Pontificia Universidad JAVERIANA Colombia

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 - Pipeline

Contenido:

1. Adquisición de Datos
2. Limpieza de Texto
3. Pre-procesamiento de Texto
4. Feature Engineering
5. Model
6. Evaluation

Pontificia Universidad
JAVERIANA
Colombia

Pipeline NLP

NLP pipeline

Componentes en la construcción de un modelo de NLP

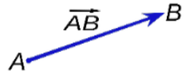
Procesamiento de texto.

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

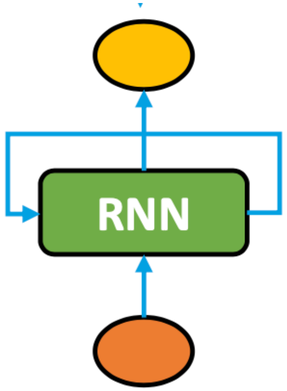
BOW



Word2Vec

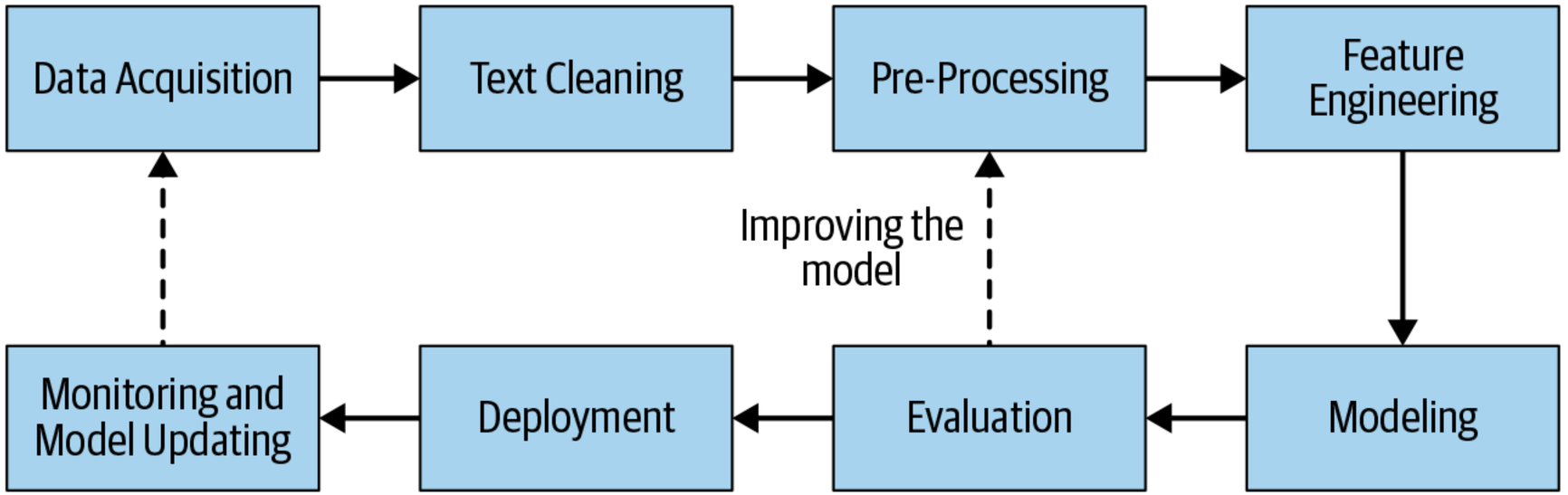
Modelamiento

- **Traducción**
- **Generación de texto**
- **Clasificación de texto**



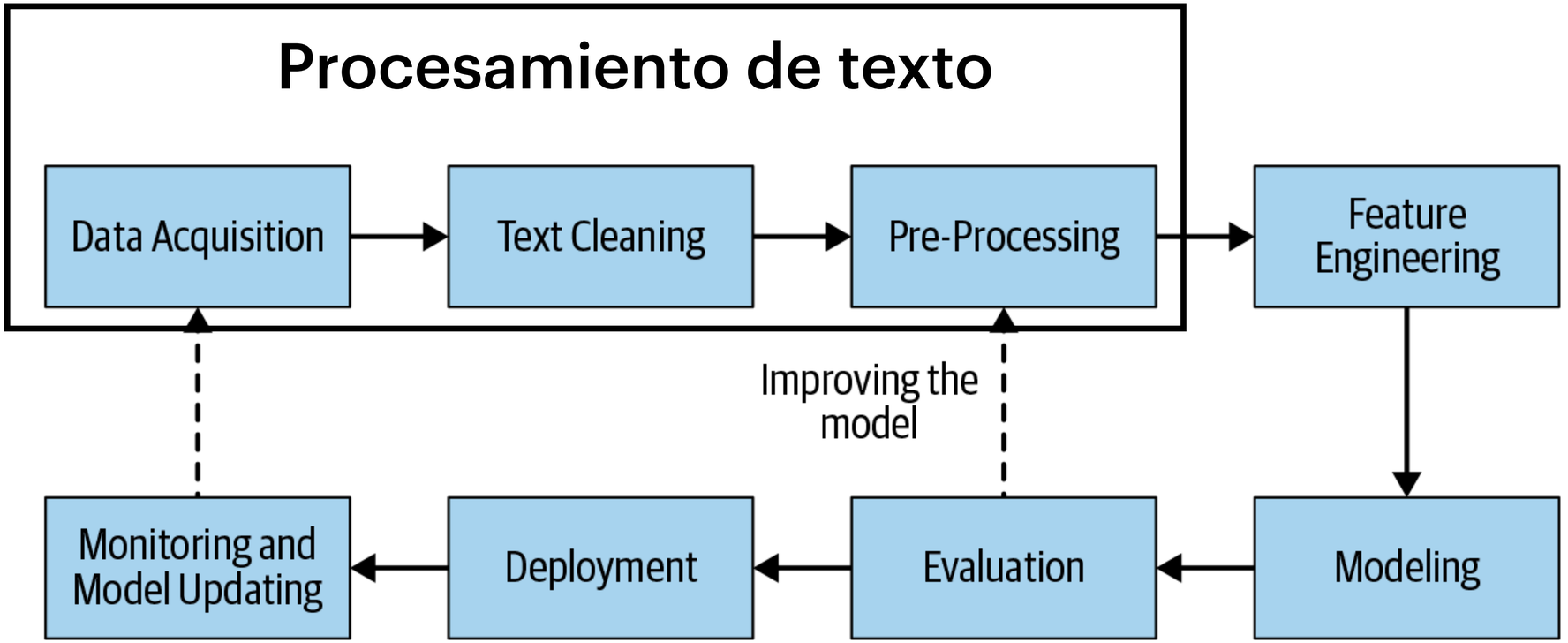
NLP pipeline

Componentes en la construcción de un modelo de NLP



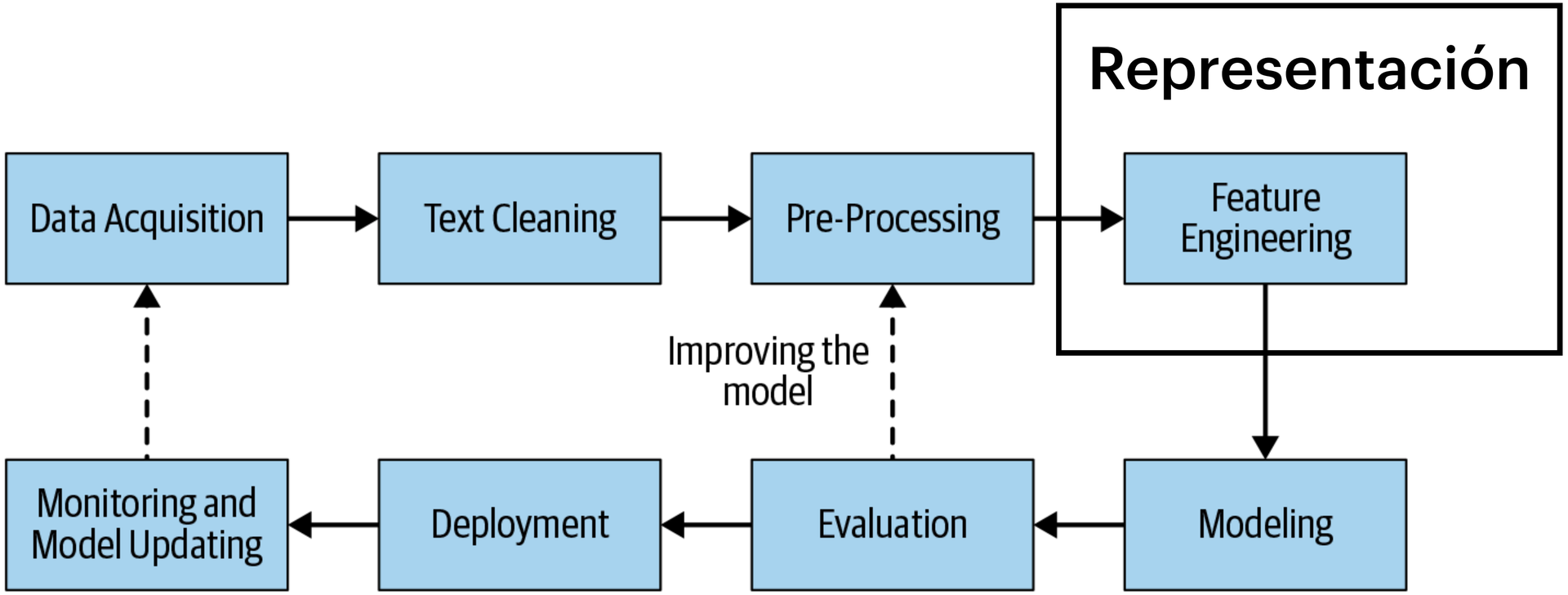
NLP pipeline

Componentes en la construcción de un modelo de NLP



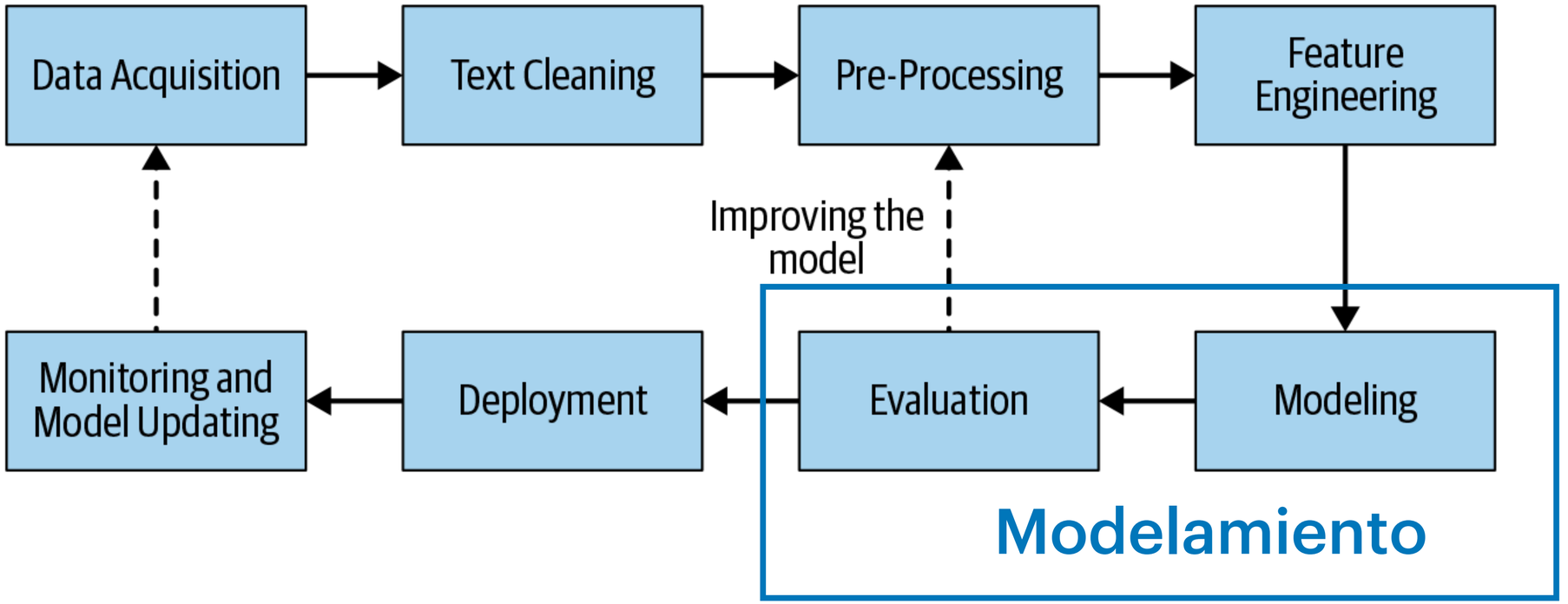
NLP pipeline

Componentes en la construcción de un modelo de NLP



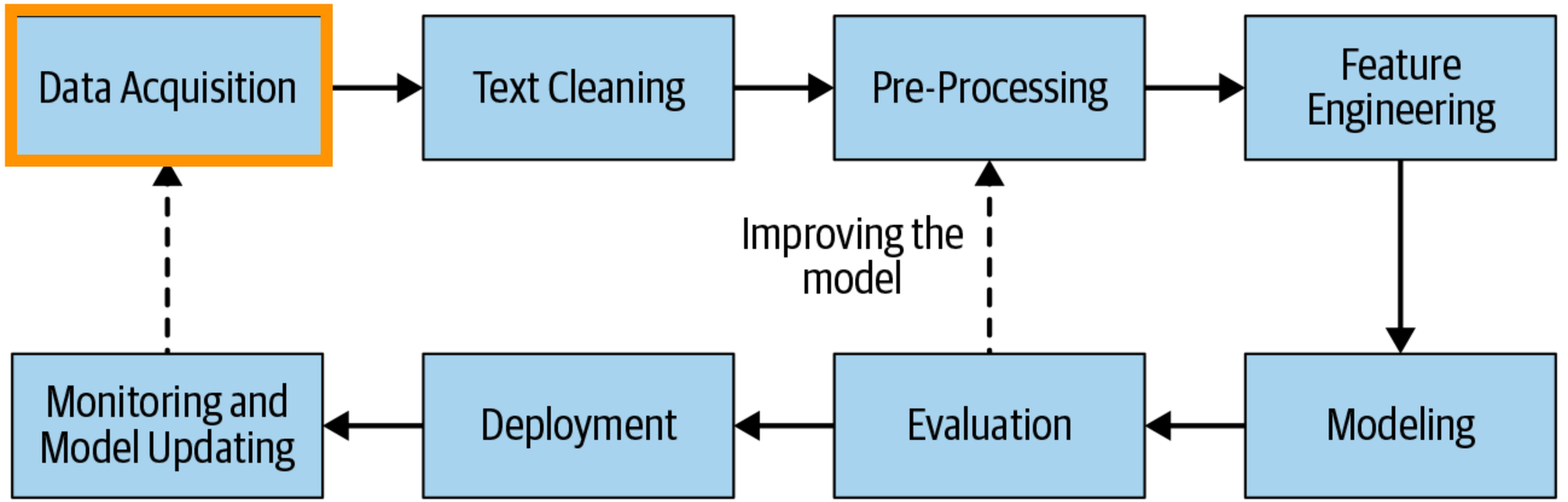
NLP pipeline

Componentes en la construcción de un modelo de NLP



NLP pipeline

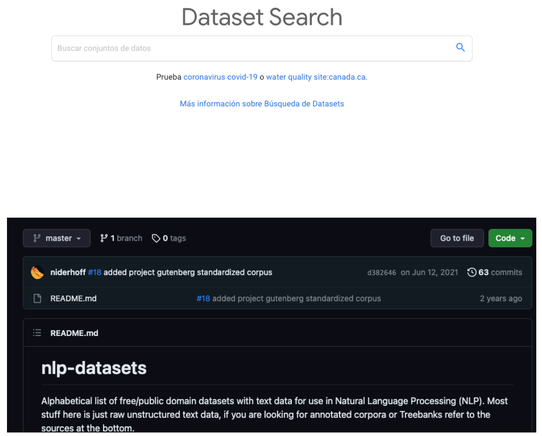
Componentes en la construcción de un modelo de NLP



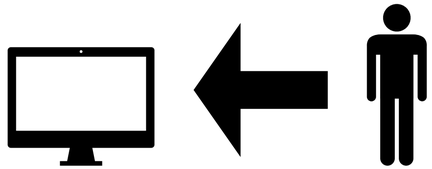
Adquisición de Datos

Componentes en la construcción de un modelo de NLP

Conjunto de datos público



‘Raspadatos’



Etiquetados de datos manual

Intervenir el Producto

Recopilación de datos a base del producto.

Las primeras implementaciones son demoradas.

Google. “Dataset Search”. Last accessed June 15, 2020.

Iderhoff, Nicolas. nlp-datasets: Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP), (GitHub repo). Last accessed June 15, 2020.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

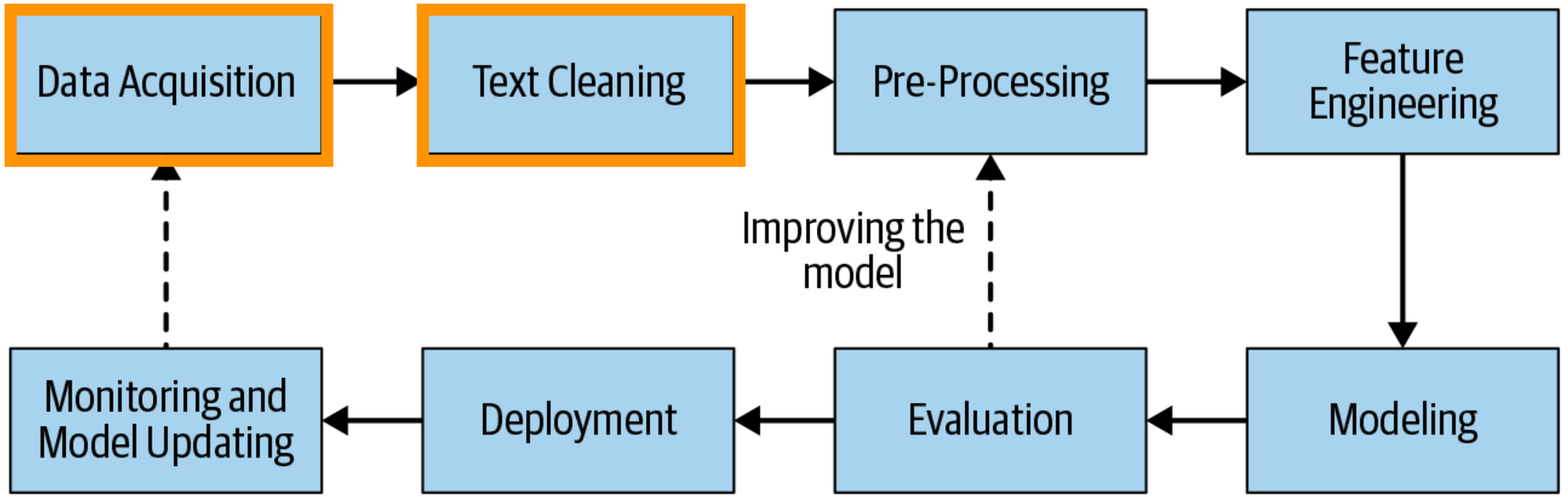
Aumentado de datos

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

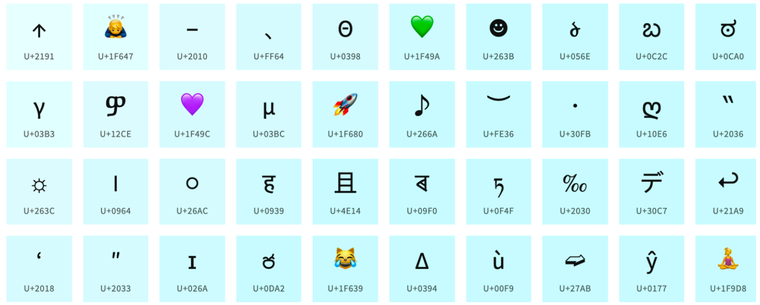


Limpieza de Texto

Componentes en la construcción de un modelo de NLP

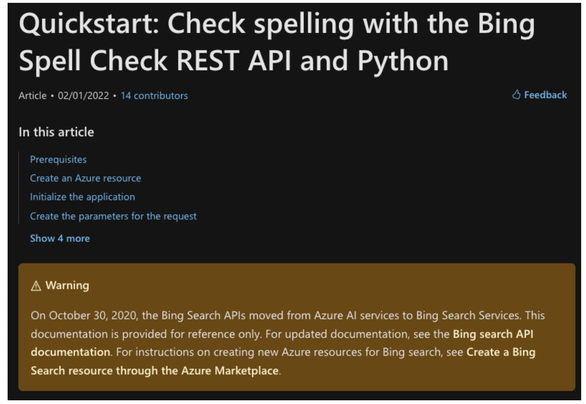
Análisis y limpieza de HTML

Normalización Unicode




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

Corrección ortográfica

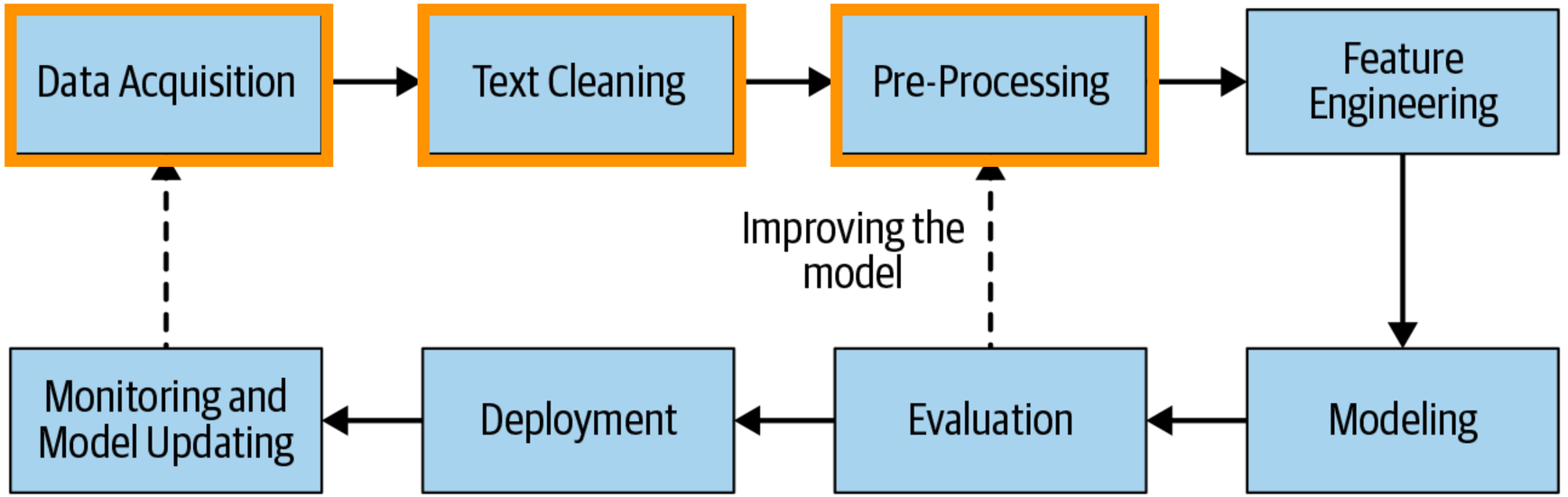


Microsoft documentation. "Quickstart: Check spelling with the Bing Spell Check REST API and Python". Last accessed June 15, 2020.

Dickinson, Markus, Chris Brew, and Detmar Meurers. *Language and Computers*. New Jersey: John Wiley & Sons, 2012. ISBN: 978-1-405-18305-5

NLP pipeline

Componentes en la construcción de un modelo de NLP

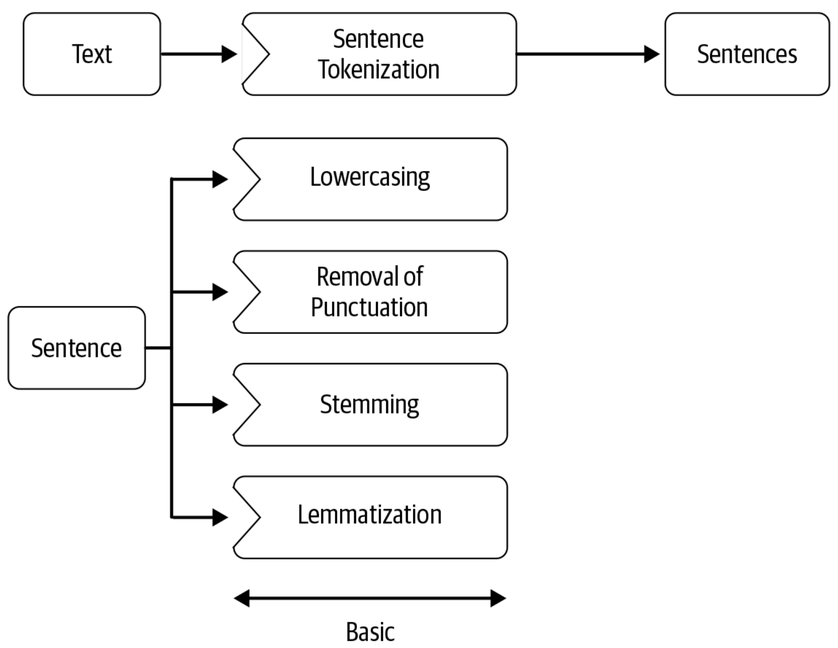


Pre-procesamiento

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.

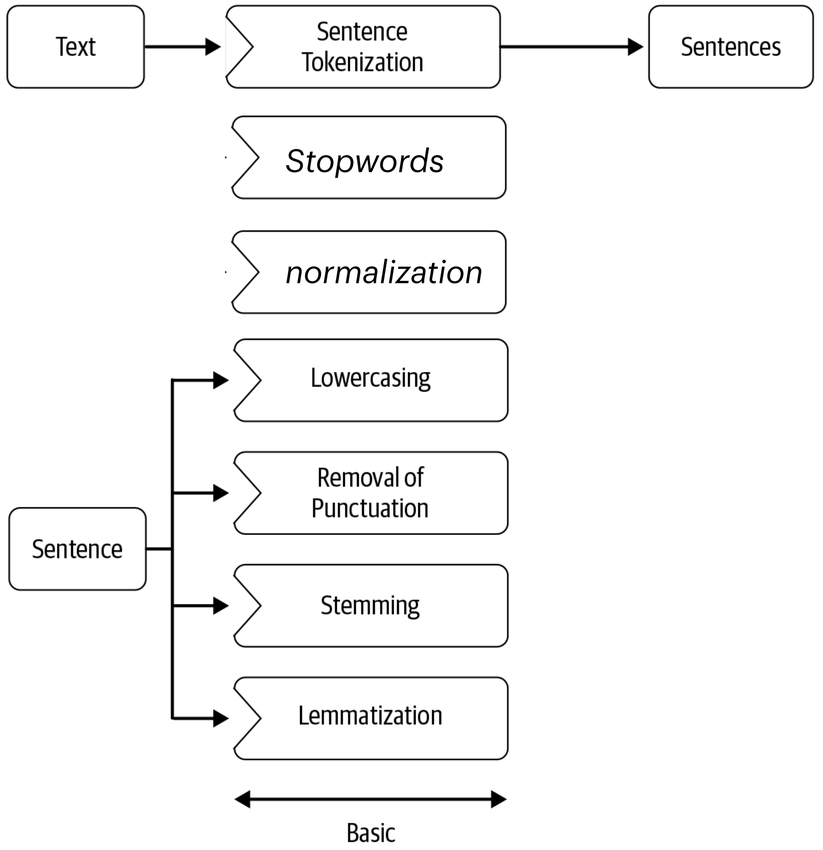


Pre-procesamiento

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: *Los amigos de Diana.*

Salida: ***Tokens*** *[Los, amigos, de, diana]*

Def. Token: *Instancia de secuencias de caracteres.*

Token es ahora un candidato para un indice...
Pero ¿cuáles se consideran tokens válidos?

Porter, Martin F. “An Algorithm For Suffix Stripping.” *Program: electronic library and information systems* 14.3 (1980): 130–137.

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

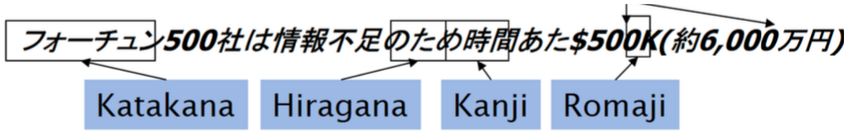
Unidad mínima para procesamiento.

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

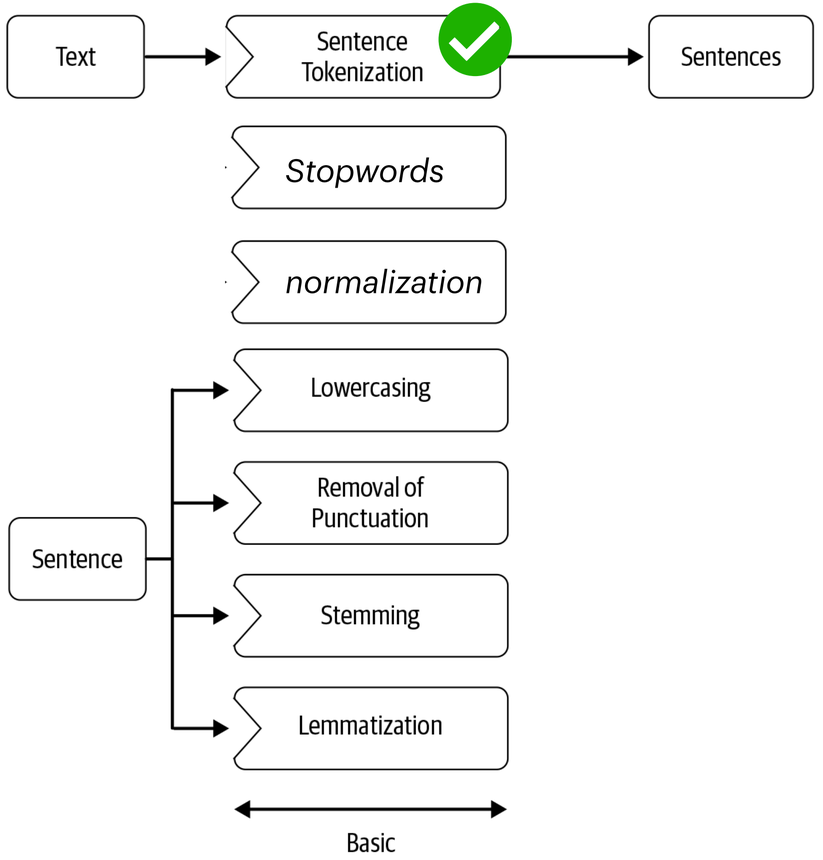
Números

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

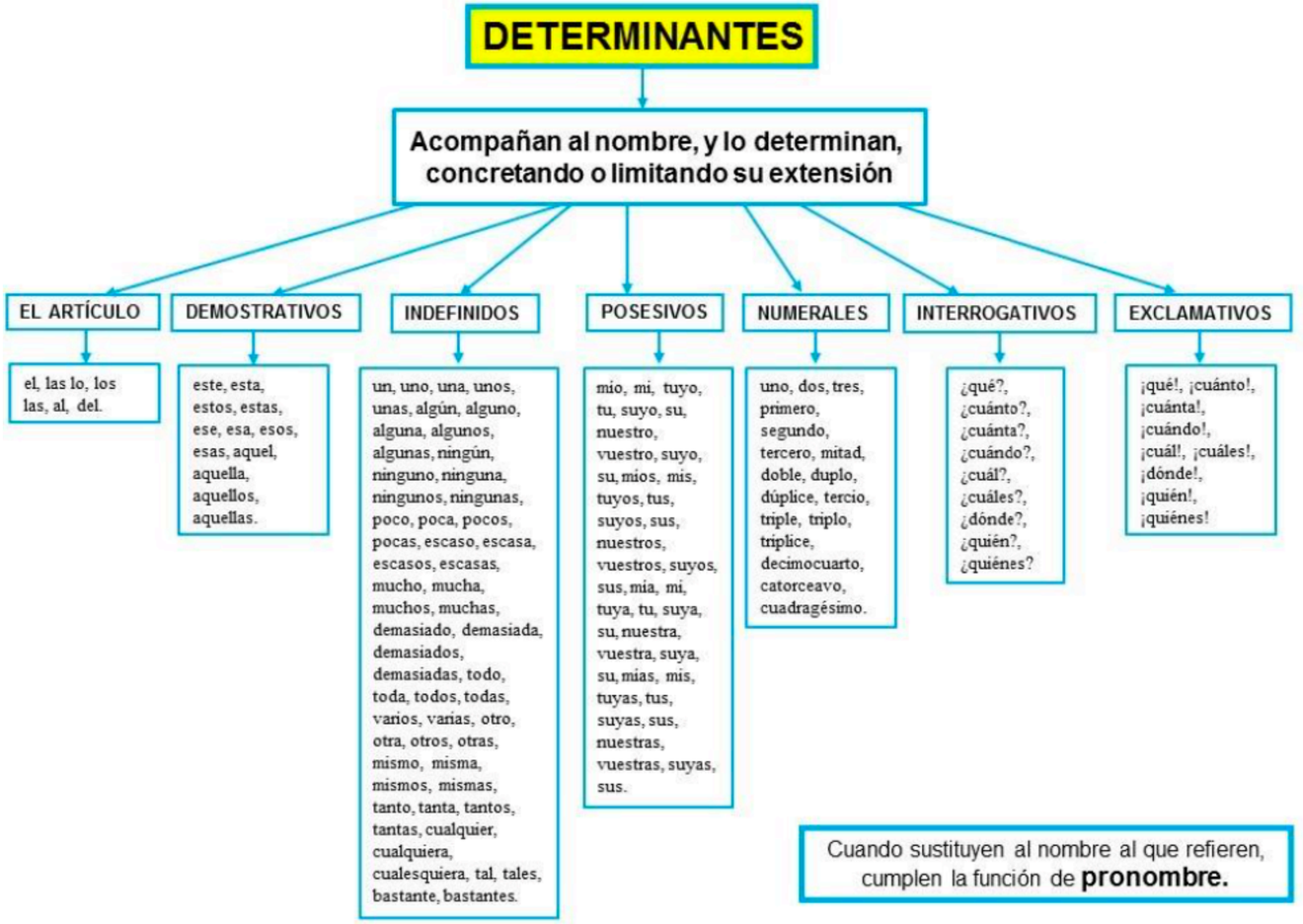
Pre-procesamiento

Componentes en la construcción de un modelo de NLP

HARVARD & SM

Stopwords

Palabras con poco contenido semántico.

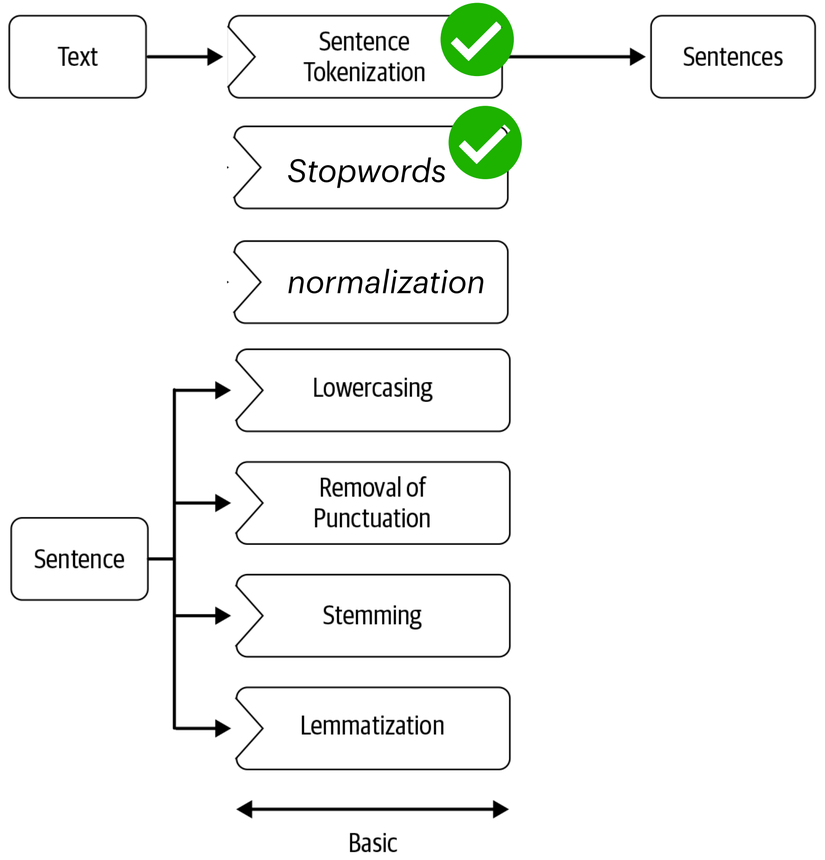


Pre-procesamiento

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*

Existen muchos de ellos en las colecciones

No sirve como criterio diferenciado de documentos.

Francés, español, currículum / curriculum

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

Salida: *término*

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

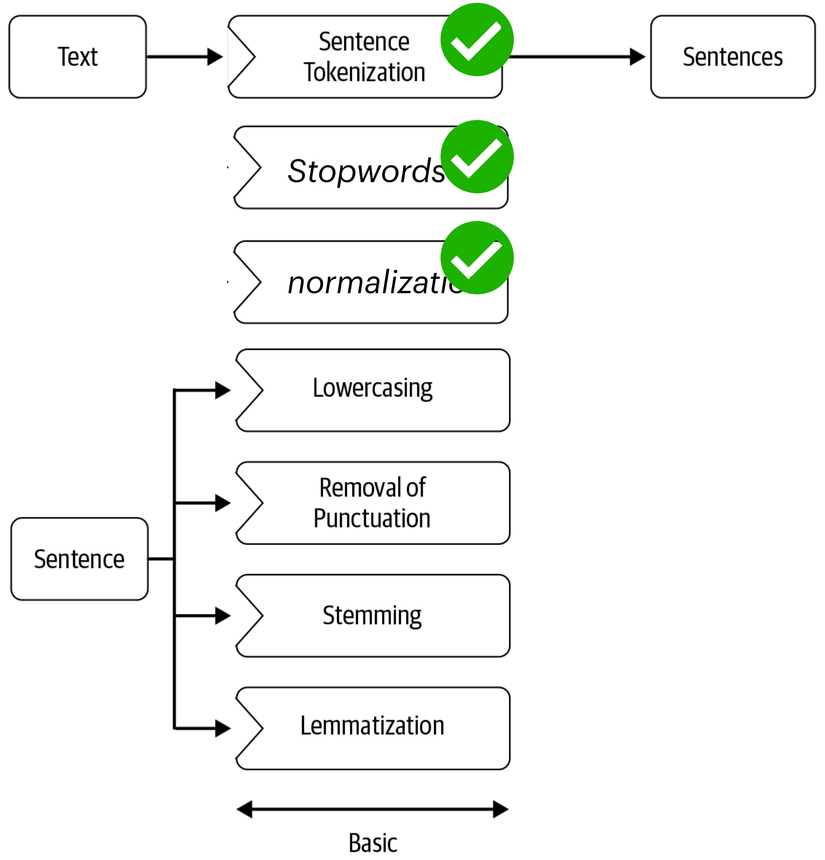
Normalización y Tokenización son dependientes del idioma y pueden estar entrelazadas con la detección del idioma.

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".

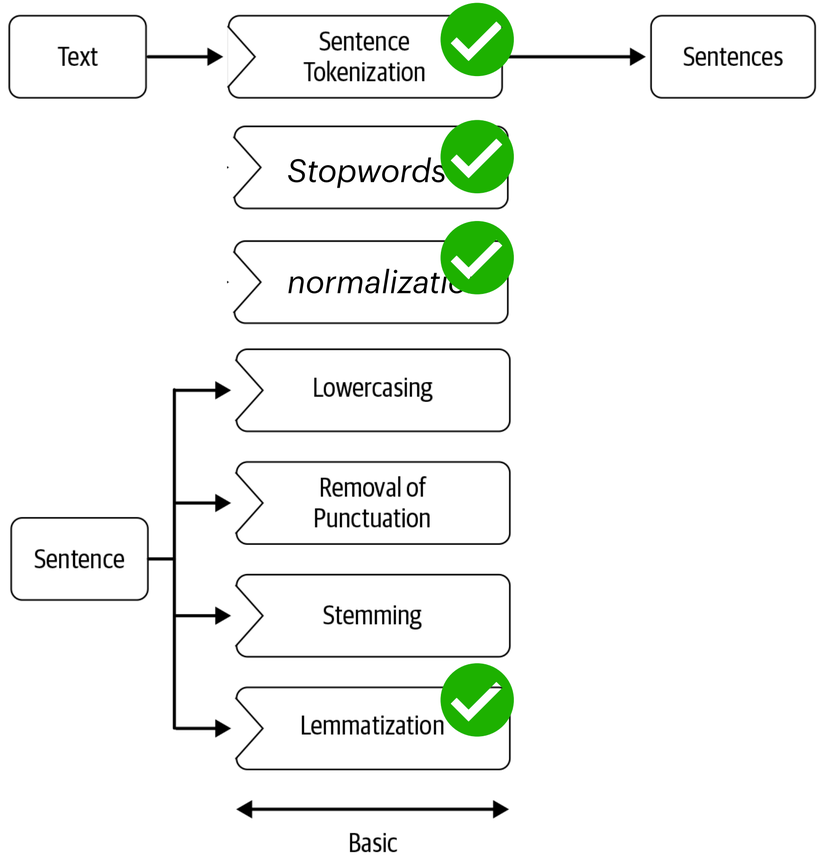
Requiere diccionarios con la morfología de las palabras.

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

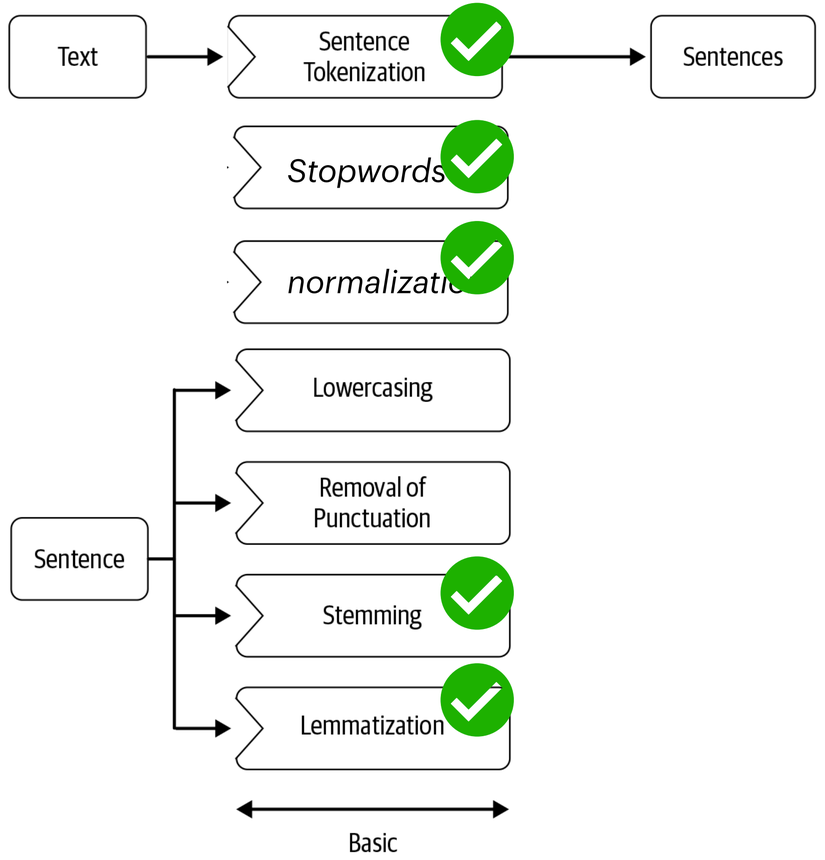
Se selecciona la regla con el sufijo más largo.

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.

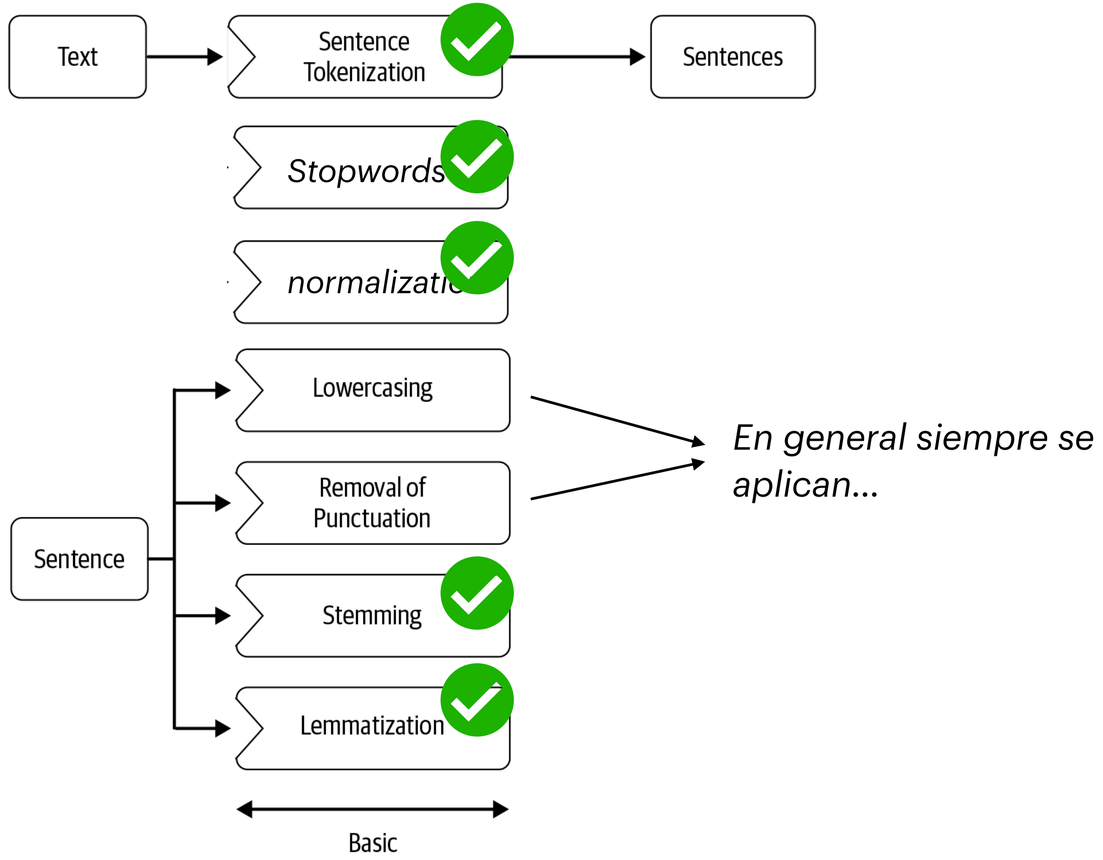


Pre-procesamiento

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.

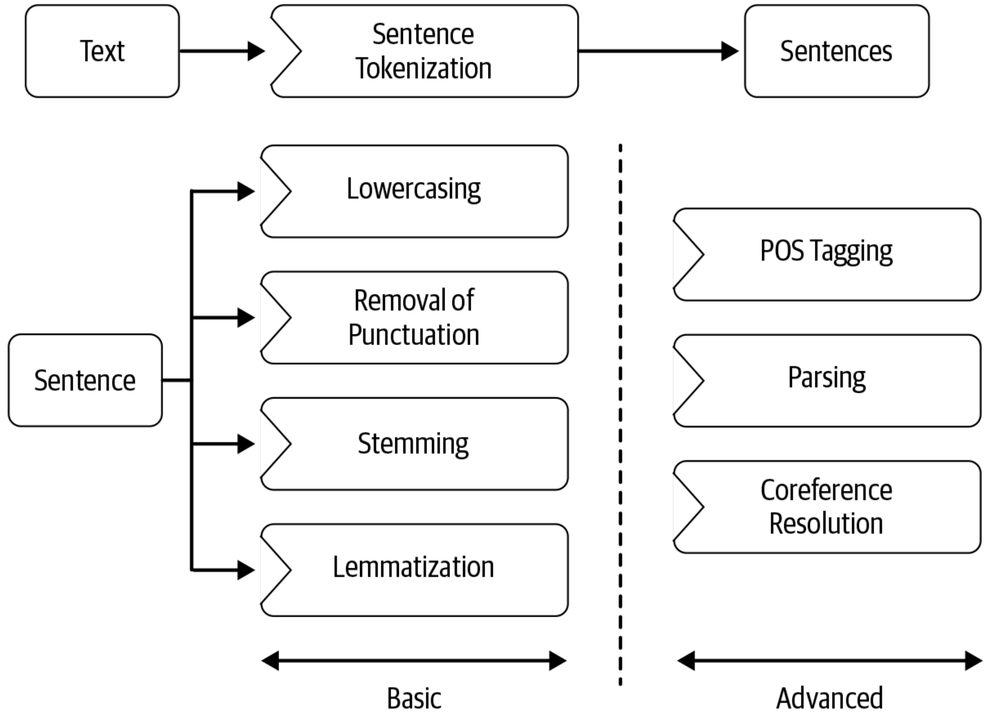


Pre-procesamiento

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.

CoreNLP

Input

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

Tokenization with Lemmatization

POS Tagging

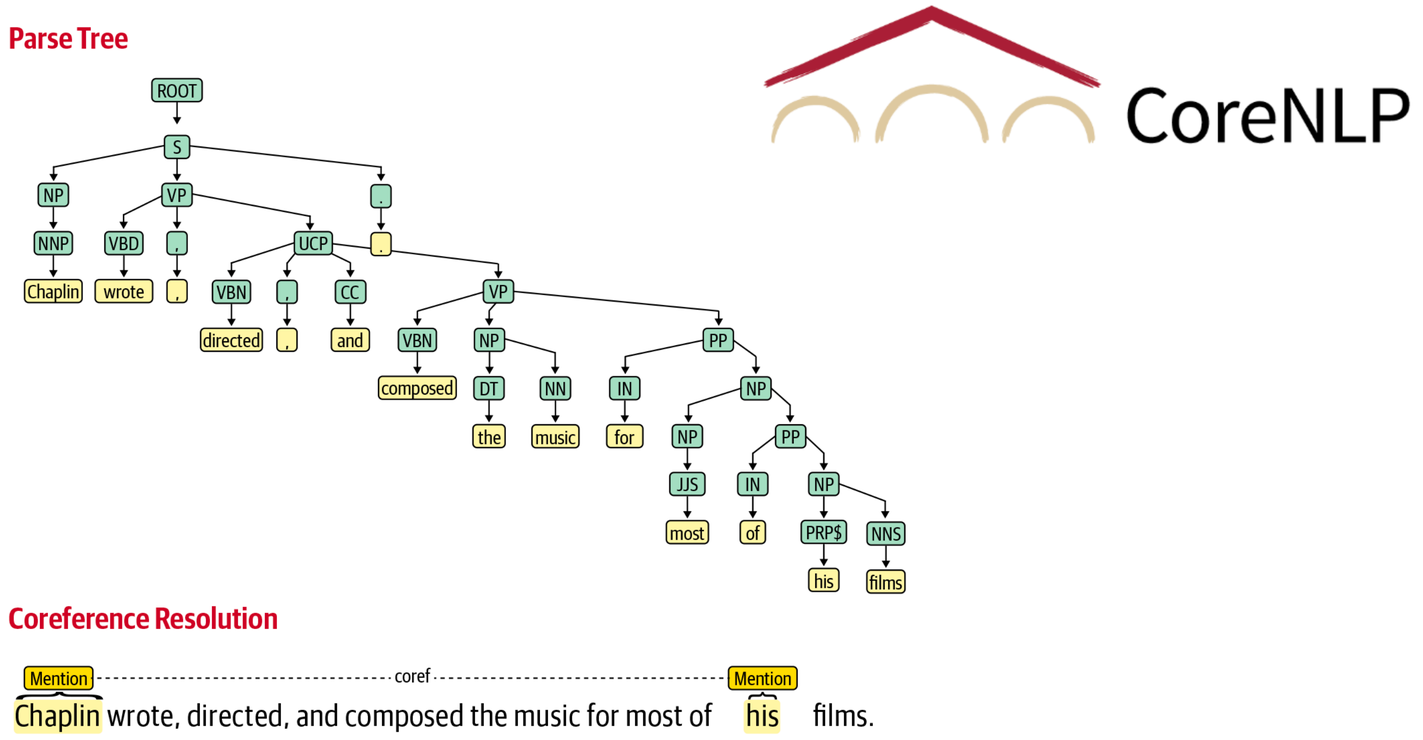
Stanford CoreNLP.Stanford CoreNLP – Natural language software. Last accessed June 15, 2020.

Pre-procesamiento

Componentes en la construcción de un modelo de NLP

Pre-procesamiento de texto

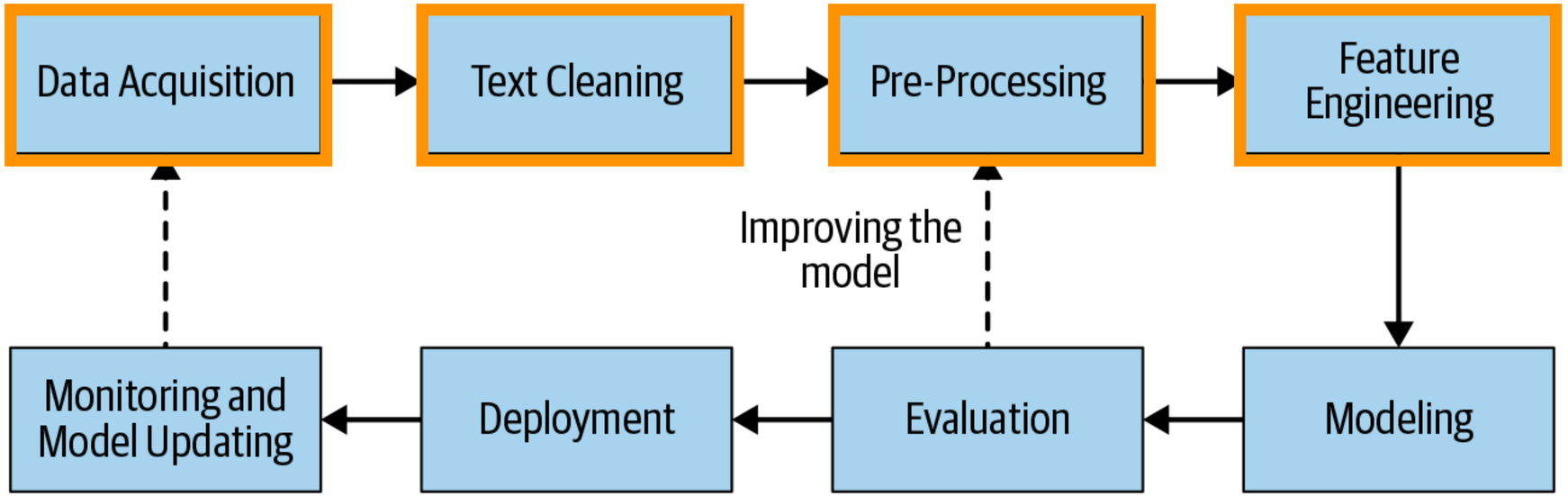
Después de la adquisición y limpieza de texto nos dejó texto sin formato. Sin embargo debemos procesar el texto antes de ser ingerido por el modelo.



Stanford CoreNLP.Stanford CoreNLP – Natural language software. Last accessed June 15, 2020.

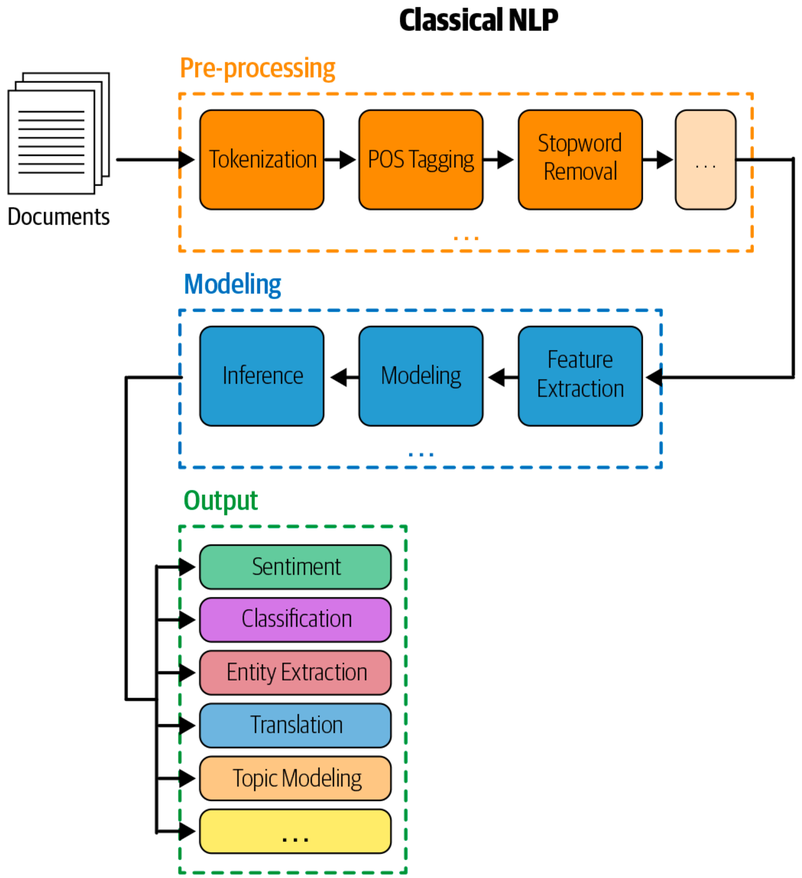
NLP pipeline

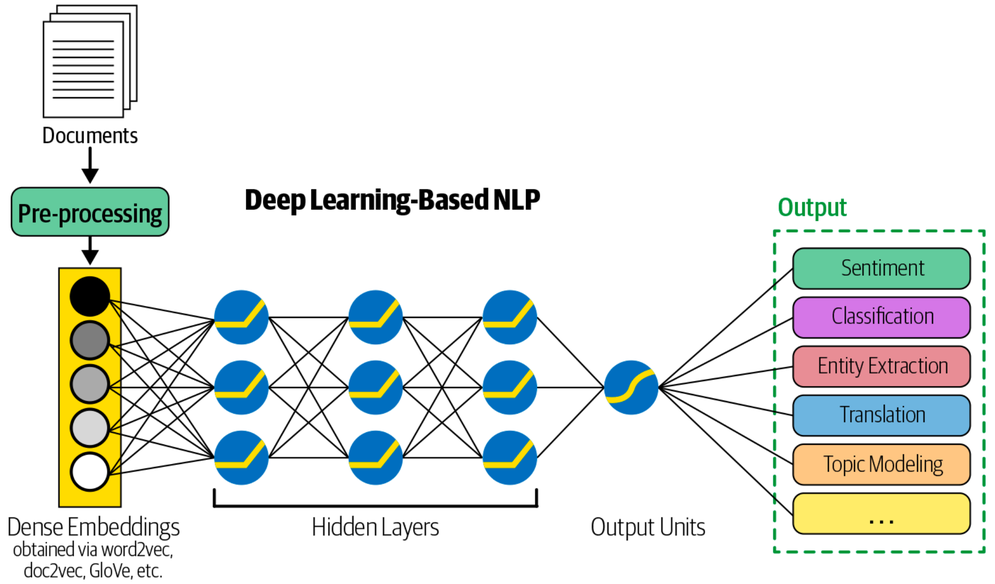
Componentes en la construcción de un modelo de NLP



Ingeniería de Características

Comparación de enfoque

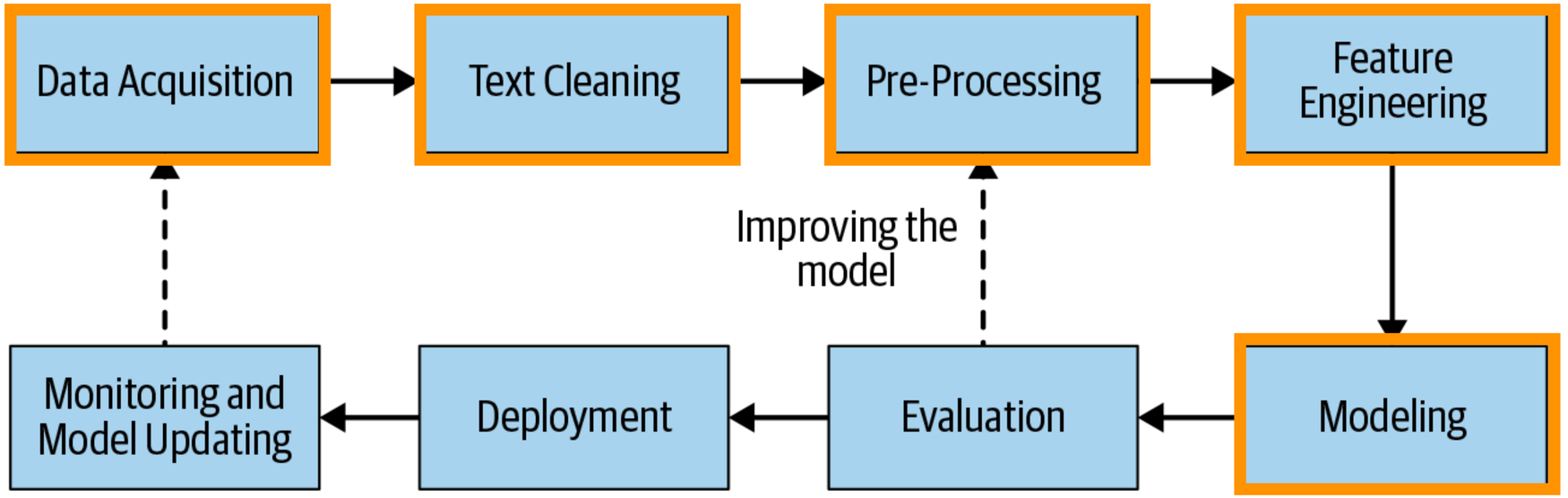




Ghaffari, Parsa. "Leveraging Deep Learning for Multilingual Sentiment Analysis". July 14, 2016.

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 paginas con dominios sospechosos / etc...

Revisa proveedores de NLP

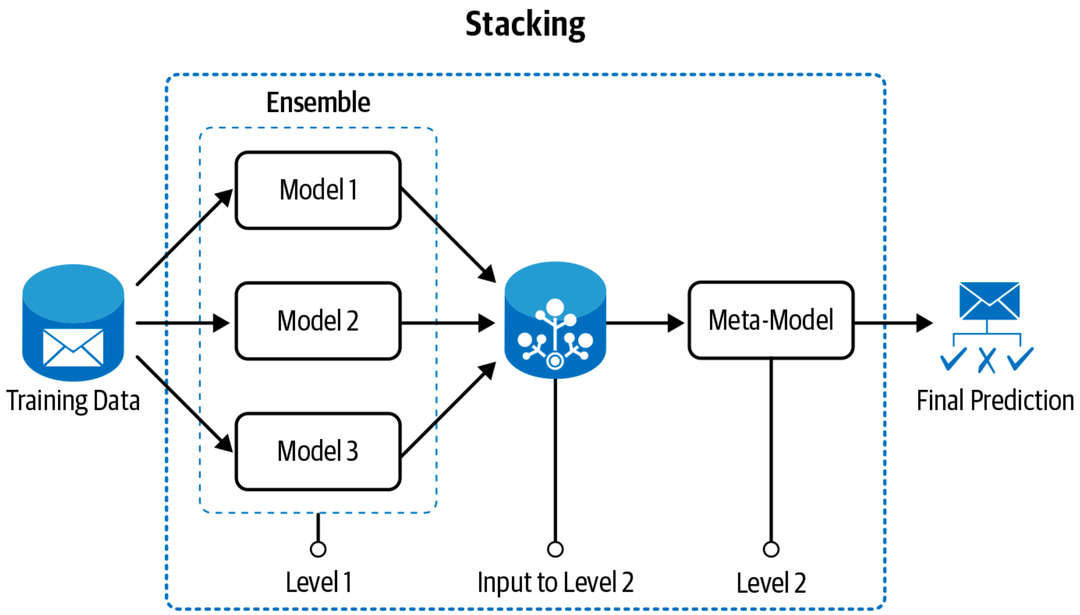
Sí es posible revisa las diferentes APIs que están listas para usar desde Microsoft, Google, IBM...

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

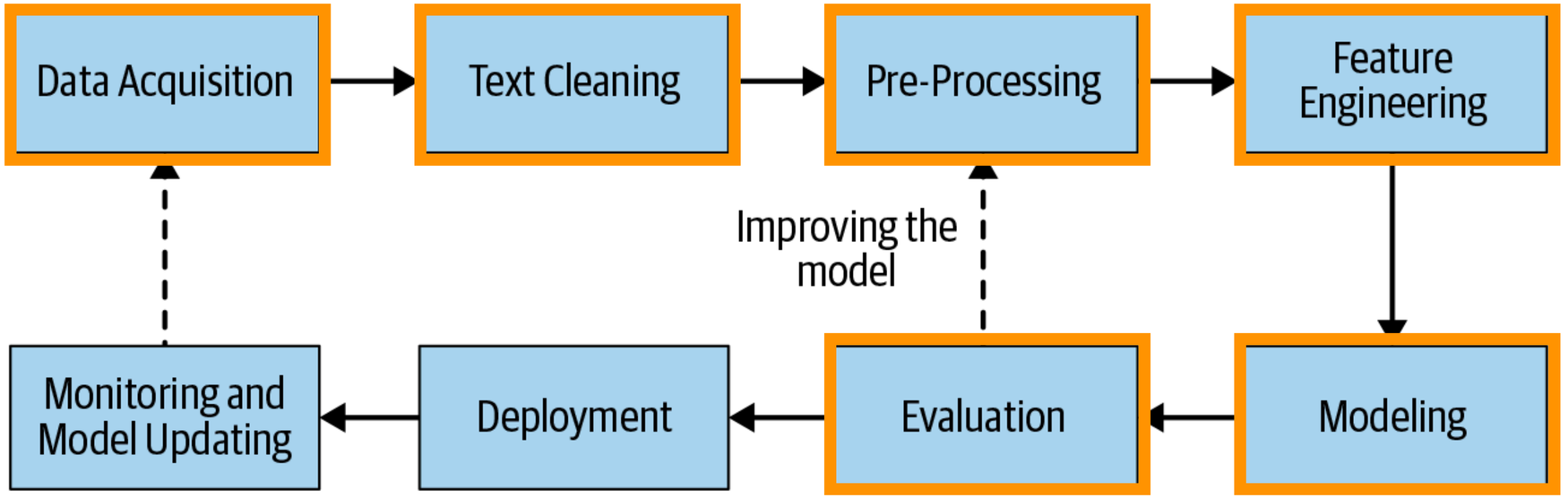
A menudo los modelos requieren de contexto externo, más allá del conjunto de datos para que el modelo comprenda bien el lenguaje y el problema

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 Extrínseca

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 semantica latente

Cosine Similarity

Distancia entre dos vectores