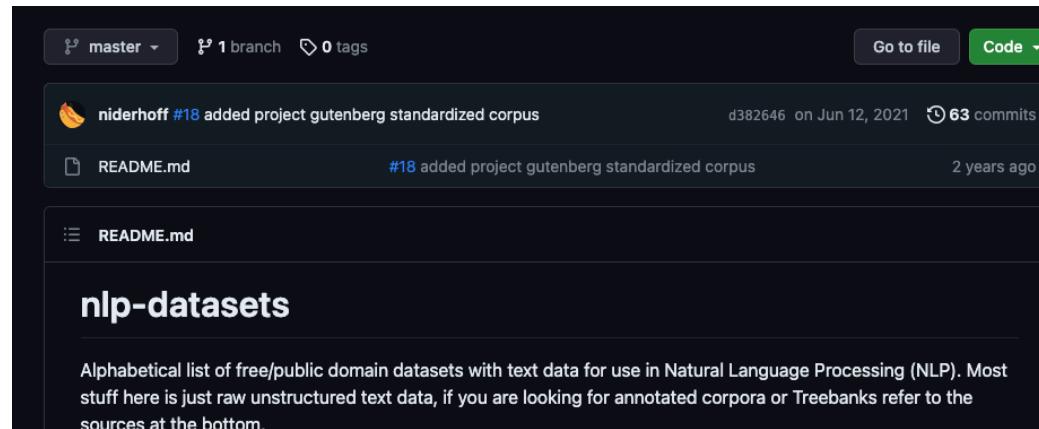
Componentes en la construcción de un modelo de NLP



Adquisición de Datos

Componentes en la construcción de un modelo de NLP

Conjunto de datos público



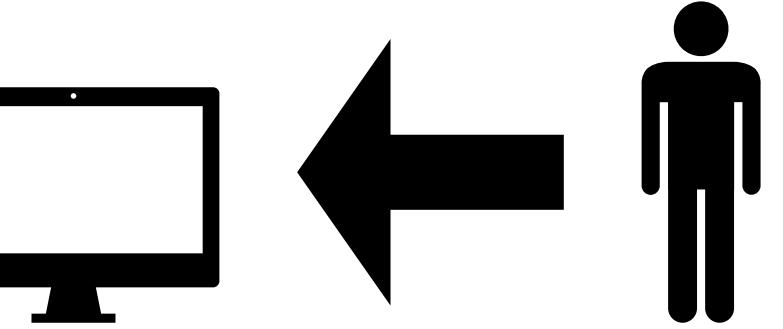
Dataset Search

Buscar conjuntos de datos



Prueba coronavirus covid-19 o water quality site:canada.ca.

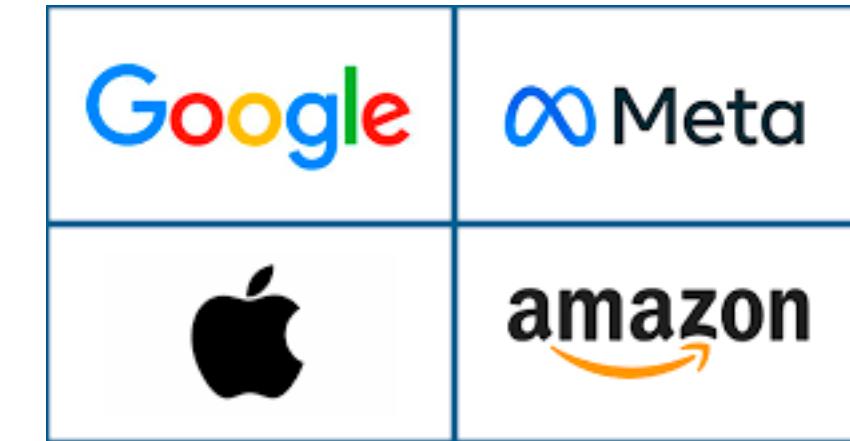
Más información sobre Búsqueda de Datasets



'Raspadatos'

Etiquetados de datos manual

Intervenir el Producto



Google

Meta



amazon

Recopilación de datos a base del producto.

Las primeras implementaciones son demoradas.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Aumentado de datos

El Procesamiento de Lenguaje Natural tiene varias técnicas para tomar un pequeño conjunto de datos y usar algunos trucos para crear más datos.

Reemplazo de sinónimos

Elegir palabras aleatoriamente para reemplazar por su sinónimo.

Traducción inversa

Traducir S1 a un segundo lenguaje, dejando S2. Luego, traducir S2 a lenguaje original.

Sustitución de entidades

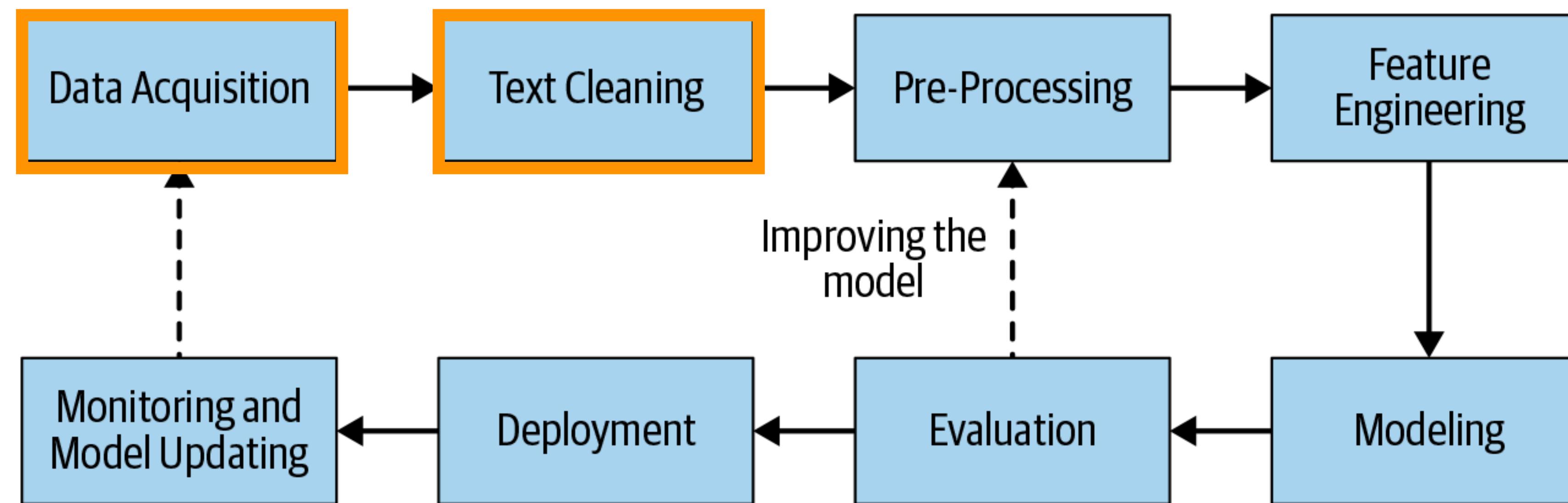
Sustituir ‘California’ con ‘Londres’. Esto depende siempre del contexto.

Agregar ruido a los datos

Fat-finger error

NLP pipeline

Componentes en la construcción de un modelo de NLP



Limpieza de Texto

Componentes en la construcción de un modelo de NLP

Análisis y limpieza de HTML



An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

Maintained by [Zyte](#) (formerly Scrapinghub) and [many other contributors](#)

pypi v2.10.0 wheel yes coverage 89% Anaconda.org 2.10.0

beautifulsoup4 4.12.2

pip install beautifulsoup4

Normalización Unicode

↑	𠂇	-	,	Θ	.GREEN HEART	☺	δ	ڦ	ڻ
γ	ϙ	♥	μ	🚀	♪	՝	.	ڙ	“
☀		○	჏	且	Ѣ	ԡ	%o	ﾃ	⟳
‘	”	I	ঠ	ঢ	ঢ	ঢ	ঢ	ঢ	ঢ

```
texto = '¡Me encanta 🍕. ¿ Reservamos un 🚗 gizza?  
Texto = texto.encode("utf-8")  
imprimir (texto)
```

Corrección ortográfica

Quickstart: Check spelling with the Bing Spell Check REST API and Python

Article • 02/01/2022 • 14 contributors

In this article

- Prerequisites
- Create an Azure resource
- Initialize the application
- Create the parameters for the request

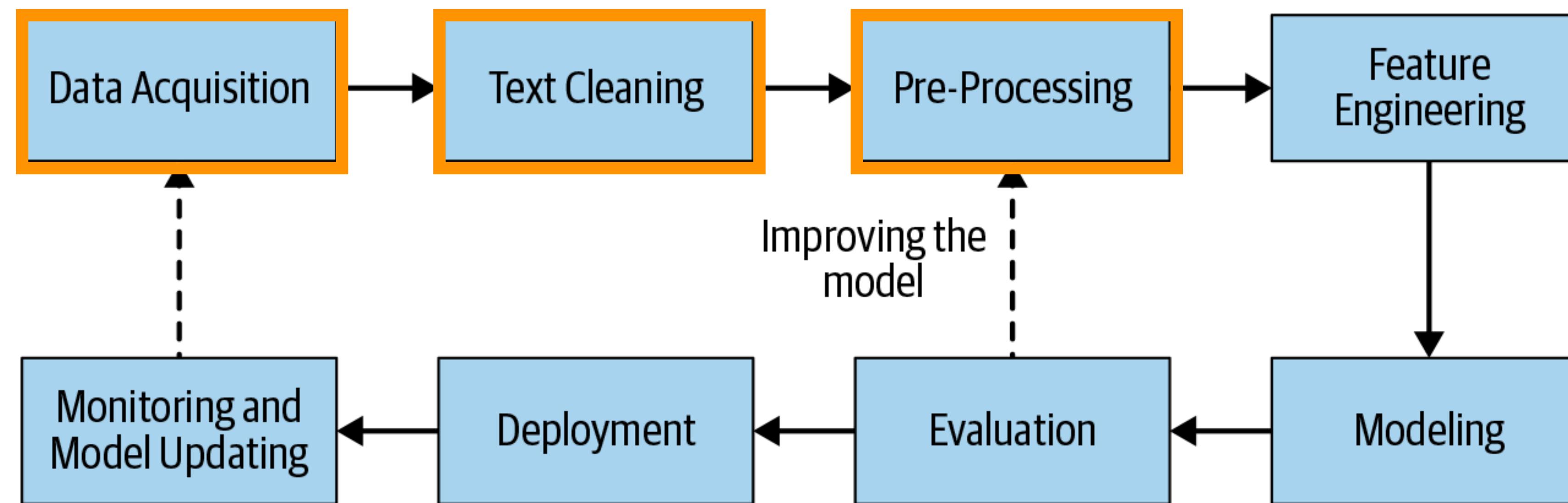
Show 4 more

⚠ Warning

On October 30, 2020, the Bing Search APIs moved from Azure AI services to Bing Search Services. This documentation is provided for reference only. For updated documentation, see the [Bing search API documentation](#). For instructions on creating new Azure resources for Bing search, see [Create a Bing Search resource through the Azure Marketplace](#).

NLP pipeline

Componentes en la construcción de un modelo de NLP

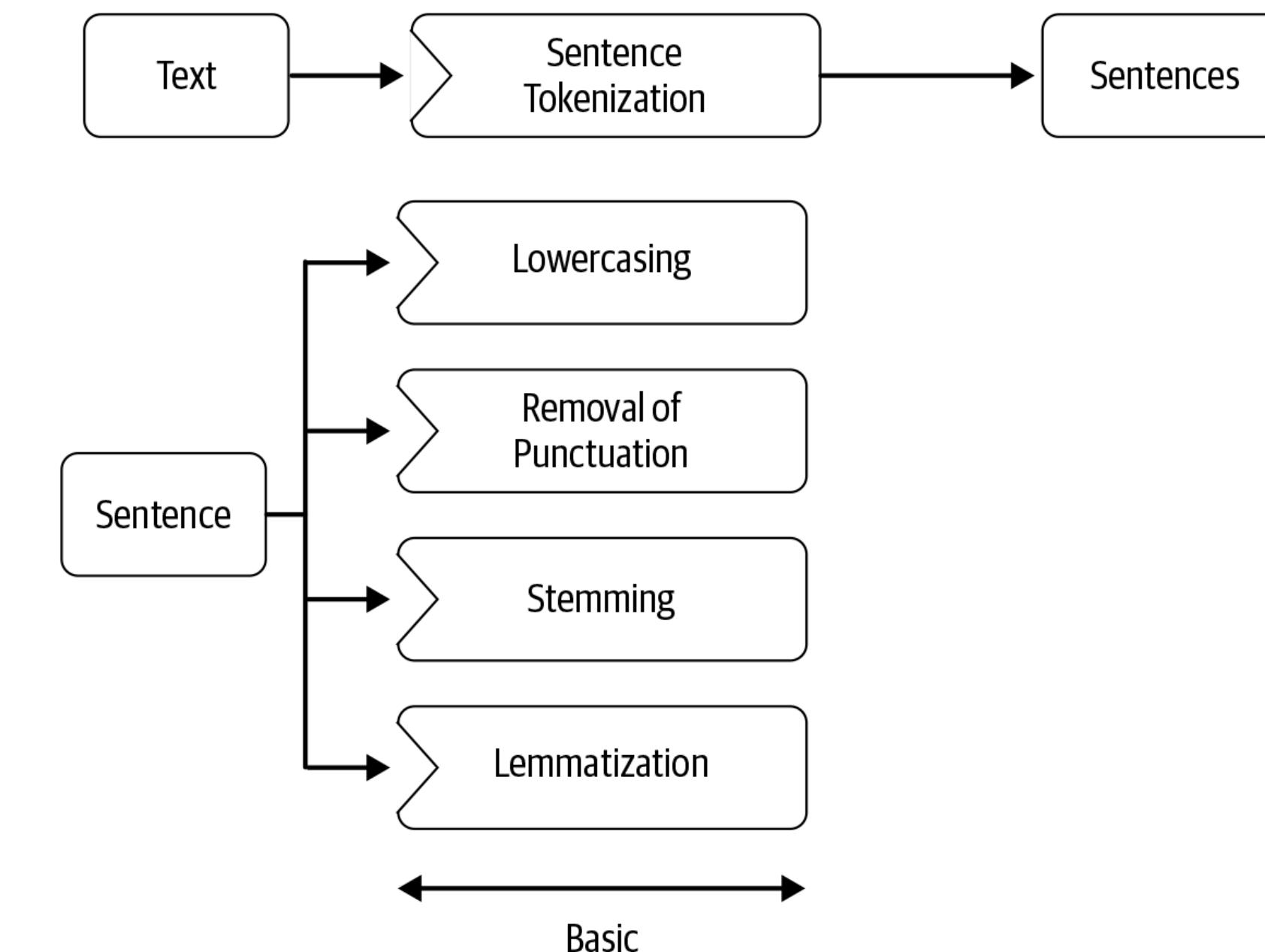


Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.

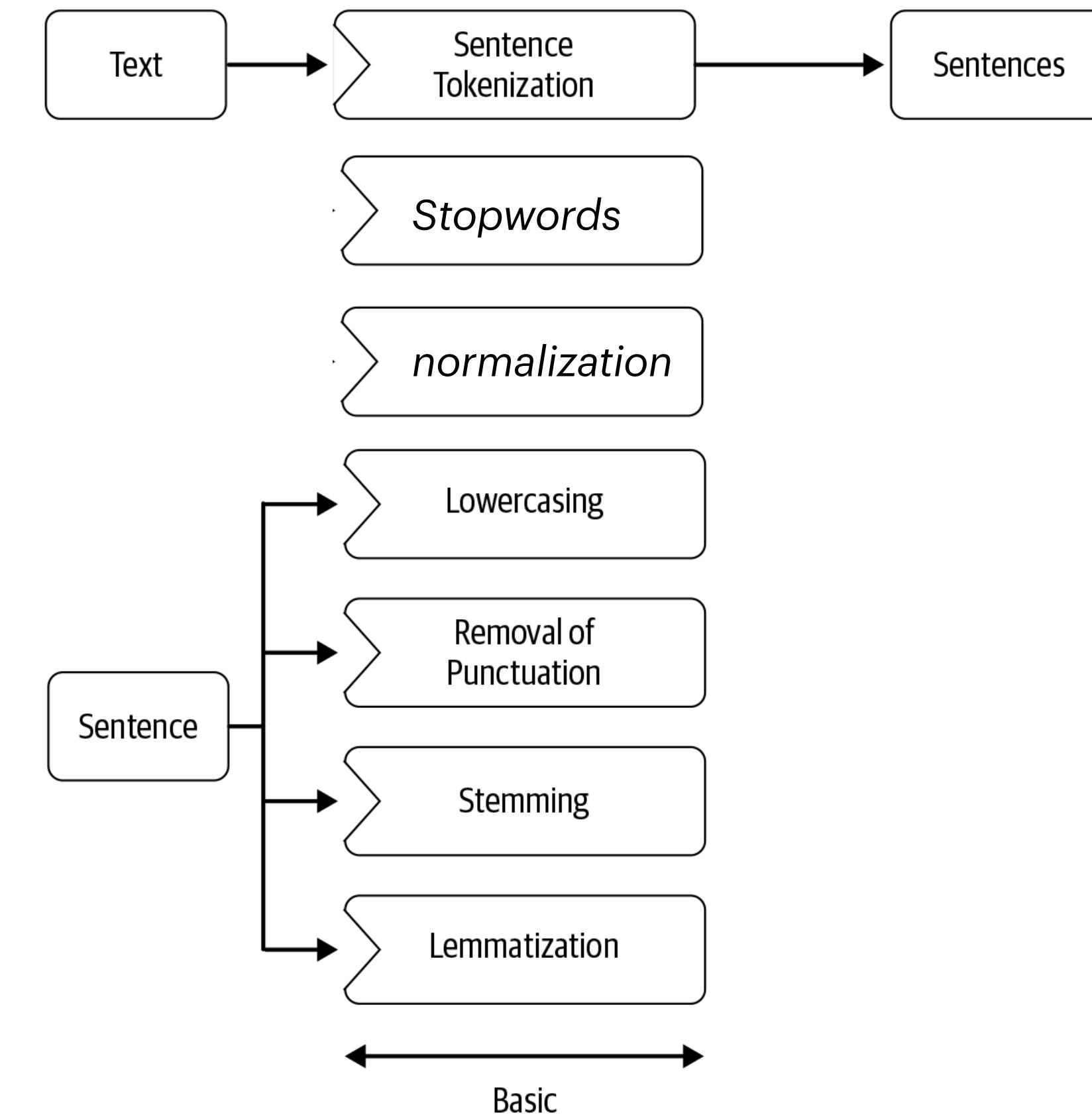


Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Tokenización

Unidad mínima para procesamiento.

Entrada:	<i>Los amigos de Diana.</i>
Salida:	Tokens [Los, amigos, de, diana]
Def. Token:	<i>Instancia de secuencias de caracteres.</i>
<i>Token es ahora un candidato para un índice... Pero ¿cuáles se consideran tokens válidos?</i>	

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Tokenización

Unidad mínima para procesamiento.

Retos:

Finland's capital

Finland AND s?

Finlands?

Finland's?

Hewlett-Packard

Hewlett y Packard?

¿Rompemos la secuencia con guiones?

Música Ligera

¿Un token o dos?

Lebensversicherungsgesellschaftsangestellter
'Life insurance company employee'

En alemán los sustantivos compuestos no se segmentan

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Tokenización

Unidad mínima para procesamiento.

Retos:

莎拉波娃现在居住在美国东南部的佛罗里达。

Chino no tiene espacio entre palabras
No siempre se garantiza una única tokenización



Japonés:
Multiples alfabetos entremezclados

(904) 265 4843

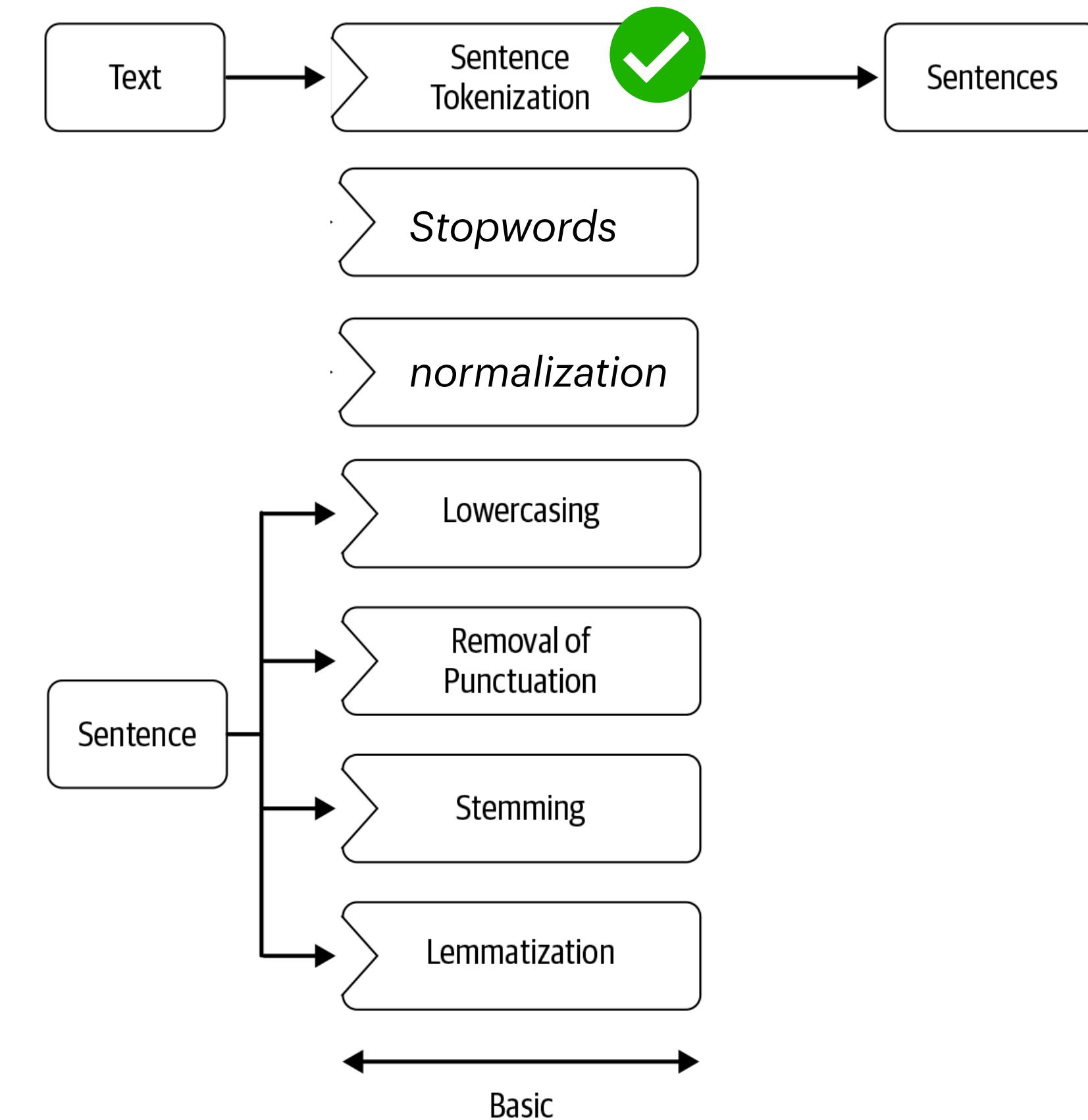
Números

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Stopwords

Palabras con poco contenido semántico.

Ej.:

[la, a, y, de, como]

*Lista de parada para excluirlas.
No sirve como criterio diferenciado de documentos.*

Tendencia:

- 1.** Hacen parte de la sintaxis de una oración correcta.
- 2.** Los embeddings contextuales requieren de esta información.

Ej.: Rey de Dinamarca
Vuelos a Bogotá

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

HARVARD & SN

Stopwords

Palabras con poco contenido semántico.

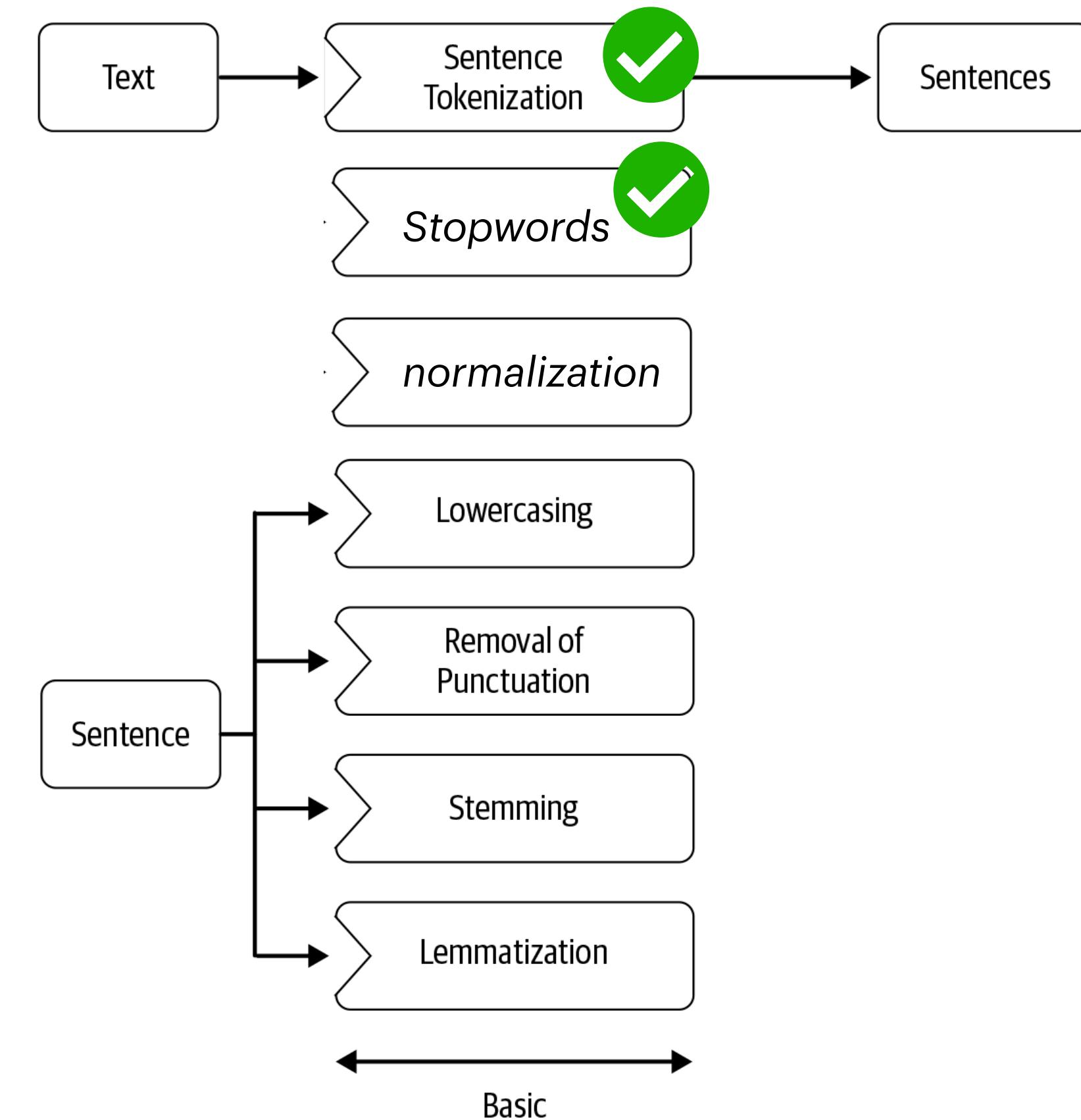


Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Normalización

En algunos casos se requiere 'normalizar' las palabras en el texto indexado y de la consulta.

Ej.:

U.S.A USA

Francés, español,
currículum / curriculum

Salida:

término

*Existen muchos de ellos en las colecciones
No sirve como criterio diferenciado de documentos.*

Es posible que si existen en el idioma, los usuarios no los escriban.

Es un tipo de palabra (normalizado), que es una entrada en el diccionario.

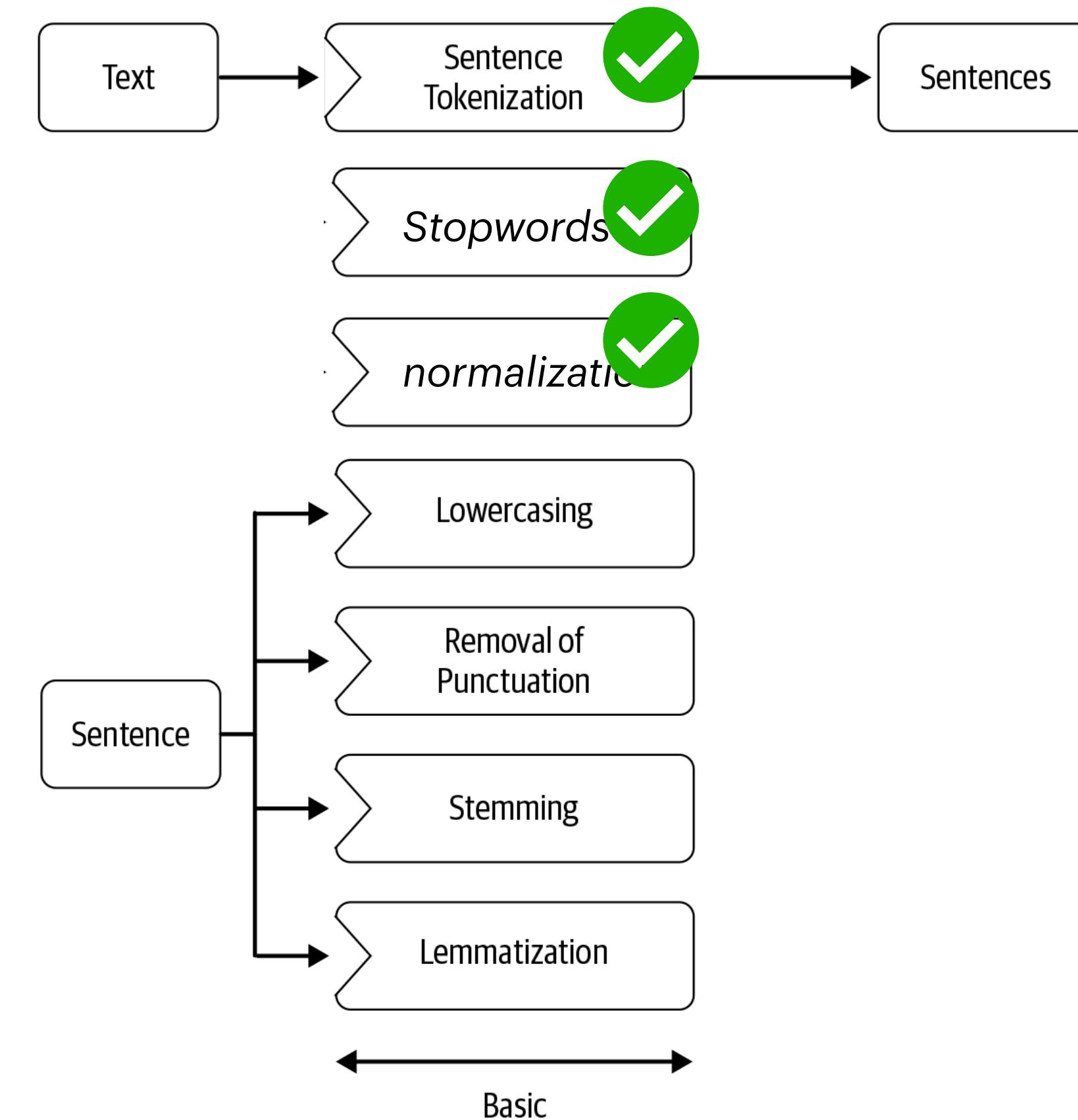
Normalización y Tokenización son dependientes del idioma y pueden estar entrelazadas con la detección del idioma.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Lematización

Reducir las formas flexivas/variantes a la forma base.

Ej.:

Pan: panadero - panadería - panecillo

Pescar: pescado - pesquero - pescador - pescadería

En inglés....

be: am - are - is (verbal)

car: car - cars - car's - cars' (nominal)

Implica hacer una reducción “adecuada”.

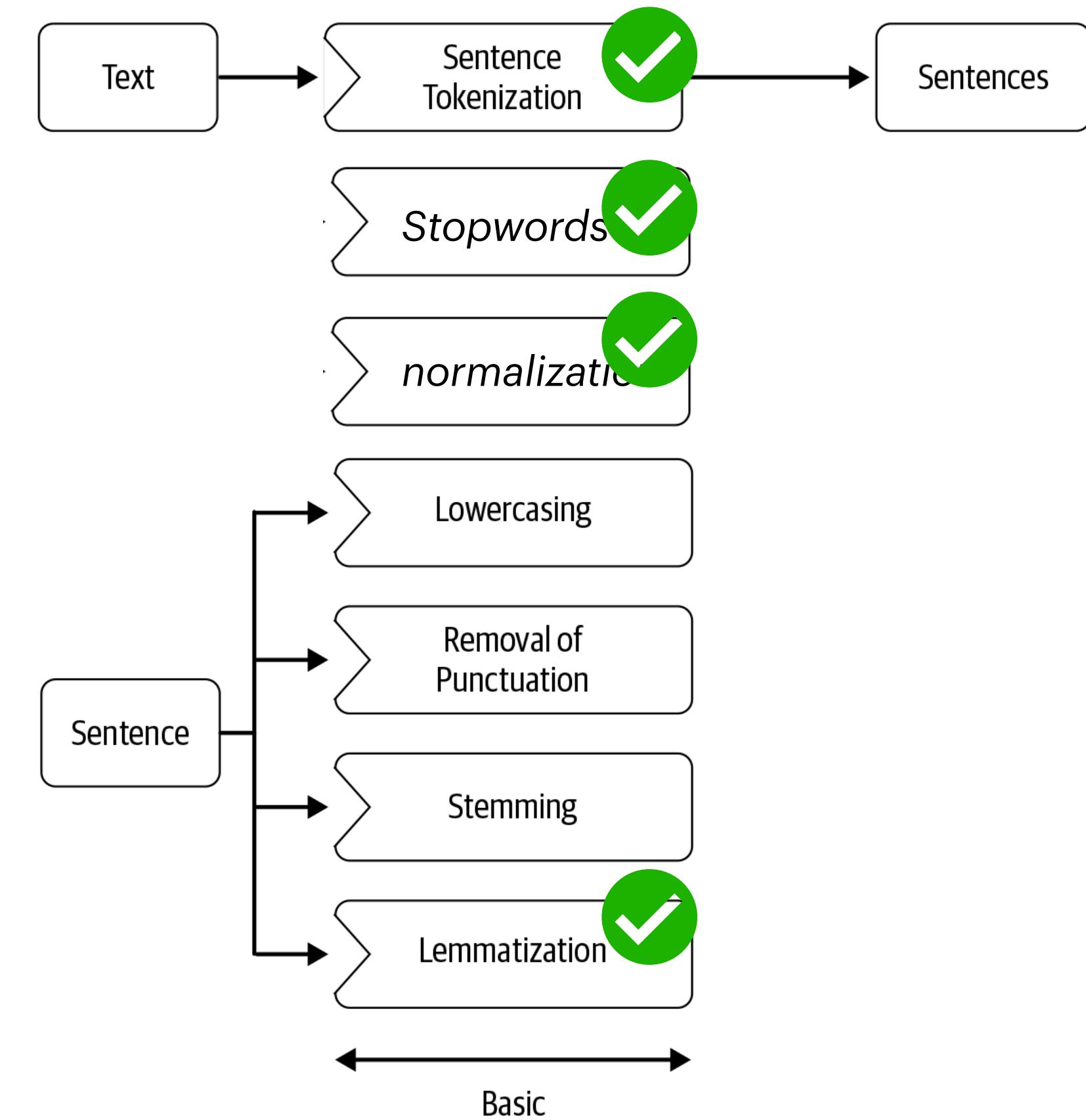
Requiere diccionarios con la morfología de las palabras.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Stemming

Reglas de corte
que se hacen para
cada lenguaje.

Ej.: *automatizar(s), automático, automatización*: todo reducido a *automat*
automate(s), automatic, automation: todo reducido a *automat*

En inglés.... Stemming <http://www.tartarus.org/~martin/PorterStemmer/>

For example compressed and
compression are both accepted
as equivalent to compress

For exampl compress and
compress ar both accept as
equival to compress

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Stemming

*Reglas de corte
que se hacen para
cada lenguaje.*

Ej.:

ATIONAL	-> ATE	relational -> relate
TIONAL	-> TION	conditional -> condition
ENCI	-> ENCE	valenci -> valence
ANCI	-> ANCE	hesitanci -> hesitance
IZER	-> IZE	digitizer -> digitize
ABLI	-> ABLE	conformabli -> conformable
ALLI	-> AL	radicalli -> radical
ENTLI	-> ENT	differentli -> different
ELI	-> E	vileli -> vile
OUSLI	-> OUS	analogousli -> analogous

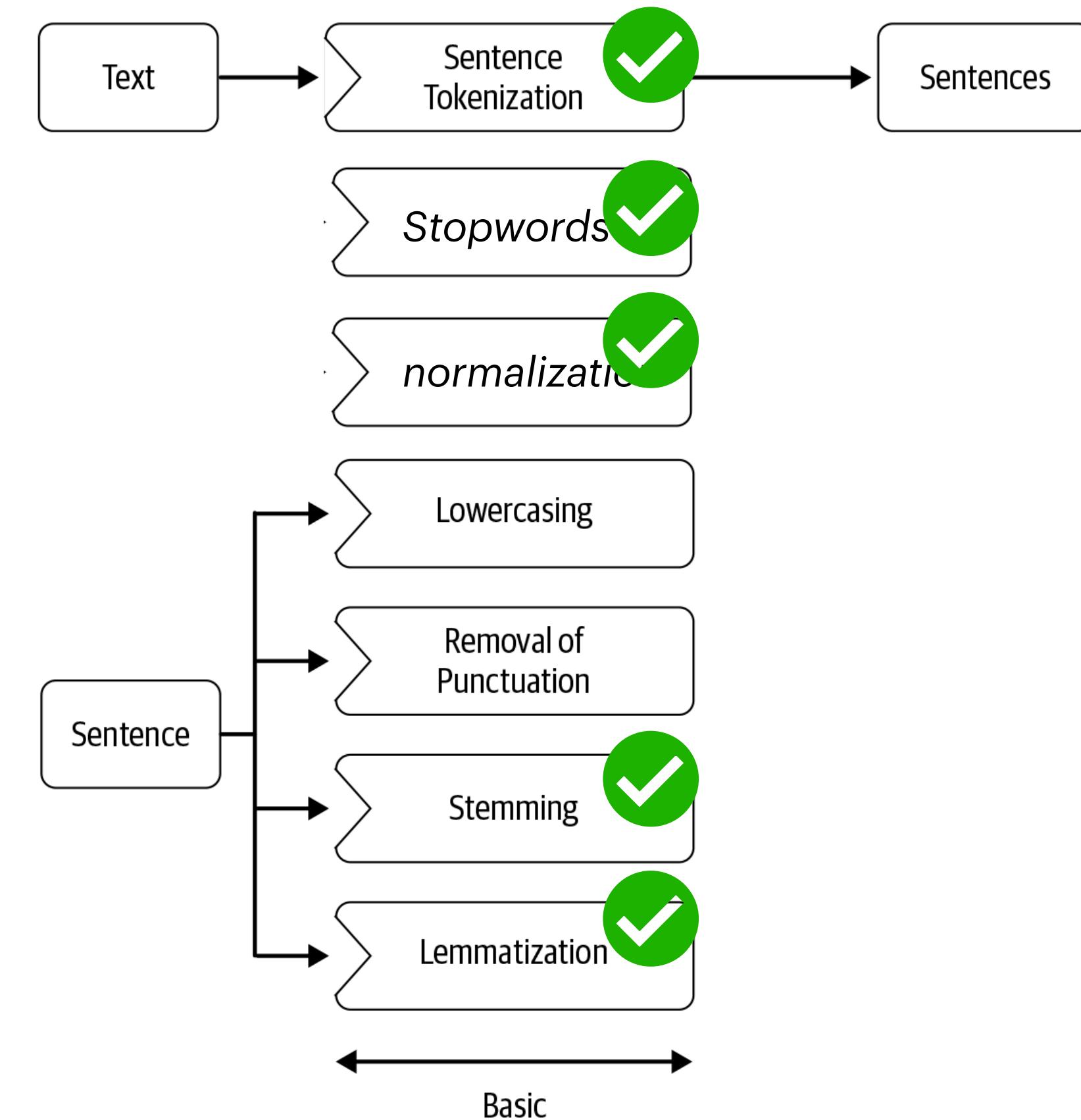
Se selecciona la regla con el sufijo más largo.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.

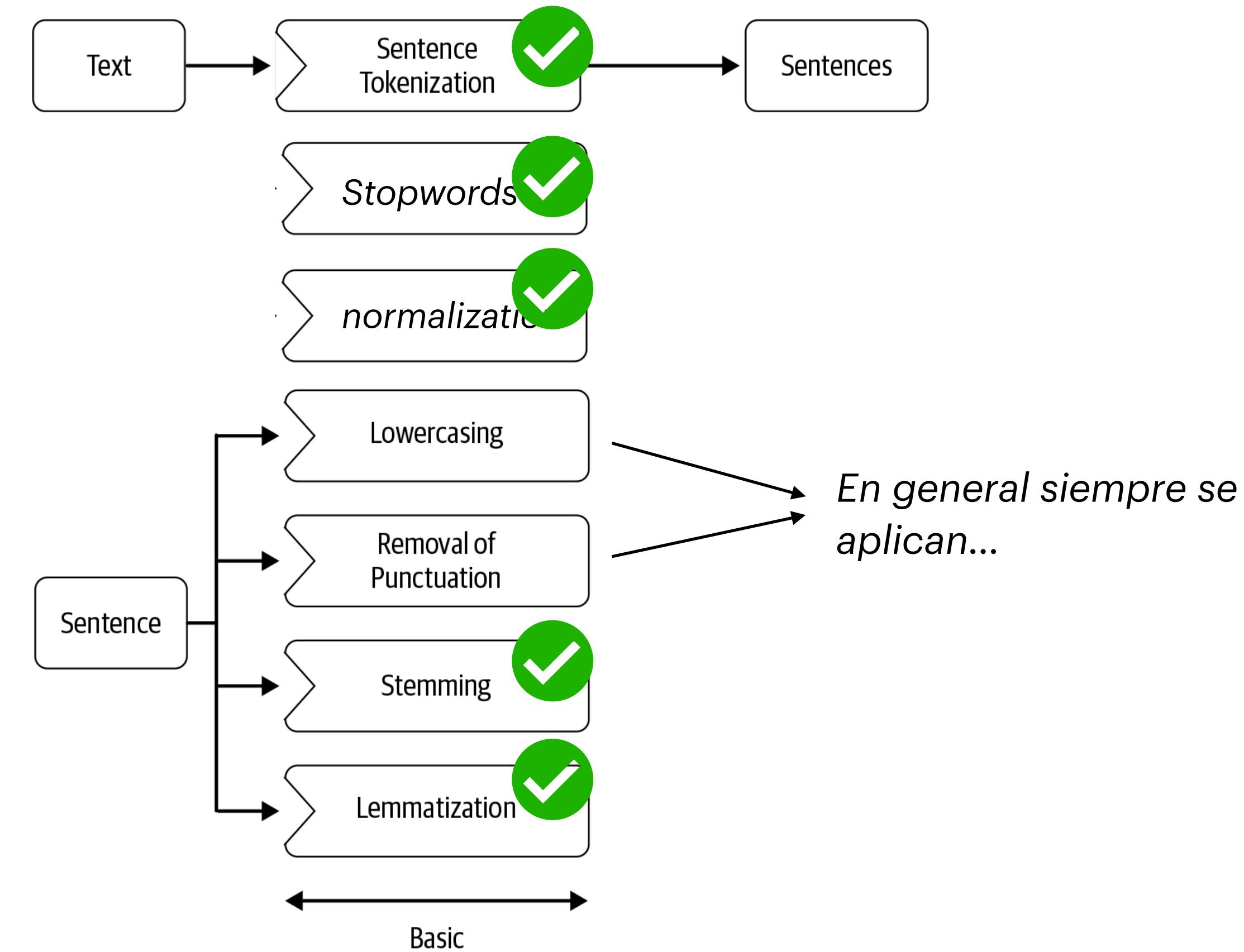


Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.

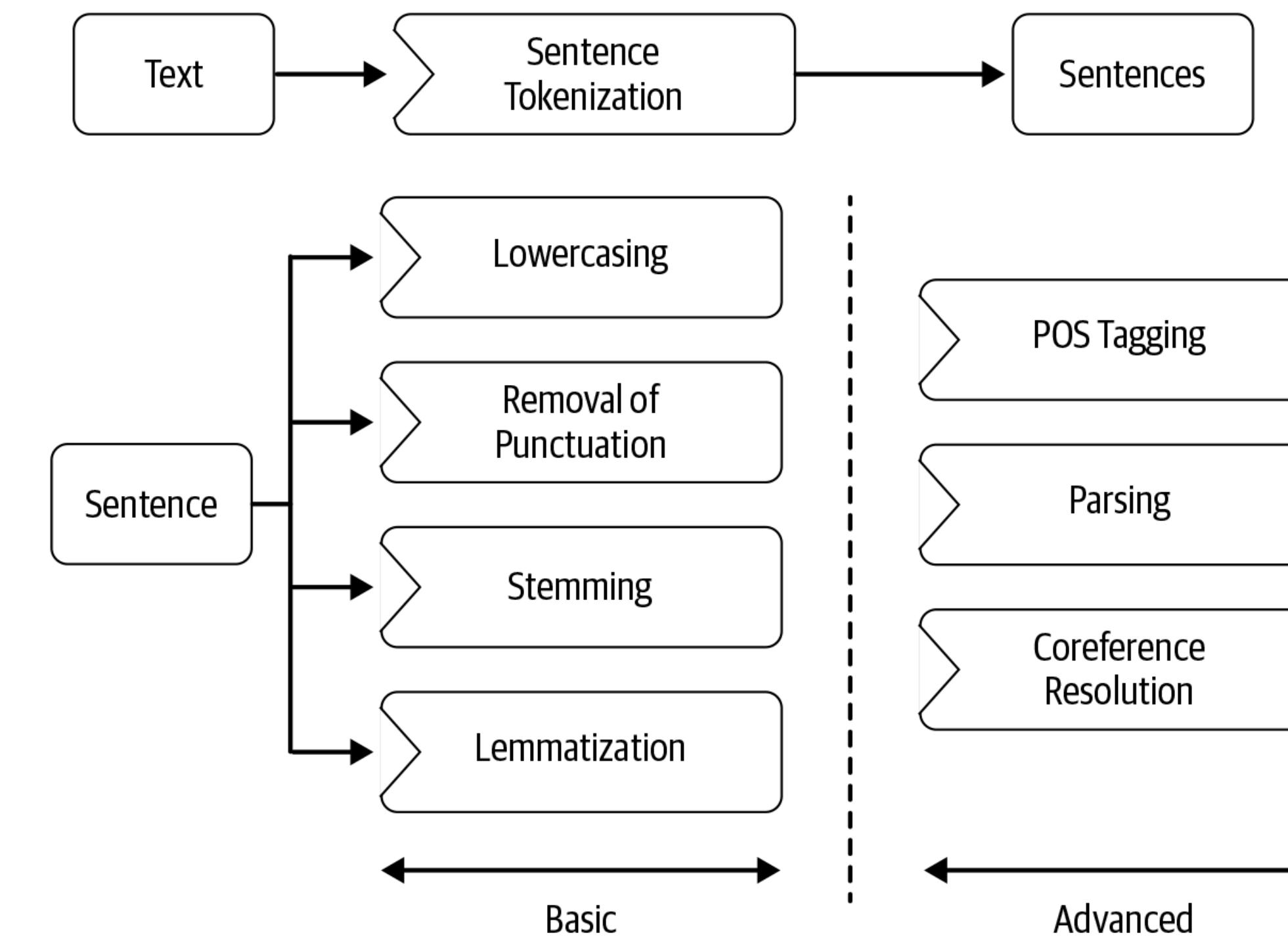


Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.

Input

Chaplin wrote, directed, and composed the music for most of his films.

Tokenization with Lemmatization

Chaplin write . direct and compose the music for most of he film .
Chaplin wrote, directed, and composed the music for most of his films.

POS Tagging

NNP VBD . VBD CC VBN DT NN IN JJS IN PRP\$ NNS .
Chaplin wrote, directed, and composed the music for most of his films.



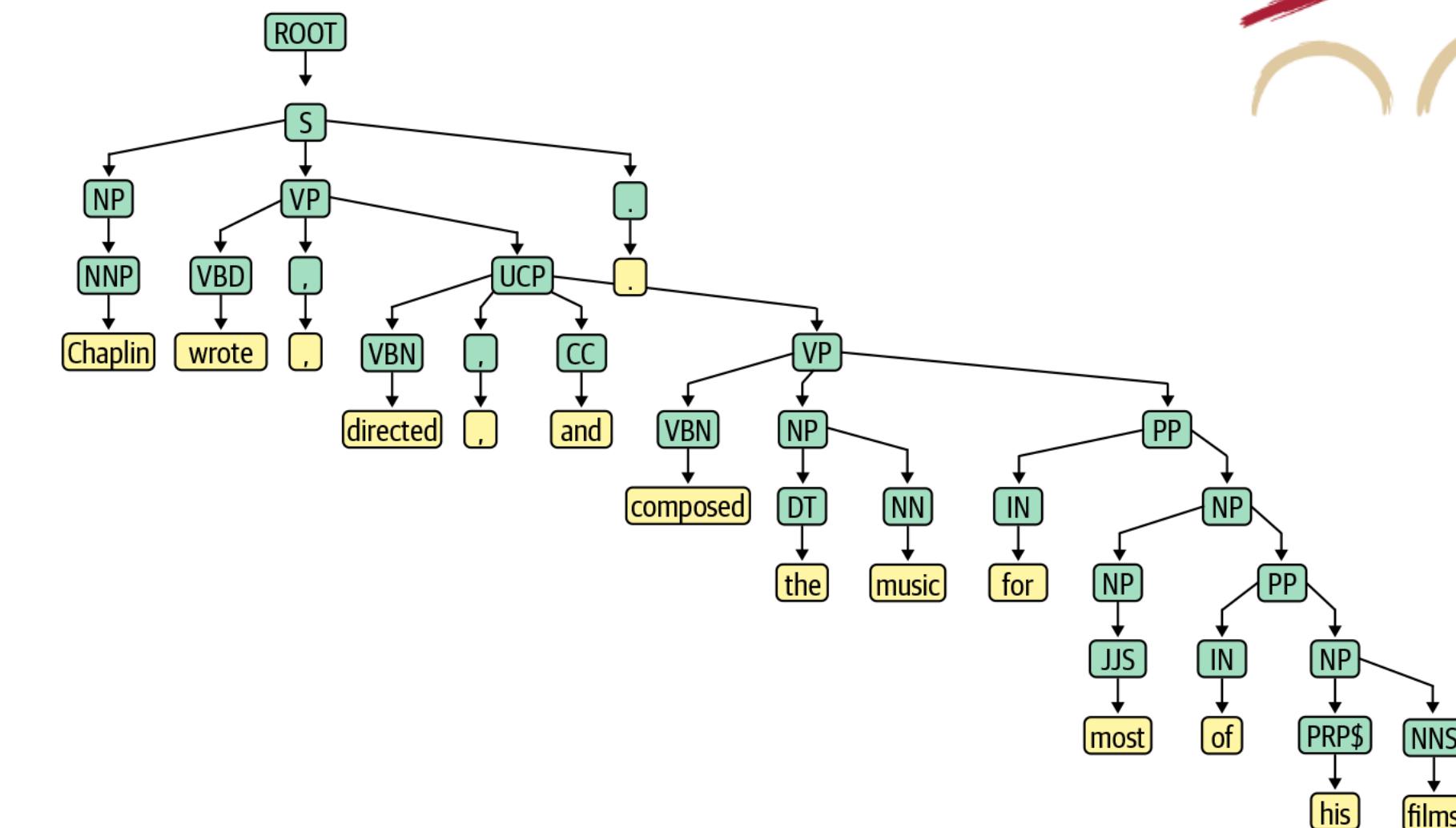
Pre-procesamiento

Componentes en la construcción de un modelo de NLP

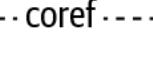
Pre-procesamiento de texto

Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.

Parse Tree



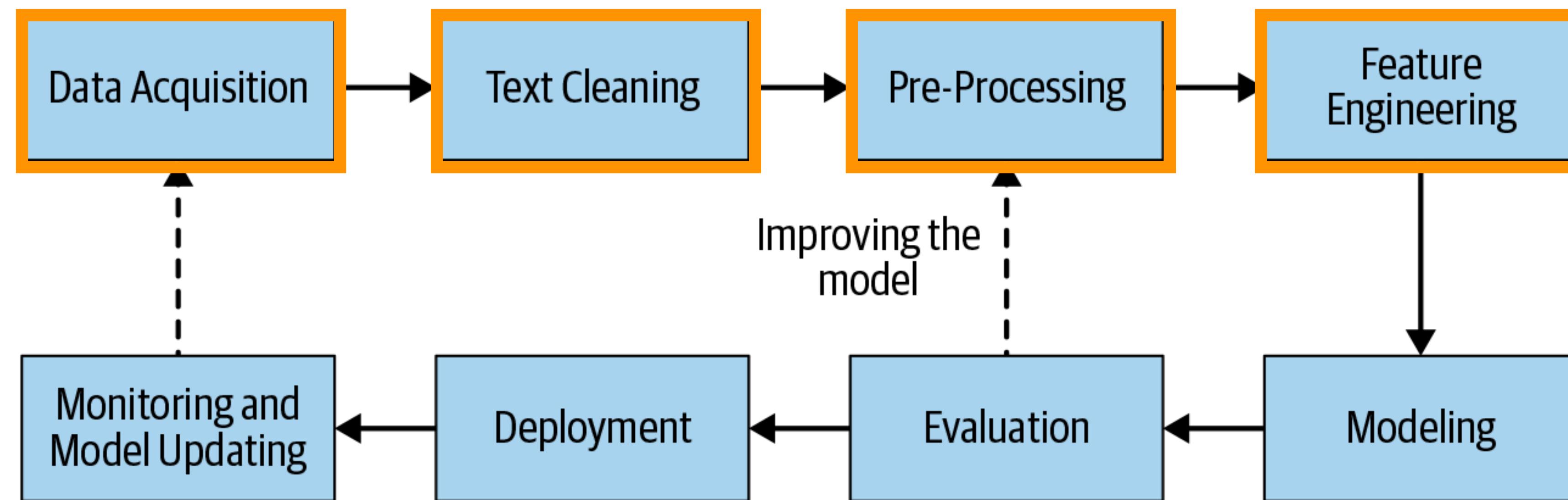
Coreference Resolution

   Chaplin wrote, directed, and composed the music for most of his films.



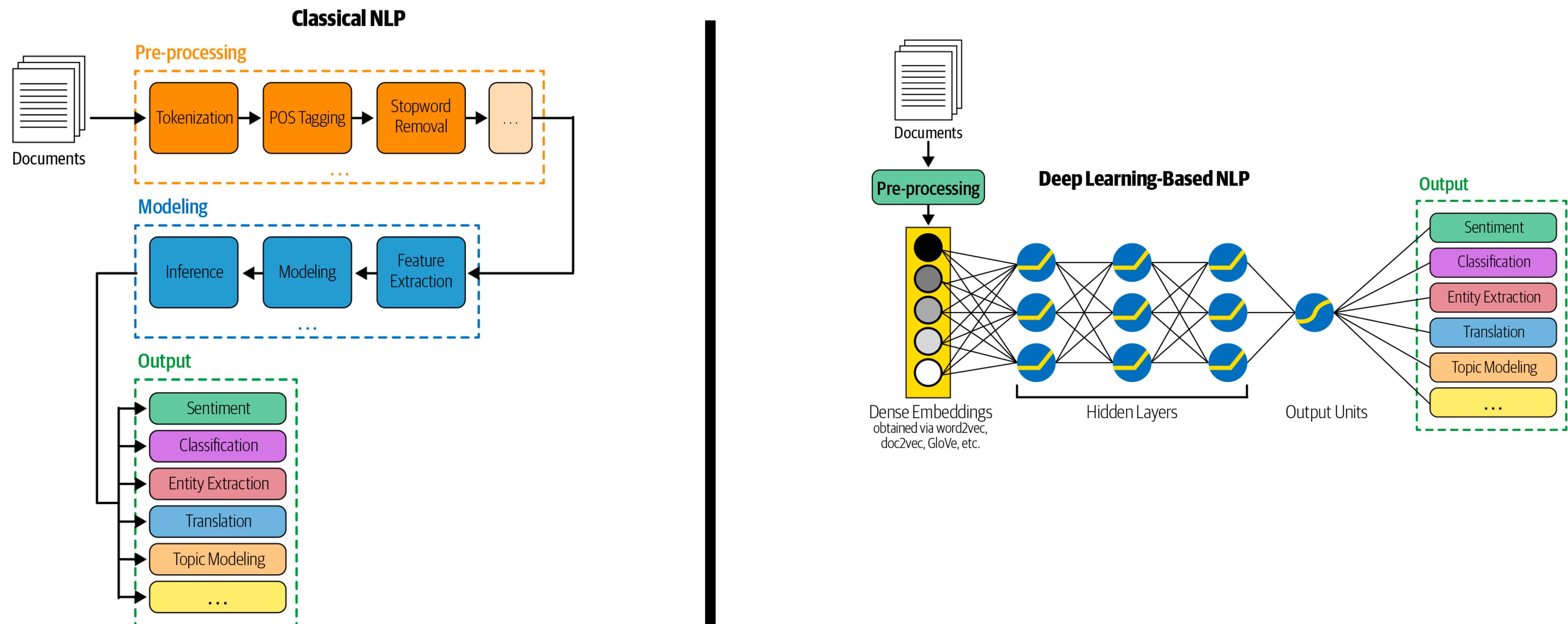
NLP pipeline

Componentes en la construcción de un modelo de NLP



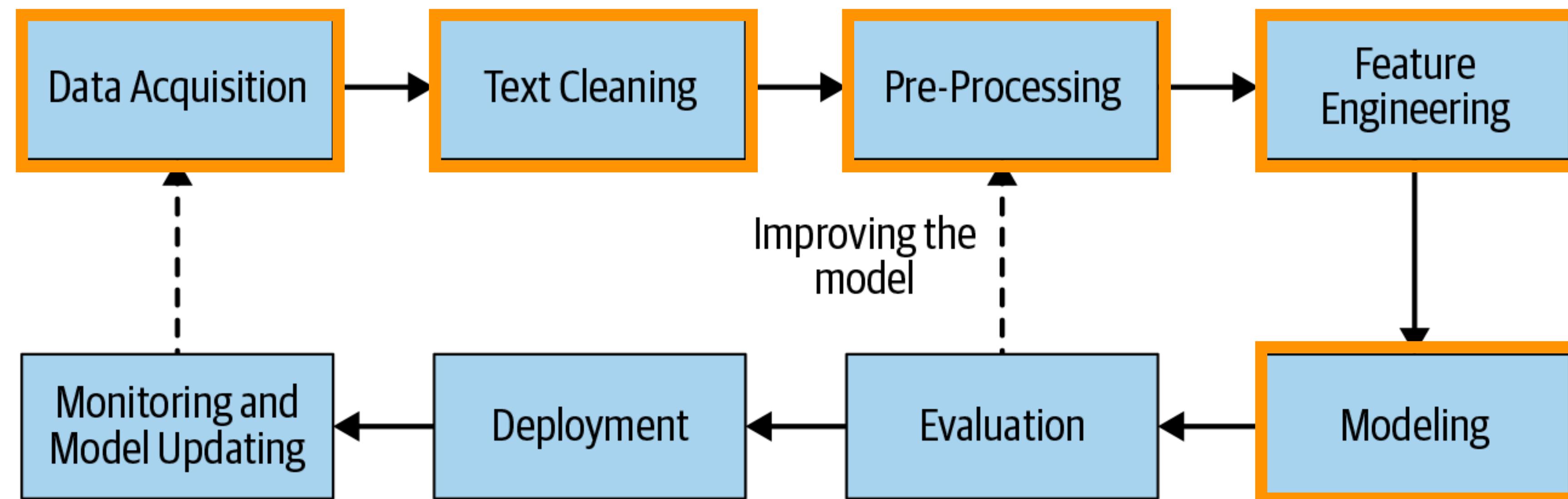
Ingeniería de Características

Comparación de enfoque



NLP pipeline

Componentes en la construcción de un modelo de NLP



Modelamiento

Componentes en la construcción de un modelo de NLP

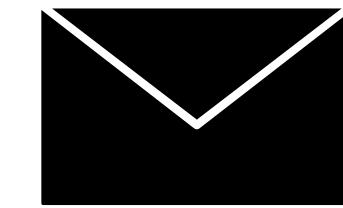
Construyendo un modelo

Después de identificar el problema analítico, selecciona el mejor modelo de acuerdo con el contexto del problema.

¡Haz un inicio rápido!

Inicia con una Heurística

Sistema de expresiones regulares / filtro de páginas con dominios sospechosos / etc...



Revisa proveedores de NLP

Sí es posible revisa las diferentes APIs que están listas para usar desde Microsoft, Google, IBM...

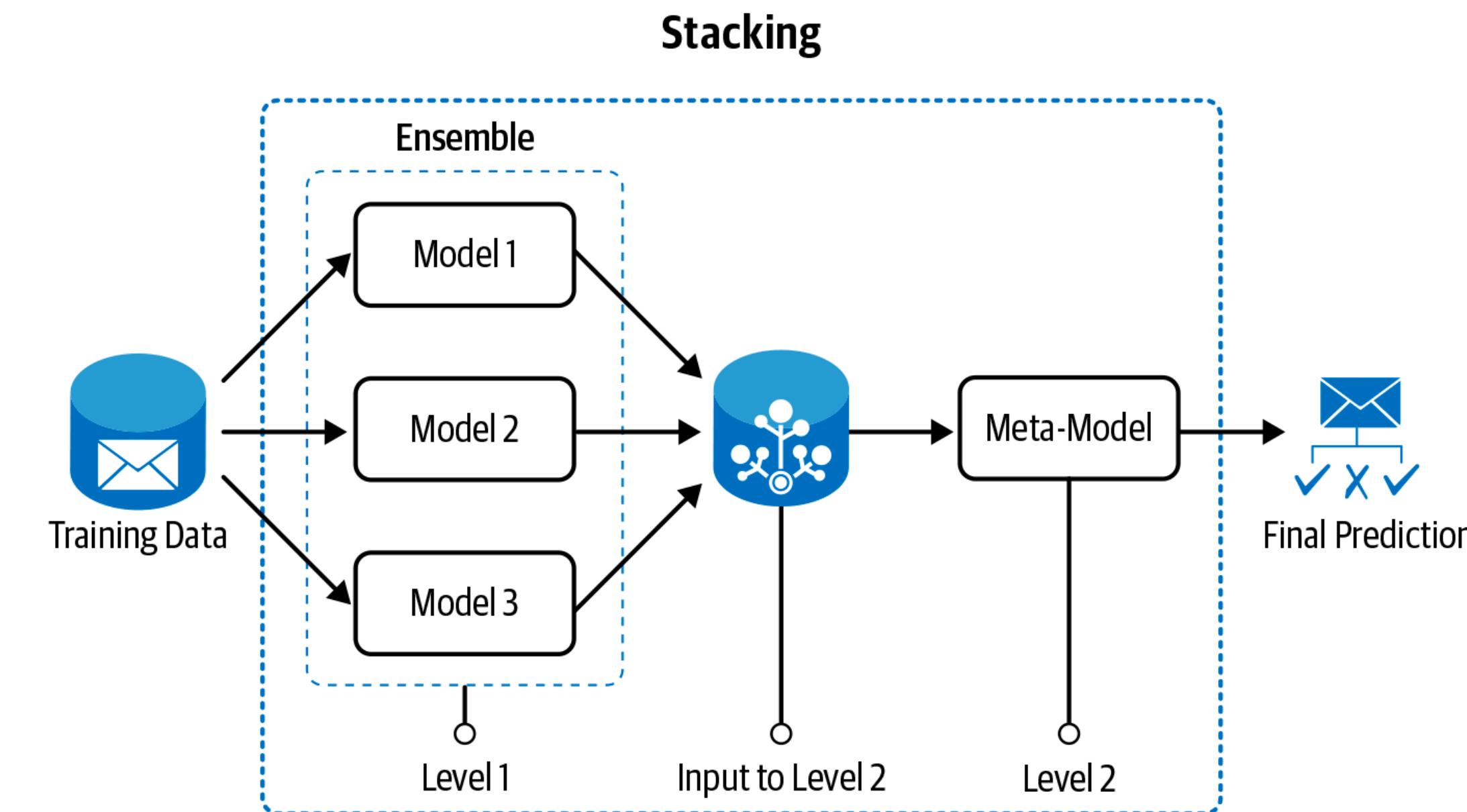


Modelamiento

Componentes en la construcción de un modelo de NLP

Construyendo un modelo

Es una práctica común no tener un solo modelo, sino usar una colección de modelos.



Modelamiento

Componentes en la construcción de un modelo de NLP

Mejora Ingeniería de características

Por lo general, al mejorar los pasos de ingeniería de característica, ya sea un nuevo paso o una nueva característica se traduce en una mejora en el rendimiento.

Aprendizaje por transferencia

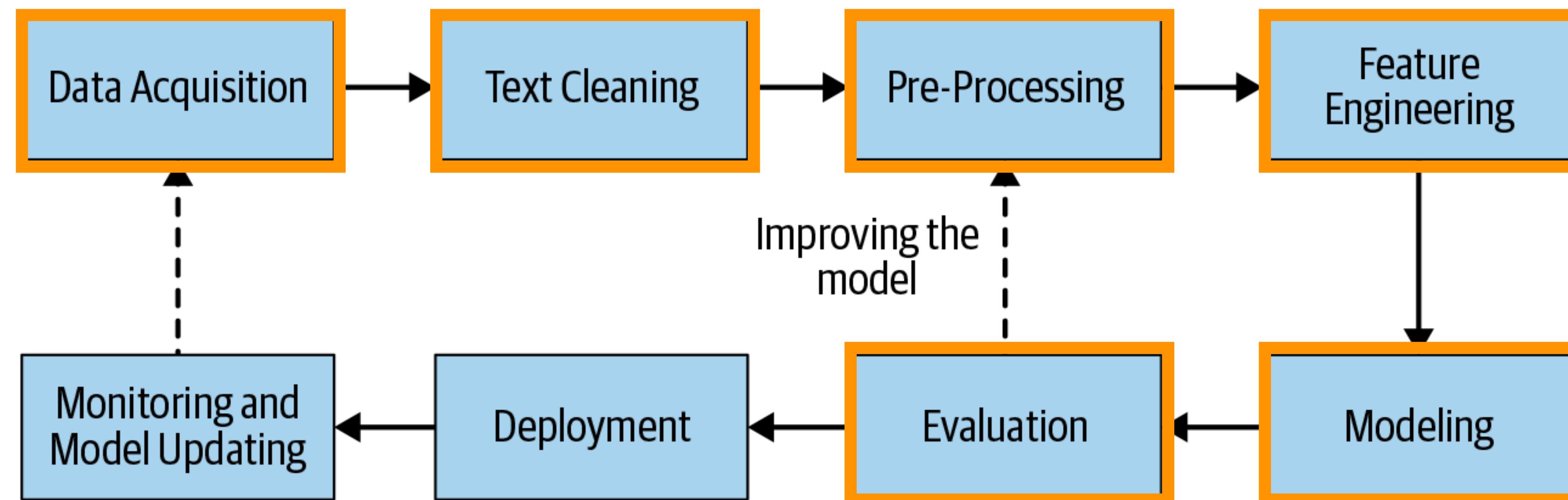
A menudo los modelos requieren de contexto externo, más allá del conjunto de datos para que el modelo comprenda bien el lenguaje y el problema

Vuelve a una heurística

Es común que los modelos llegan a caer en un error, mientras se sutura o mejora, cúbrelo con una heurística.

NLP pipeline

Componentes en la construcción de un modelo de NLP



Evaluación

Componentes en la construcción de un modelo de NLP

Evaluación Intrínseca

Métricas del modelo

- *AUC*
- *Precision*
- *Recall*

Evaluación Intrínseca

Métricas del negocio

- *Open rate*
- *Ordenes Inc.*
- *Buyers Inc.*

Caso de estudio

Componentes en la construcción de un modelo de NLP

COTA by Uber

TF-IDF

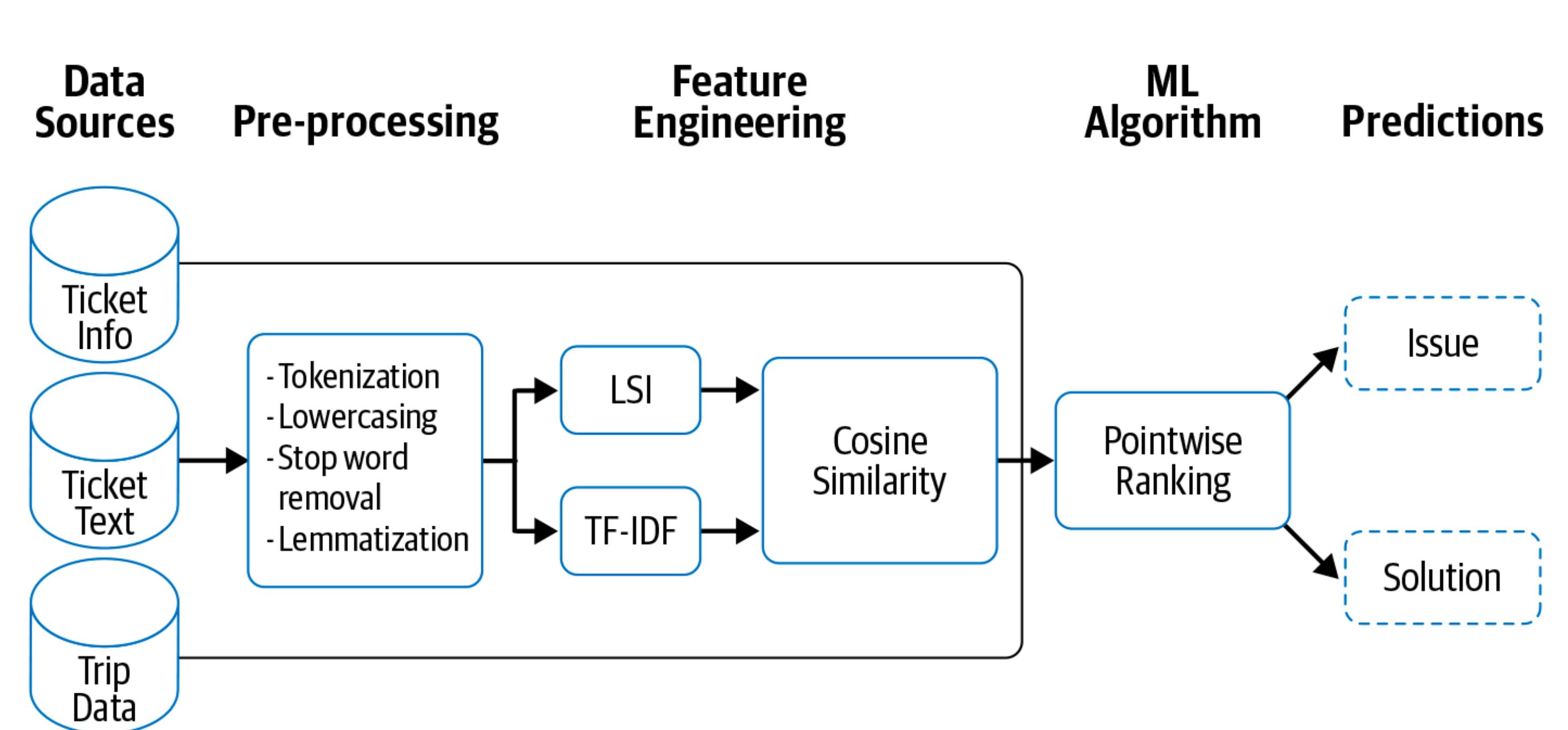
Frecuencia de termino y frecuencia de documento inversa

LSI

Indexación semántica latente

Cosine Similarity

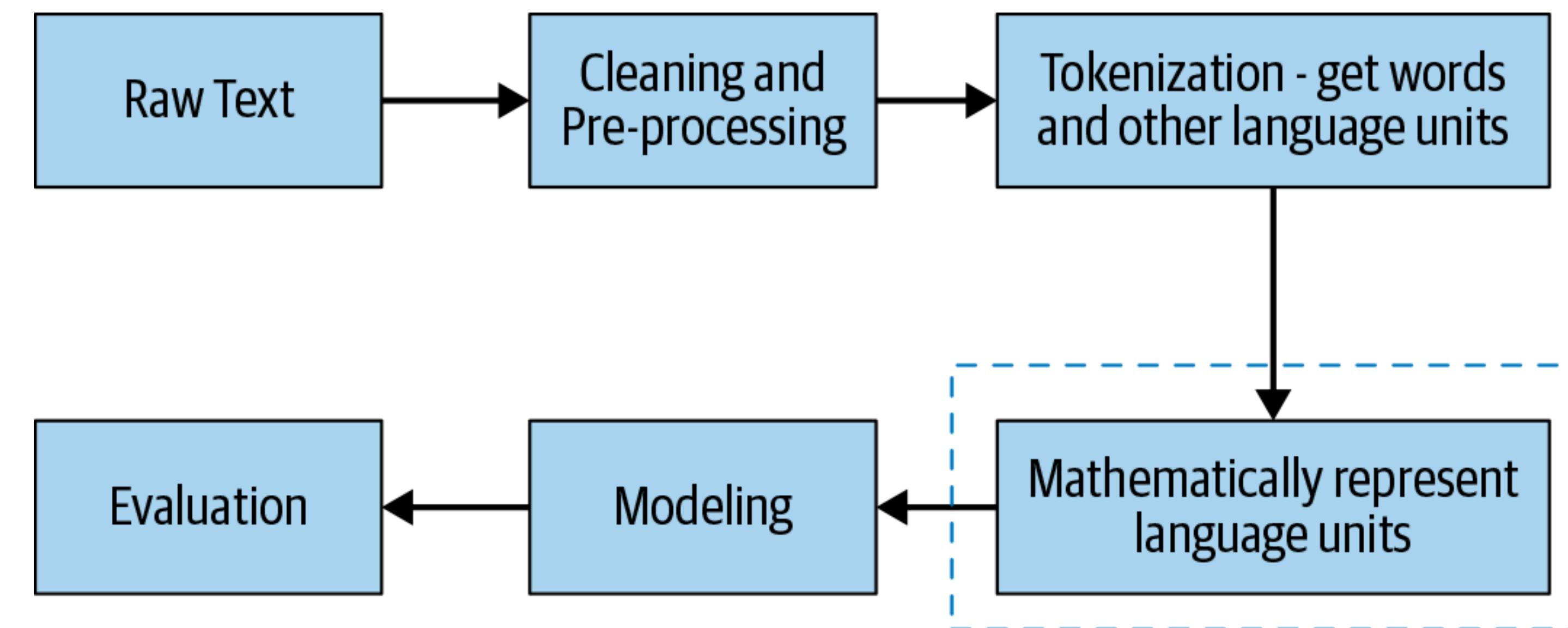
Distancia entre dos vectores



Text representation

Representación de Texto

Feature Engineering



Representación de Texto

Feature Engineering



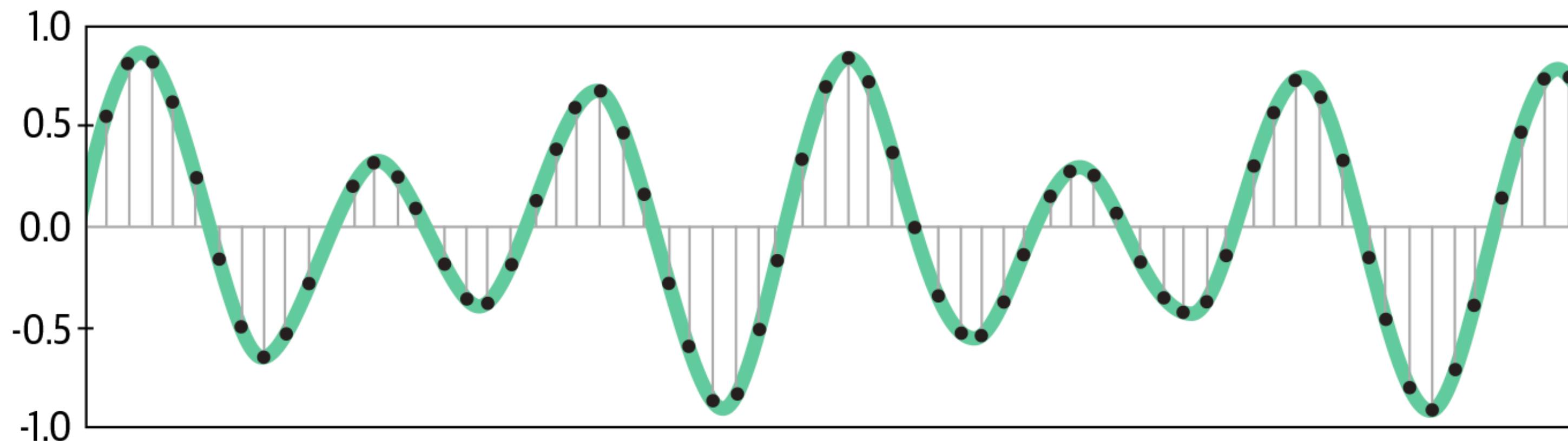
What We See

08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48

What Computers See

Representación de Texto

Feature Engineering



```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41,
-169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448,
-397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451,
1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499,
-488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148,
-1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325,
350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

Text Representation

Feature Engineering

1. Space Vector Models

*Todas las representaciones
de texto son SVM.*

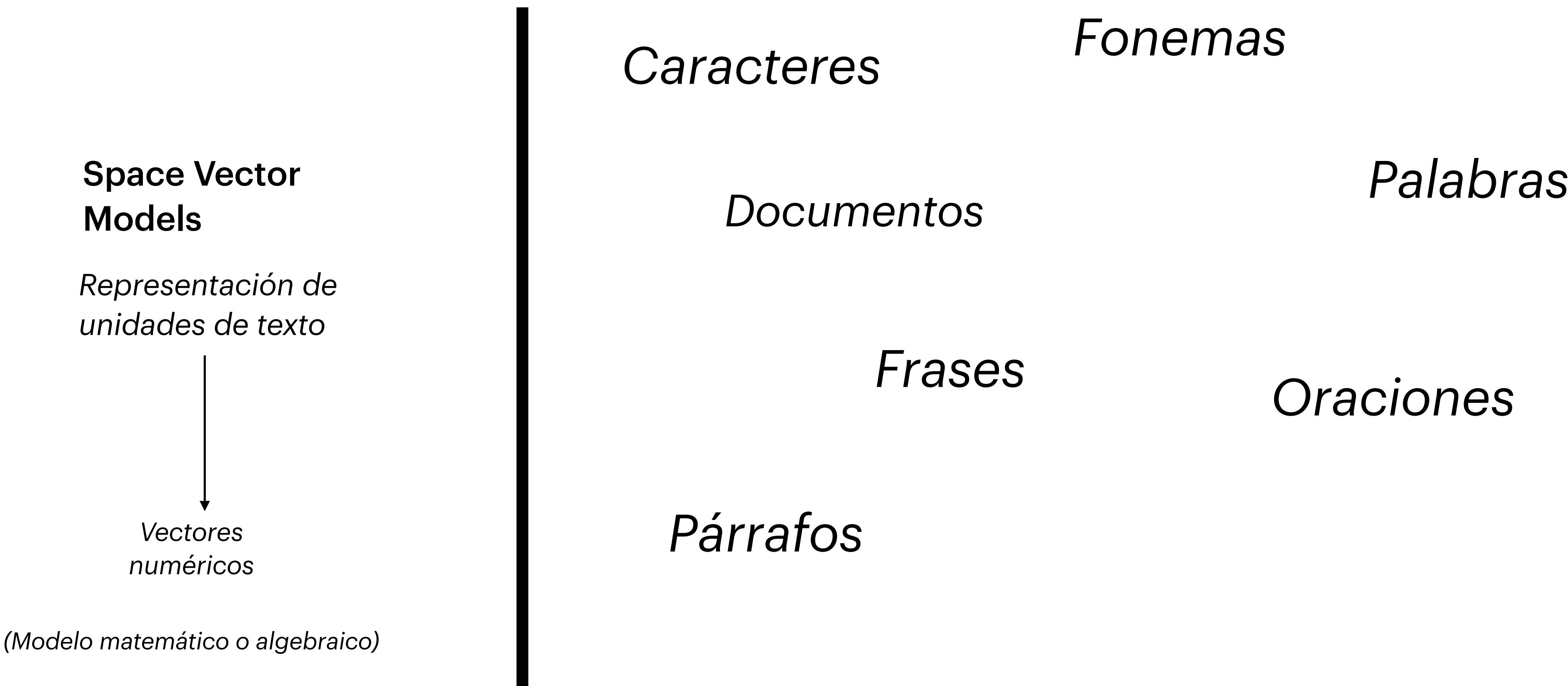
2. Basic Vectorization Approaches

3. Word Embeddings

4. Visualizing Embeddings***

Text Representation

Space Vector Models



Text Representation

Space Vector Models

Space Vector Models

Representación de
unidades de texto



Vectores
numéricos

(Modelo matemático o algebraico)

Distancia del coseno:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Donde A_i y B_i son las i -ésimas componentes de los vectores A y B .

Text Representation

Feature Engineering

- 1.**
Space Vector Models

Todas las representaciones de texto son SVM.
- 3.**
Word Embeddings

- 2.**
Basic Vectorization Approaches
- 4.**
Visualizing Embeddings***

Text Representation

Basic Vectorization Approaches

Basic Vectorization Approaches

Ej. asignación de cada palabra en el vocabulario (V) del corpus de texto aun ID única (valor entero)

- D1 Perro muerde a hombre.
- D2 Hombre muerde a perro.
- D3 El perro come carne.
- D4 El hombre come comida.

Vocabulario del corpus:

[perro, muerde, hombre, come, carne, comida]



Vector de tamaño 6

Texto procesado:
minúsculas, sin puntuación, etc..

Texto tokenizado:
Cadena de texto dividida en tokens.

Text Representation

Basic Vectorization Approaches

One-Hot Encoding

Vocabulario
basado en
asignación de Id
por palabras.

W Palabra del texto

V Conjunto del vocabulario del corpus

W_{id} Índice de palabras del texto, donde W_{id} entre $\{1, |V|\}$

Representación binaria de cada palabra en un vector de
tamaño $|V|$

Index = W_{id}

Text Representation

Basic Vectorization Approaches

One-Hot Encoding

Vocabulario
basado en
asignación de Id
por palabras.

W Palabra del texto

V Conjunto del vocabulario del corpus

W_{id} Índice de palabras del texto, donde W_{id} entre $\{1, |V|\}$

perro = 1, muerde = 2, hombre = 3, carne = 4, comida = 5, come = 6

Ej.: D1: “perro muerde a hombre”

$[[1\ 0\ 0\ 0\ 0\ 0]]$ perro

$[0\ 1\ 0\ 0\ 0\ 0]$ muerde

$[0\ 0\ 1\ 0\ 0\ 0]$ hombre

Ej.: D4: “hombre come comida”

$[[0\ 0\ 1\ 0\ 0\ 0]]$ hombre

$[0\ 0\ 0\ 0\ 1\ 0]$ comida

$[0\ 0\ 0\ 0\ 0\ 1]$ come

Text Representation

Basic Vectorization Approaches

One-Hot Encoding

*Vocabulario
basado en
asignación de Id
por palabras.*

Pros

*Intuitiva y fácil
de implementar*

;

;

;

;

Text Representation

Basic Vectorization Approaches

One-Hot Encoding

Vocabulario basado en asignación de Id por palabras.

Pros

Intuitiva y fácil de implementar

*Tienen la misma distancia:
[correr, corre, manzana]*

Ej. Pasarle la palabra 'fruta' al modelo.

Contra

1. *Tamaño del vector proporcional al tamaño del vocabulario*
2. *No proporciona una longitud fija intra documentos.*
3. *Palabras como unidades atómicas y no tiene noción de (des)similitud.*
4. *No maneja un esquema fuera del vocabulario (OOV).*

Text Representation

Basic Vectorization Approaches

Bag of Words (BOW)

Calificamos cada palabra en V por su recuento de ocurrencias en el documento.

↓
Recuento de palabras

W Palabra del texto

V Conjunto del vocabulario del corpus

W_{id} Índice de palabras del texto, donde W_{id} entre $\{1, |V|\}$

perro = 1, muerde = 2, hombre = 3, carne = 4, comida = 5, come = 6

Ej.: D1: "perro muerde a hombre"

[1 1 1 0 0 0]	perro
	muerde
	hombre

Ej.: D4: "hombre come comida"

[0 0 1 0 1 1]	hombre
	comida
	come

Text Representation

Basic Vectorization Approaches

Bag of Words (BOW)

Calificamos cada palabra en V por su recuento de ocurrencias en el documento.

↓
Recuento de palabras

Pros

1. *Intuitiva y fácil de implementar*
2. *Captura la similitud semántica de los documentos*
3. *Codificación de longitud fija para cualquier oración arbitraria*

Ej. espacio euclíadiano entre D_1 y D_2 es 0. En comparación con D_1 y D_4 que es 2.

Text Representation

Basic Vectorization Approaches

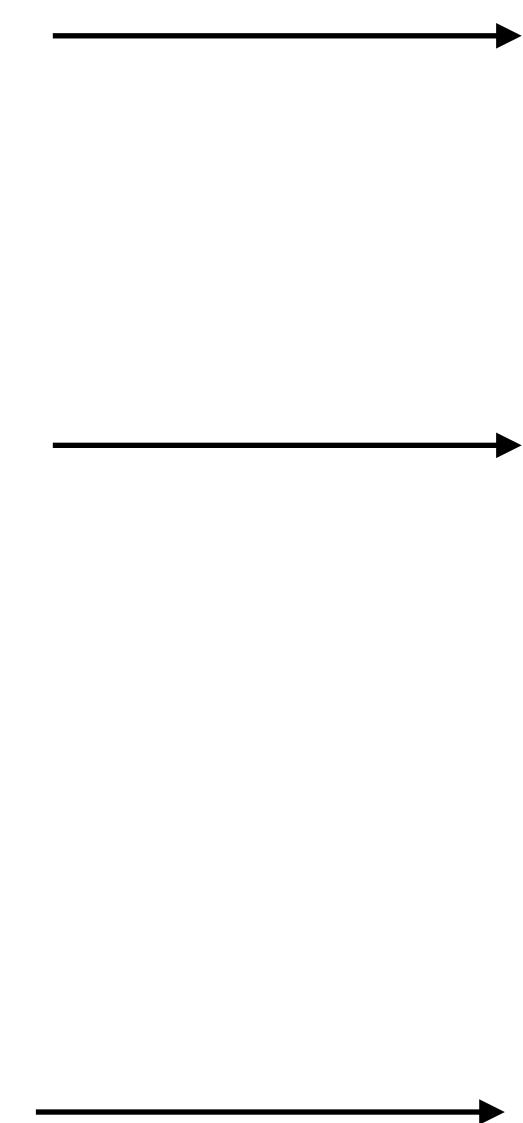
Bag of Words (BOW)

Calificamos cada palabra en V por su recuento de ocurrencias en el documento.

↓
Recuento de palabras

Contra

1. *Tamaño del vector aumenta con el tamaño del vocabulario.*
2. *No Captura la similitud semántica de palabras que significan lo mismo*
3. *No tiene forma de manejar palabras fuera del vocabulario.*
4. *El orden de las palabras se pierde*



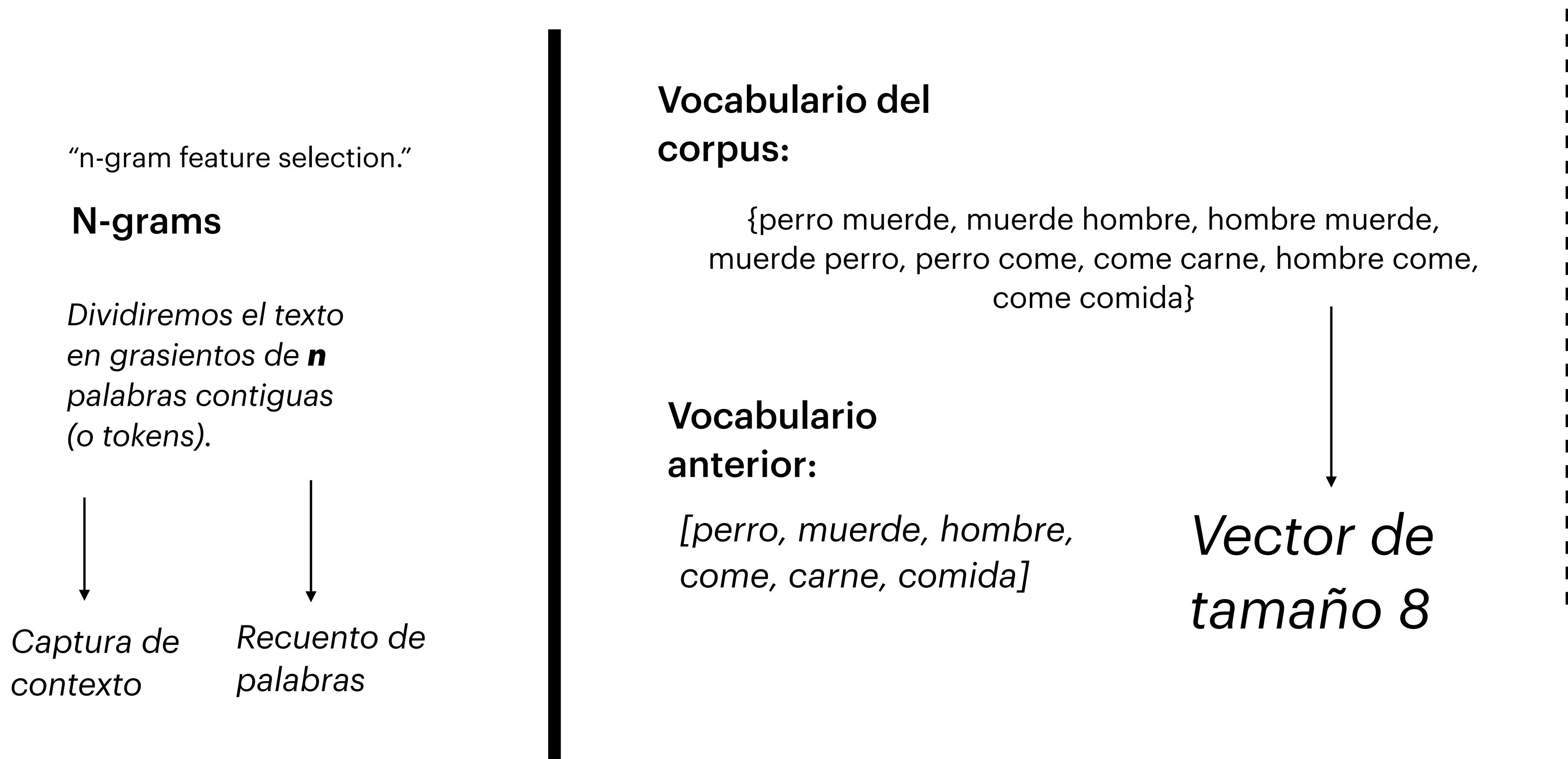
Hace necesario limitar vocabulario a n número de palabras más frecuentes.

Ej.: "I ran", "I run" y "I ate"

D1 y D2 tendrán la misma representación en este esquema.

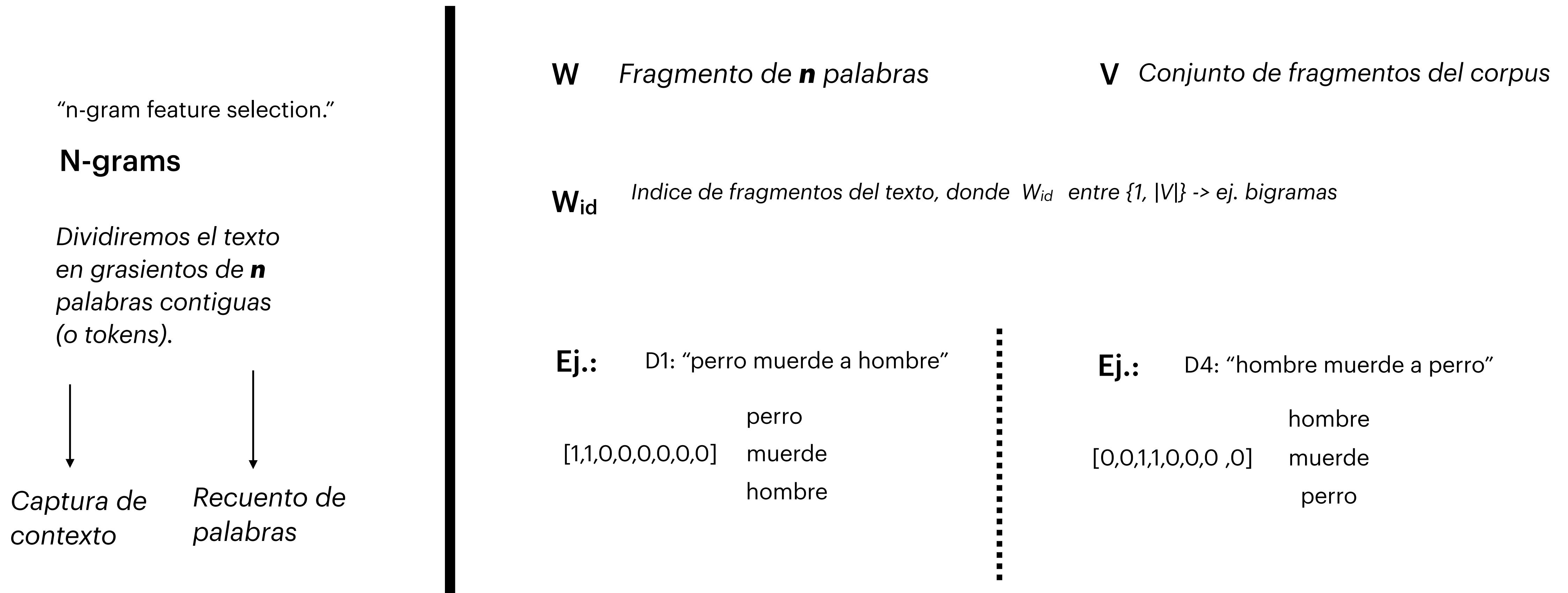
Text Representation

Basic Vectorization Approaches



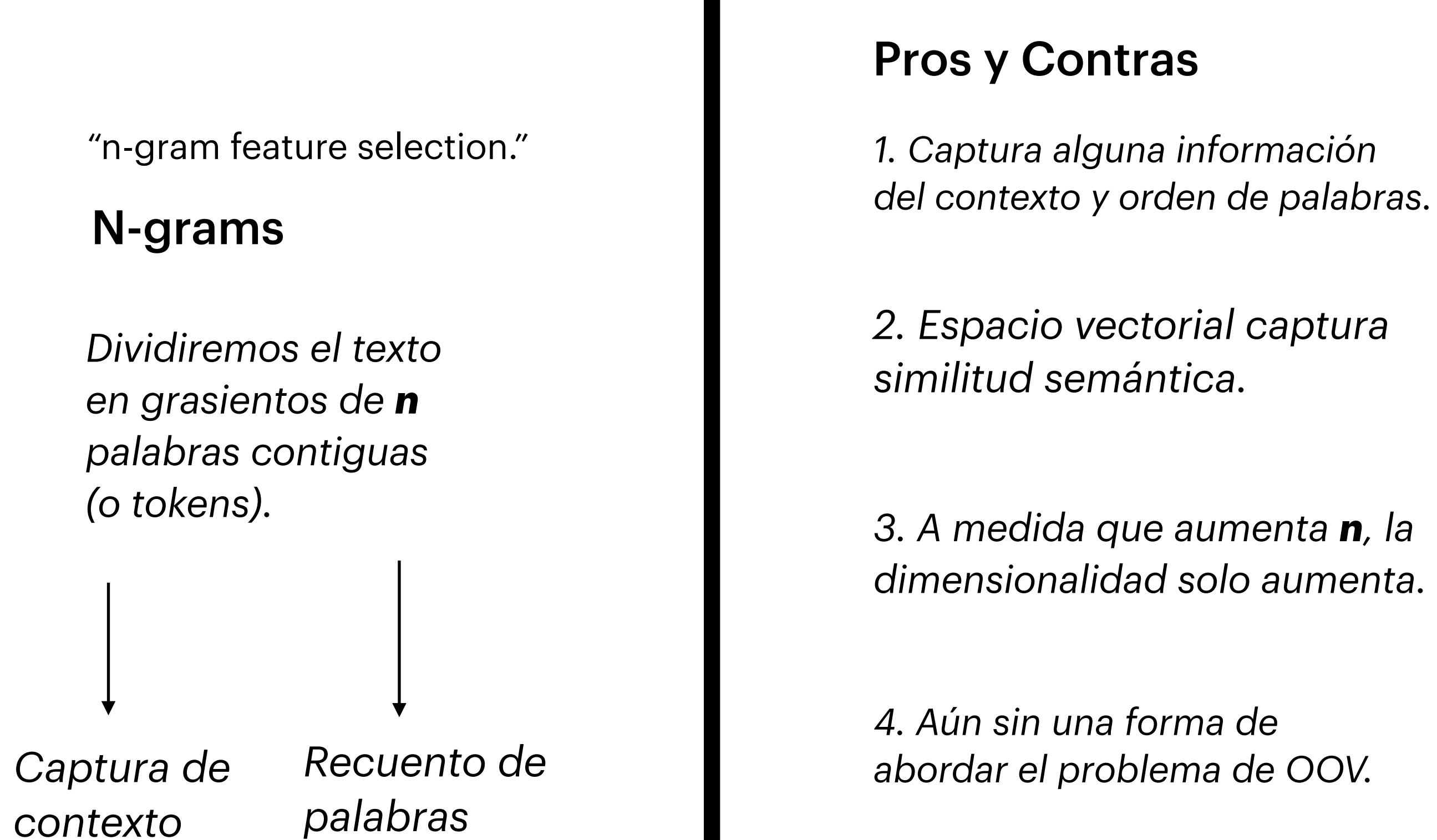
Text Representation

Basic Vectorization Approaches



Text Representation

Basic Vectorization Approaches



Pros y Contras

1. Captura alguna información del contexto y orden de palabras.
2. Espacio vectorial captura similitud semántica.
3. A medida que aumenta **n**, la dimensionalidad solo aumenta.
4. Aún sin una forma de abordar el problema de OOV.

Text Representation

Basic Vectorization Approaches

t : Término

d : Documento

Frecuencia de término (TF).

Frecuencia de un término o una palabra en un documento.

TF-IDF

Término de frecuencia - frecuencia de documento inversa, cuantifica la importancia de una palabra en relación con otras palabras en el documento y en el corpus.



Importancia de palabras

$$\text{TF}(t, d) = \frac{\text{(Number of occurrences of term } t \text{ in document } d)}{\text{(Total number of terms in the document } d)}$$

Frecuencia de documento inversa (IDF).

Mide la importancia de término en el corpus

$$\text{IDF}(t) = \log_e \frac{\text{(Total number of documents in the corpus)}}{\text{(Number of documents with term } t \text{ in them)}}$$

Text Representation

Basic Vectorization Approaches

TF-IDF

Término de frecuencia - frecuencia de documento inversa, cuantifica la importancia de una palabra en relación con otras palabras en el documento y en el corpus.



Importancia de palabras

Word	TF score	IDF score	TF-IDF score
dog	$\frac{1}{3} = 0.33$	$\log_2(4/3) = 0.4114$	$0.4114 * 0.33 = 0.136$
bites	$\frac{1}{6} = 0.17$	$\log_2(4/2) = 1$	$1 * 0.17 = 0.17$
man	0.33	$\log_2(4/3) = 0.4114$	$0.4114 * 0.33 = 0.136$
eats	0.17	$\log_2(4/2) = 1$	$1 * 0.17 = 0.17$
meat	$1/12 = 0.083$	$\log_2(4/1) = 2$	$2 * 0.083 = 0.17$
food	0.083	$\log_2(4/1) = 2$	$2 * 0.083 = 0.17$

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

La representación del vector TF-IDF es el mismo puntaje de TF-IDF, para cada término en ese documento.

Ejemplo, D1:

Perro	muerde	hombre	come	carne	alimento
0.136	0.17	0.136	0	0	0

Text Representation

Basic Vectorization Approaches

TF-IDF

Término de frecuencia - frecuencia de documento inversa, cuantifica la importancia de una palabra en relación con otras palabras en el documento y en el corpus.



Importancia
de palabras

Pros y Contras

1. Mejor que los otros métodos de vectorización vistos anteriormente.
2. Sufre de la maldición de la alta dimensionalidad.

TF-IDF en SKLEARN:

1. Ligera modificación en la fórmula IDF.
2. Disposiciones para dividir por cero.
3. No ignora por completo los términos que aparecen en todos los documentos.



Text Representation

Basic Vectorization Approaches

Representaciones discretas:

Tratan a las unidades del lenguaje (palabras, n-grams, etc.) como unidades atómicas.

Lo que dificulta la capacidad para captar relaciones entre palabras.

Dimensionalidad:

Los vectores generalmente son dispersos y de alta dimensión.

La dimensionalidad aumenta con el tamaño del vocabulario, siendo la mayoría de los valores cero.

- 1. Dificulta la capacidad de aprendizaje.*
- 2. Alta dimensionalidad, hace inefficientes los modelos*

OOV:

No pueden manejar palabras fuera del vocabulario (OOV).

Text Representation

Feature Engineering

- 1.**
Space Vector
Models
- 2.**
Basic Vectorization
Approaches

- 3.**
Word Embeddings

- 4.**
Visualizing
Embeddings***

***opcional

Text Representation

Feature Engineering

Text Representation

Feature Engineering