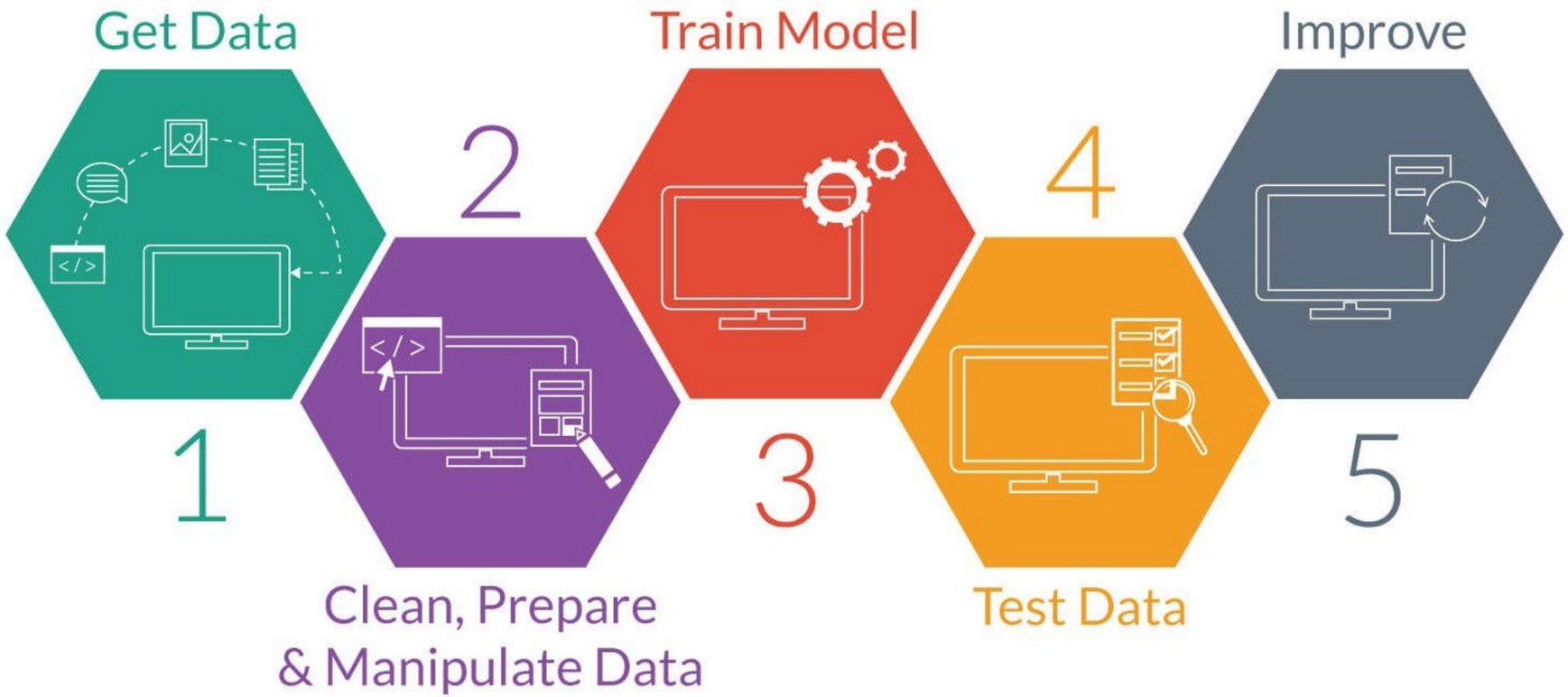
Introducción al Procesamiento del Lenguaje Natural

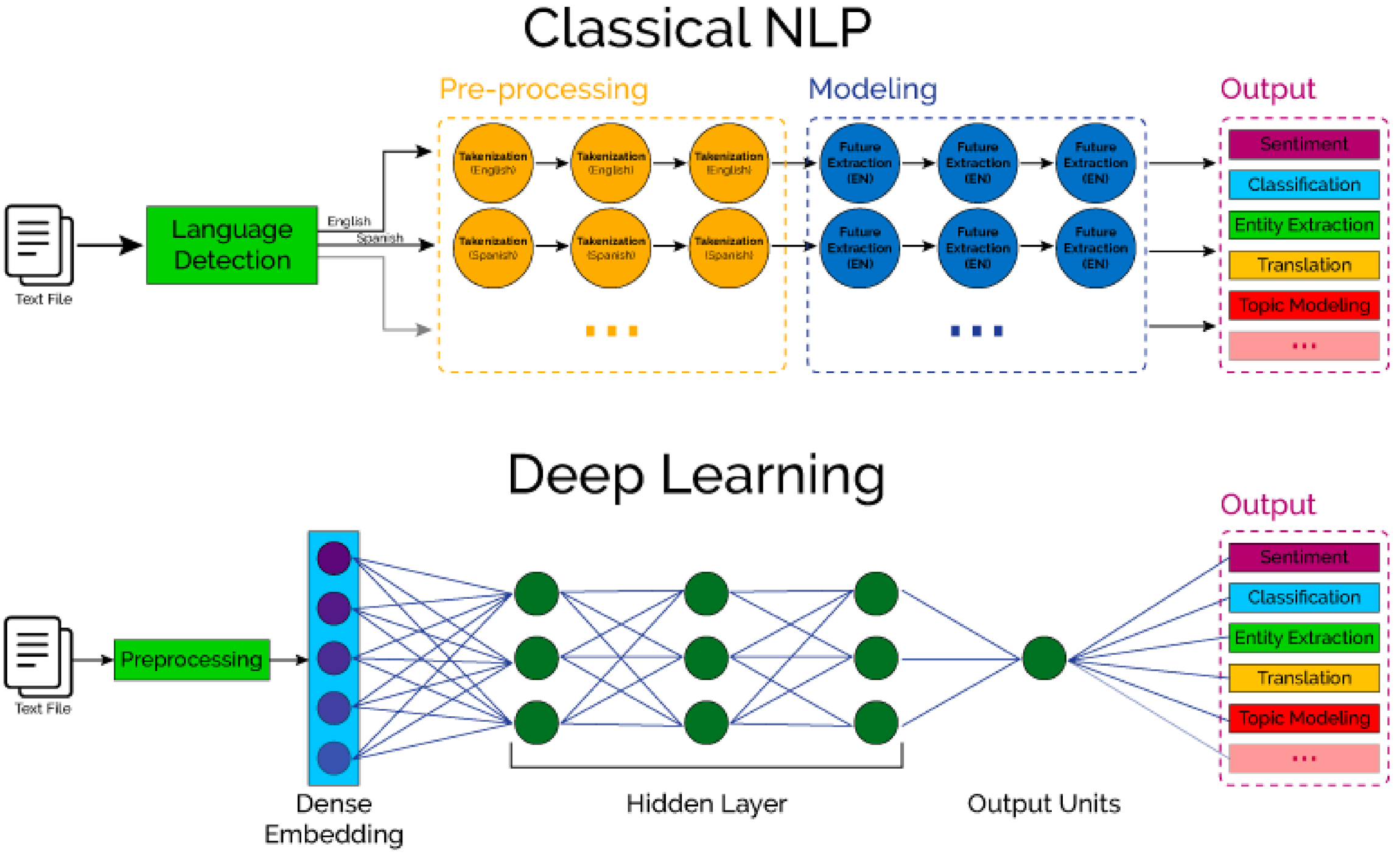
¿Qué es PLN?

Rama de la inteligencia artificial que que combina lingüística y las ciencias de la computación para analizar el lenguaje y desarrollar tecnologías alrededor de este.

A través de diversos algoritmos de **aprendizaje de máquina** (*machine learning*) se desarrollan programas que puedan hacer inferencias sobre corpus desconocidos, a partir de corpus ya analizados y procesados.

Aprendizaje de máquina



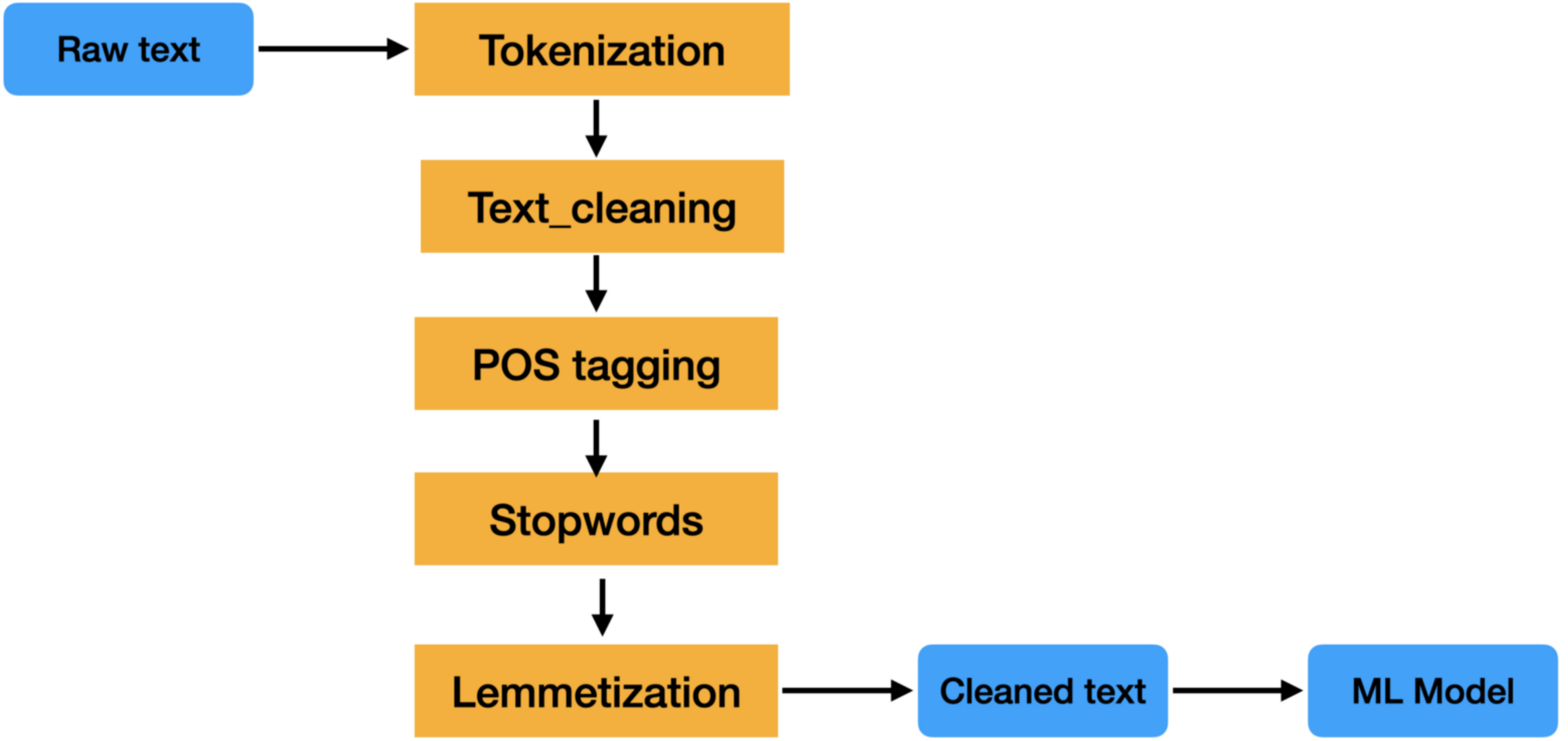

Classical NLP
Pre-processing
Modeling
Output
Text File
Language Detection
English
Spanish
Tokenization (English)
Tokenization (Spanish)
Future Extraction (EN)
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling
Deep Learning
Output
Text File
Preprocessing
Dense Embedding
Hidden Layer
Output Units
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling

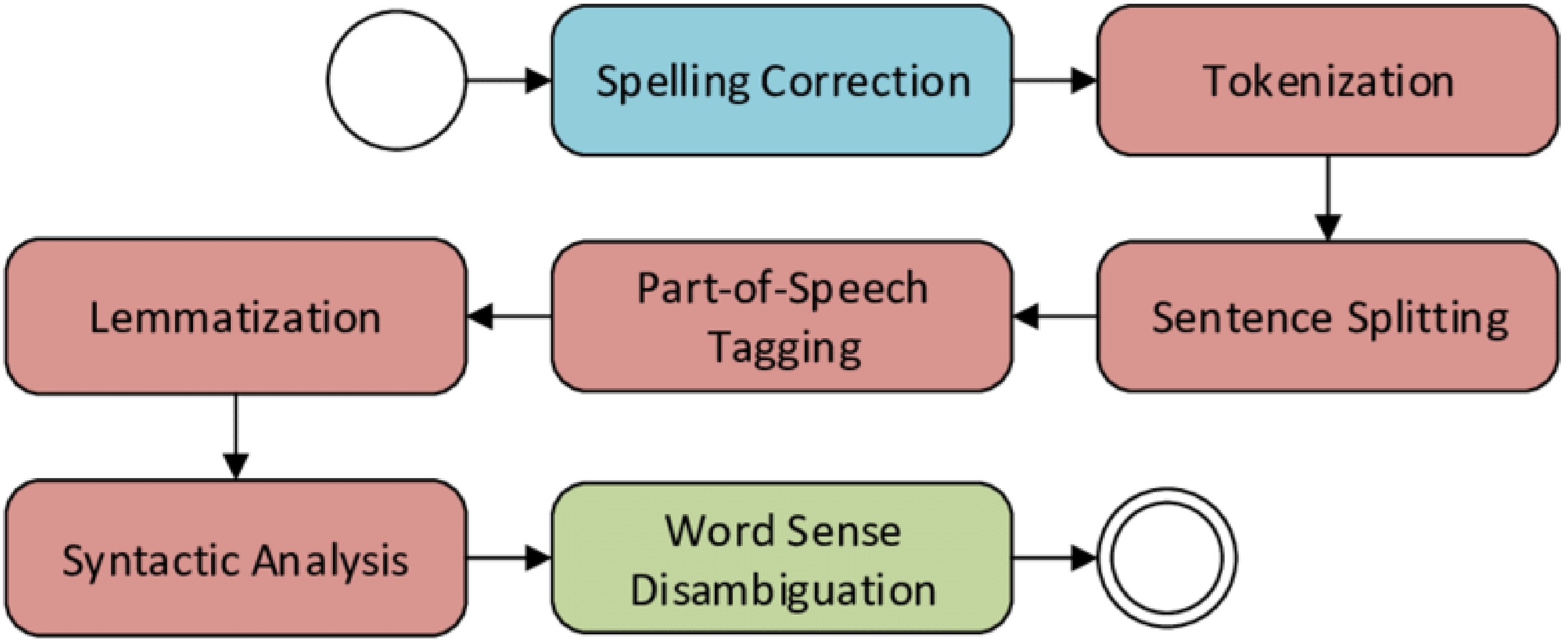
Objetivo

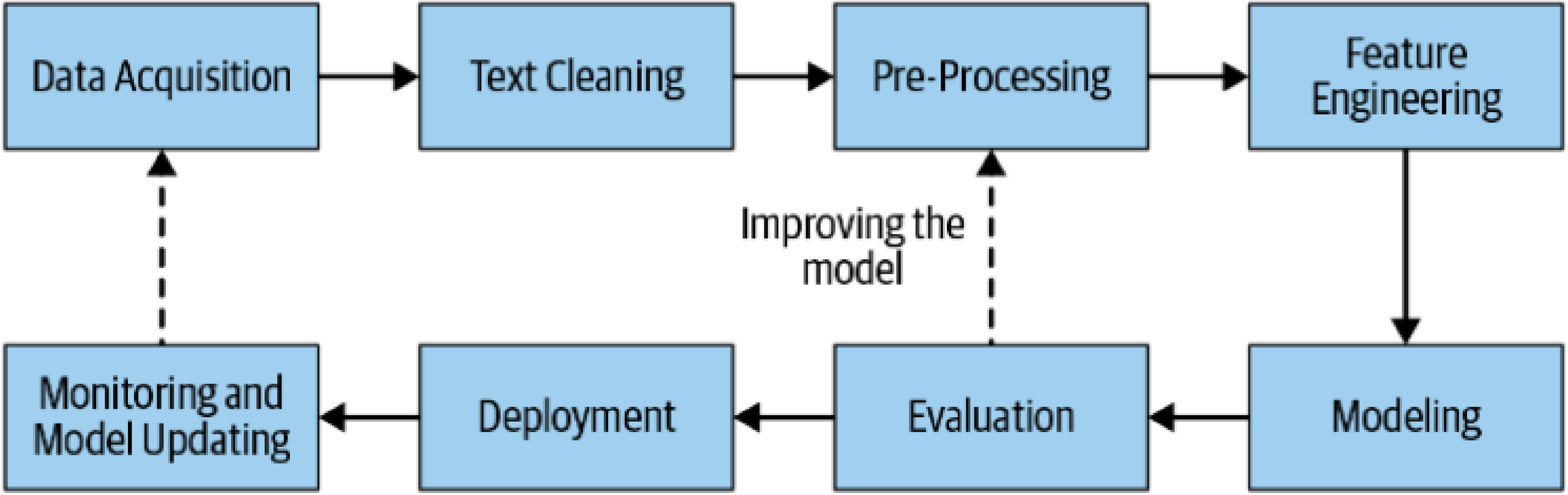
El objetivo último del PLN es lograr que las computadoras utilicen el lenguaje natural tan efectivamente como los humanos.

Principales herramientas del PLN

1. Expresiones regulares: Cadenas de caracteres que expresan un patrón de búsqueda en un texto.
2. Normalización de textos: Convertir un texto a una forma más conveniente para el procesamiento de los textos.
3. Tokenización: Separar palabras en tokens para su procesamiento.
4. Lematización: Determinar si dos palabras tienen (o no) la misma raíz.
5. Stemming: Elimiar los sufijos al final de las palabras.
6. Segmentación de oraciones: Separar oraciones individuales utilizando puntuación.
7. Distancia de edición: Medir qué tan similares son dos cadenas de caracteres basados en el número de cambios que hay que realizar a uno para generar el otro.


Raw text
Tokenization
Text_cleaning
POS tagging
Stopwords
Lemmetization
Cleaned text
ML Model


Spelling Correction
Tokenization
Sentence Splitting
Part-of-Speech Tagging
Lemmatization
Syntactic Analysis
Word Sense Disambiguation


Data Acquisition
Text Cleaning
Pre-Processing
Feature Engineering
Improving the model
Monitoring and Model Updating
Deployment
Evaluation
Modeling

Expresiones Regulares

Especifican patrones de búsqueda específicos en un texto o un corpus.
Regresan todas las instancias que cumplen con el patrón de búsqueda establecido.

RE	Example Patterns Matched
/woodchucks/	"interesting links to woodchucks and lemurs"
/a/	"Mary Ann stopped by Mona's"
/!/	"You've left the burglar behind again!" said Nori

Figure 2.1 Some simple regex searches.

RE	Match	Example Patterns
/[wW]oodchuck/	Woodchuck or woodchuck	"Woodchuck"
/[abc]/	'a', 'b', *or* 'c'	"In uomini, in soldati"
/[1234567890]/	any digit	"plenty of 7 to 5"

Figure 2.2 The use of the brackets [] to specify a disjunction of characters.

https://regexlearn.com/learn

$User_1$:	Men are all alike.
$ELIZA_1$:	IN WHAT WAY
$User_2$:	They're always bugging us about something or other.
$ELIZA_2$:	CAN YOU THINK OF A SPECIFIC EXAMPLE
$User_3$:	Well, my boyfriend made me come here.
$ELIZA_3$:	YOUR BOYFRIEND MADE YOU COME HERE
$User_4$:	He says I'm depressed much of the time.
$ELIZA_4$:	I AM SORRY TO HEAR YOU ARE DEPRESSED

```
s/.* I'M (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/
s/.* I AM (depressed|sad) .*/WHY DO YOU THINK YOU ARE \1/
s/.* all .*/IN WHAT WAY/
s/.* always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/
```

ELIZA (1966)

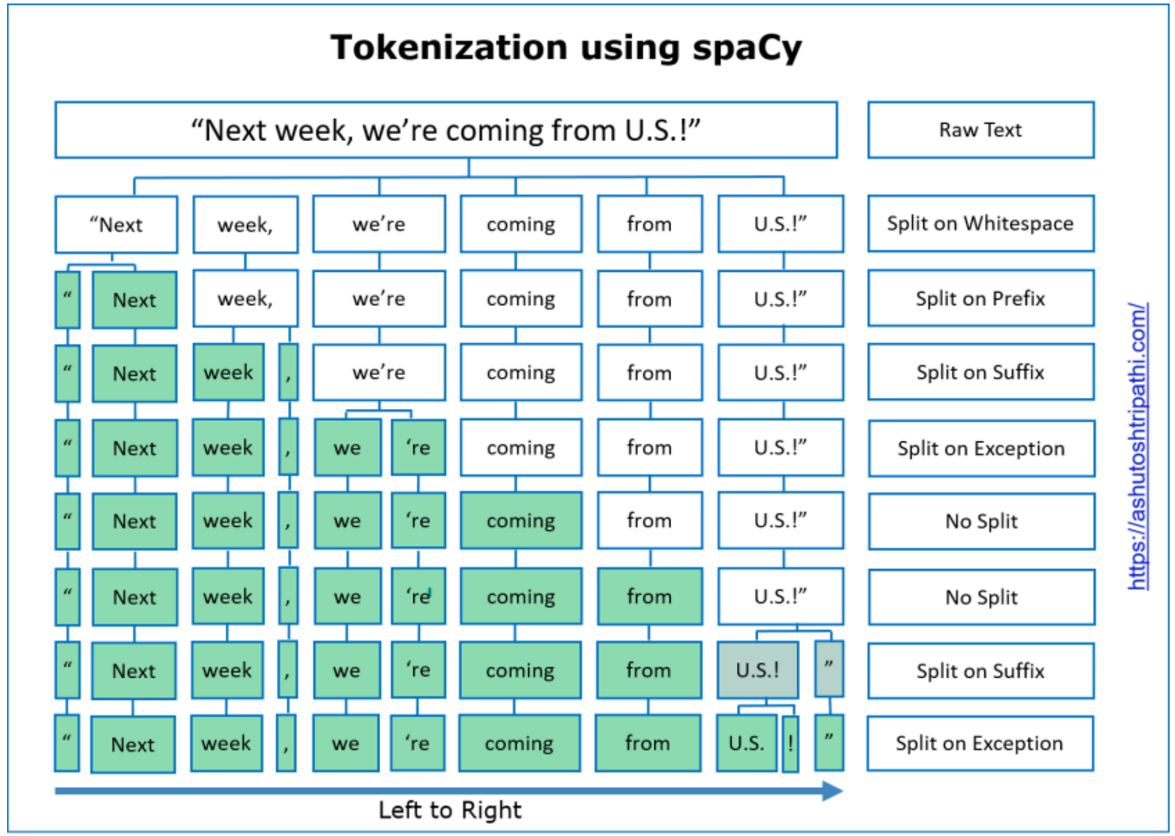
- Funciona gracias a una serie de sustituciones por medio de expresiones regulares, en donde cada una de estas cambia cierta parte del texto de entrada.
- Primero, las entradas son pasadas a mayúsculas.
- Luego, sustituye todas las instancias de MY a YOUR, y I'M a YOU ARE.
- Así se siguen haciendo sustituciones.

Palabras

¿Qué cuenta como una **palabra**?

¿Qué se hace con palabras como *hmm*?

Tokens



Tipos de corpus

- Wikipedia
- Textos religiosos
- Ficción
- Notas periodísticas
- Conversaciones telefónicas
- Cámaras de video
- Transcripts de entrevistas
- Textos legales
- Datos de internet
- etc

Hoja de Datos o Declaración de Datos

Motivation: Why was the corpus collected, by whom, and who funded it?

Situation: When and in what situation was the text written/spoken? For example, was there a task? Was the language originally spoken conversation, edited text, social media communication, monologue vs. dialogue?

Language variety: What language (including dialect/region) was the corpus in?

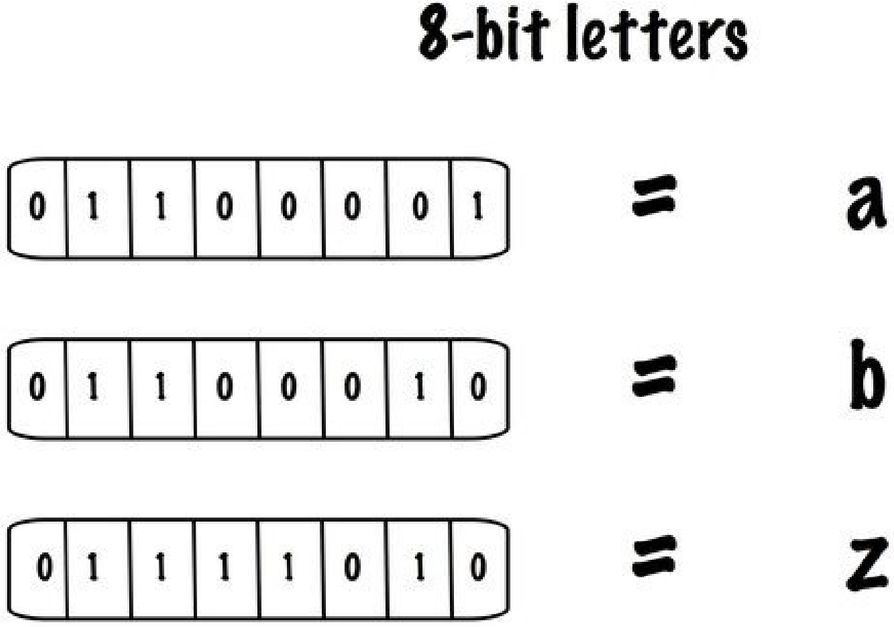
Speaker demographics: What was, e.g., age or gender of the authors of the text?

Collection process: How big is the data? If it is a subsample how was it sampled? Was the data collected with consent? How was the data pre-processed, and what metadata is available?

Annotation process: What are the annotations, what are the demographics of the annotators, how were they trained, how was the data annotated?

Distribution: Are there copyright or other intellectual property restrictions?

¿Cómo es que las computadoras entienden el lenguaje?



https://javaconceptoftheday.com/how-the-strings-are-stored-in-the-memory/

Representaciones de palabras

Para representar palabras se usan **word embeddings.**
Estos consisten en hacer un mapeo de palabras a espacios matemáticos que permiten hacer operaciones aritméticas.

https://towardsdatascience.com/introduction-to-word-embeddings-4cf857b12edc

Entrenamiento de modelos

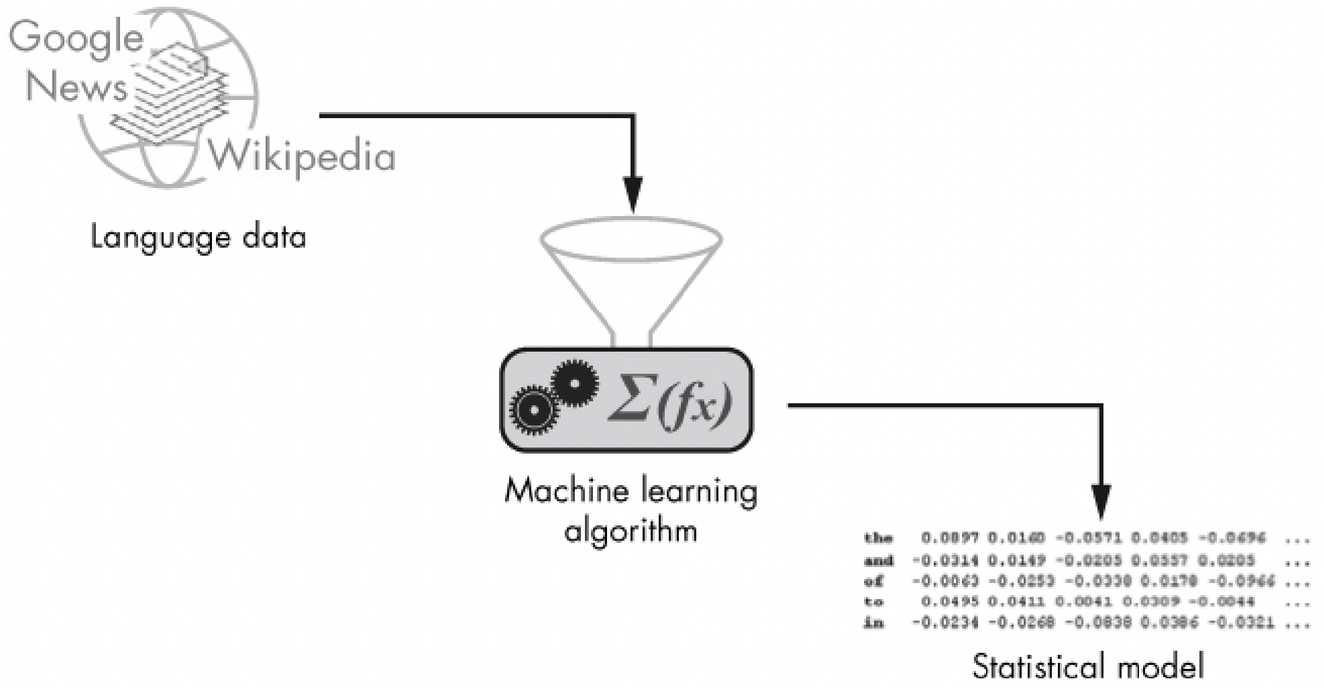


Figure 1-1: Generating a statistical model with a machine learning algorithm using a large volume of text data as input

Haciendo predicciones

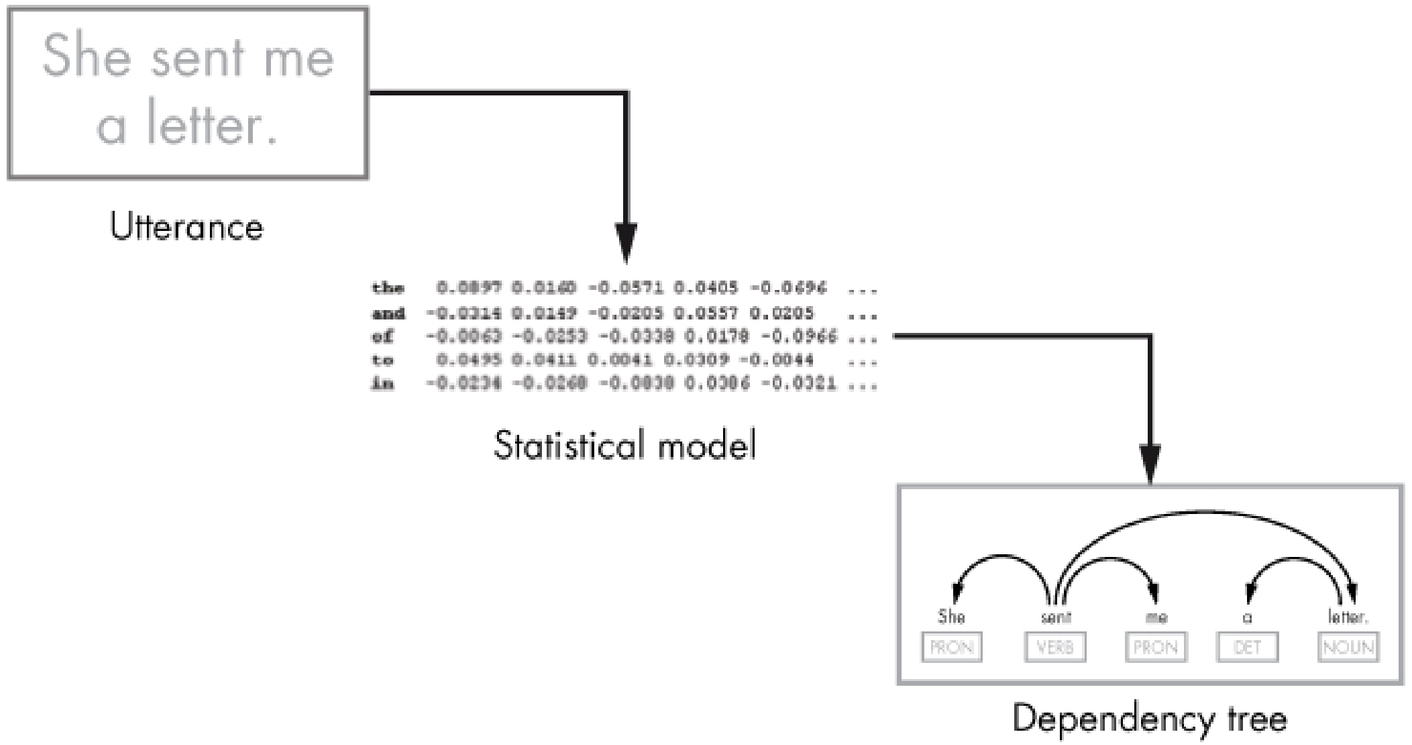


Figure 1-2: Predicting a dependency tree structure for an utterance using a statistical model

¿Por qué usar aprendizaje de máquina para PLN?

- El número de palabras en una lengua es demasiado extenso como para hacer una correspondencia manual entre significado y representación computacional.
- Las palabras adquieren significado de acuerdo a su contexto.
- La complejidad de las lenguas impide crear reglas formales exactas para describirlas.

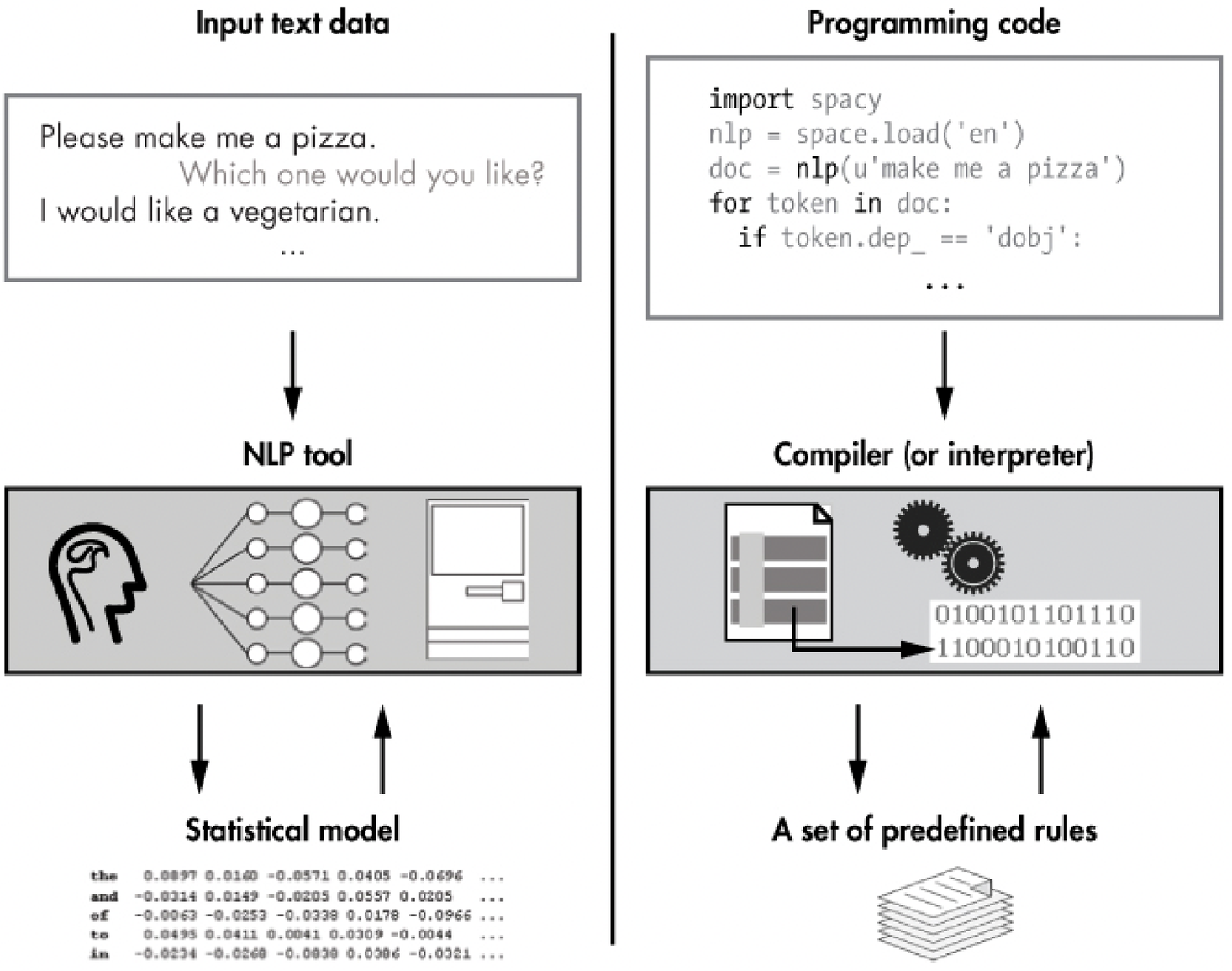
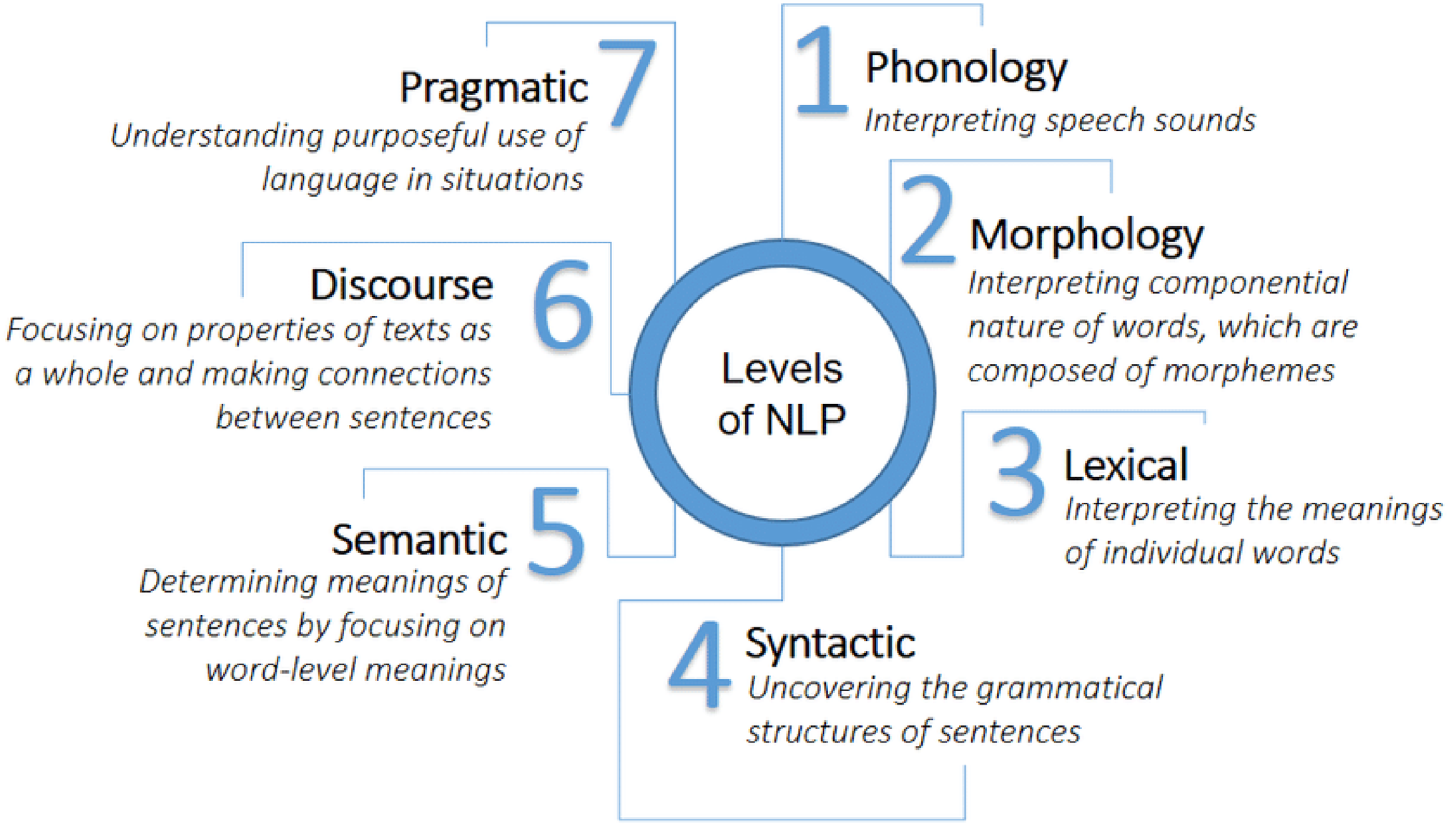
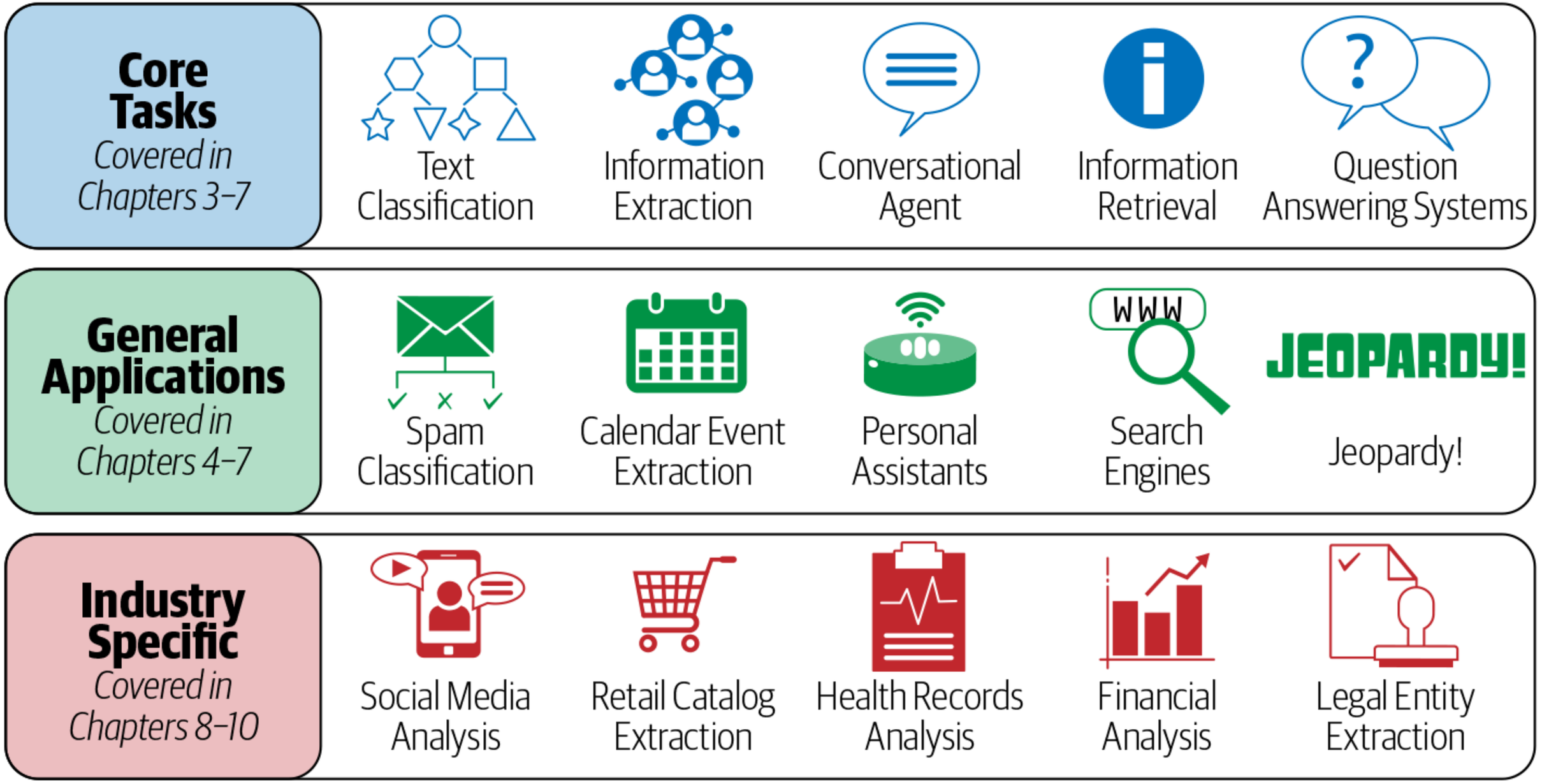


Figure 1-3: On the left, a basic workflow for processing natural language; on the right, a basic workflow for processing a programming language


Levels of NLP
1 Phonology
Interpreting speech sounds
2 Morphology
Interpreting componential nature of words, which are composed of morphemes
3 Lexical
Interpreting the meanings of individual words
4 Syntactic
Uncovering the grammatical structures of sentences
5 Semantic
Determining meanings of sentences by focusing on word-level meanings
6 Discourse
Focusing on properties of texts as a whole and making connections between sentences
7 Pragmatic
Understanding purposeful use of language in situations


Core Tasks
Covered in Chapters 3–7
Text Classification
Information Extraction
Conversational Agent
Information Retrieval
Question Answering Systems
General Applications
Covered in Chapters 4–7
Spam Classification
Calendar Event Extraction
Personal Assistants
WWW
Search Engines
JEOPARDY!
Jeopardy!
Industry Specific
Covered in Chapters 8–10
Social Media Analysis
Retail Catalog Extraction
Health Records Analysis
Financial Analysis
Legal Entity Extraction

https://aclanthology.org/events/emnlp-2021/