Introducing Natural Language Processing

NLP OVERVIEW · CHALLENGES · NLP PIPELINE · TEXT ANALYSIS · TRANSCRIBE · POLLY · TRANSLATE · COMPREHEND · LEX

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

01 **NLP, defined: NLP develops computational algorithms to automatically analyze and represent human language.** It evaluates language's hierarchy — **words** → **phrases** → **sentences** → **ideas** — so ML systems can process large sets of words and phrases. NLP **predates machine learning** (early speech-to-text, screen readers), and its development **maps directly to the ML pipeline**.

02 **Why it is hard:** Human language is **imprecise and context-dependent** — the same words carry different meanings. Every NLP system must tackle four challenges: **discovering the structure** of text, **labeling data** (tagging parts of speech), **representing context**, and **applying grammar** — the challenge where machine learning has the biggest impact.

03 **The pipeline: Preprocess** the text (remove stop words, normalize with stemming and lemmatization, standardize), **tokenize** it into a DataFrame, engineer **features** (bag of words, TF-IDF), then **analyze and extract meaning** (classification, similarity, NER). Libraries such as **NLTK** for Python handle much of this.

04 **Five managed services: Transcribe** (speech → text), **Polly** (text → speech), **Translate** (language → language), **Comprehend** (insights from text — entities, key phrases, language, sentiment, parts of speech), and **Lex** (conversational voice and text bots powered by the same engine as Amazon Alexa).

05 **Direction and composition:** Mind the direction — **Transcribe and Polly are opposites**. The services **compose**: Transcribe can label streaming audio and feed **Comprehend**; Translate integrates with Comprehend, Transcribe, and Polly to build multilingual, spoken, analyzed pipelines; Lex scales on demand with **AWS Lambda**.

02 KEY TERMS

TERM	DEFINITION
Natural language processing (NLP)	Computational algorithms that automatically analyze and represent human language by evaluating its structure
Stop words	Common, low-value words (such as "this" or "is") that are removed from input text before analysis
Normalization	Converting similar word forms (run, runner, ran, running) into one common form, using stemming and lemmatization
Standardization	Removing or replacing words that are not in the analysis dictionary — acronyms, slang, and special characters
Tokenization	Converting words into items in a data collection such as a DataFrame; for example, NLTK's `word_tokenize` function
Bag of words	A vector model that represents a document by the frequency (count) of each word; it ignores word order
TF-IDF	Term frequency–inverse document frequency: weights words by how often they appear in a document versus across many documents; words common to many documents get a lower weight
Named entity recognition (NER)	Extracting entities from text — identify noun phrases, classify them (for example, with Word2Vec), then disambiguate with a knowledge graph
Knowledge graph	Combines subject matter expertise with ML to disambiguate entities and derive meaning; the Amazon recommendations engine is an example
Edit distance	The Levenshtein distance algorithm for measuring text similarity; used for autocorrect, spell check, and grammar check
Coreference resolution	The process used to derive relationships between different words or phrases within a body of text

TERM	DEFINITION
SSML	Speech Synthesis Markup Language — markup that gives Amazon Polly special instructions, such as inserting a pause between two words

03 THE NLP PIPELINE

each stage maps to the ML development process

01 **Formulate** the problem, then collect and label (tag) data → 02 **Preprocess**: remove stop words, normalize, standardize → 03 **Tokenize** the text into a DataFrame → 04 Engineer **features**: bag of words, TF-IDF → 05 **Analyze and extract meaning**: classify, match, NER

04 TWO JOBS BEFORE YOU MODEL

clean the text, then turn it into numbers

PREPROCESS — CLEAN THE TEXT

make raw text analyzable

- Remove **stop words** ("this", "is")
- **Normalize** word forms — stemming and lemmatization
- **Standardize** acronyms, slang, special characters
- Also: encoding, spelling and grammar checks

FEATURE ENGINEERING — MAKE NUMBERS

convert cleaned text into model input

- **Tokenize** words into a DataFrame
- **Bag of words** — per-document word frequency
- **TF-IDF** — weight by frequency across documents
- Feed features into classification or sentiment models

05 AWS LANGUAGE SERVICES AT A GLANCE

one job each — exam gold

SERVICE	WHAT IT DOES	DIRECTION / REMEMBER
Amazon Transcribe	Recognizes speech in audio files and produces a transcription; custom vocabularies, WebSockets, real-time subtitles	Speech → text
Amazon Polly	Converts text into lifelike speech from plain text or SSML; outputs MP3, Vorbis, and PCM audio	Text → speech
Amazon Translate	High-quality, on-demand text translation; integrates with Comprehend, Transcribe, and Polly	Language → language
Amazon Comprehend	Uses NLP to extract insights — entities, key phrases, language, sentiment, parts of speech	Insights from text
Amazon Lex	Builds conversational voice and text interfaces using the Amazon Alexa engine; scales with AWS Lambda	Chatbots / conversation

06 TEXT ANALYSIS: THREE CATEGORIES

know the signature technique of each

CATEGORY	WHAT IT DOES	TECHNIQUE / EXAMPLE
Classifying text	Extract features, then run them through an ML algorithm and a classifier model to infer a category	NLTK classification (e.g., document topics)
Discovering similarities	Match text for autocorrect, spell check, and grammar check	Edit distance (Levenshtein distance)
Deriving relationships	Relate different words or phrases within the text	Coreference resolution

07 EXAM TRAPS

where the knowledge check gets you

- × "NLP is a machine learning invention." — NLP **predates ML**; early speech-to-text and screen readers used no ML. Many NLP systems now use ML, but NLP is the older domain.
- × "Amazon Polly turns speech into text." — Backwards. **Polly is text → speech**; **Transcribe** is speech → text. The two run in opposite directions.
- × "Removing stop words and normalizing are the same step." — No. Removing stop words **deletes** low-value words ("this", "is"); normalizing **collapses** word forms (run, ran, running → run) via stemming and lemmatization.
- × "Bag of words captures word order and meaning." — It only records **word frequency** per document as a vector; order is ignored ("is" scores 2 simply because it appears twice).
- × "In TF-IDF, common words get the highest weight." — Reversed. Words that appear in **many documents** receive a **lower** weight; the model favors more distinctive words.
- × "Amazon Comprehend can transcribe audio." — Comprehend works on **text** (entities, key phrases, language, sentiment, parts of speech). To start from audio, run **Transcribe** first, then feed the text to Comprehend.
- × "You must build your own conversational engine to use Amazon Lex." — Lex provides the **same engine that powers Amazon Alexa**, and it scales automatically with **AWS Lambda**.
- × "NER and a knowledge graph are the same thing." — **NER** extracts and classifies the entities; a **knowledge graph** then disambiguates them and derives meaning (the Amazon recommendations engine is a knowledge graph).

08 SELF-QUIZ

answers are upside-down below

Q1 Which came first as a computing domain, NLP or machine learning?

Q2 In the language hierarchy that NLP evaluates, what is the lowest-level unit, and what do groups of them form?

Q3 Name the four main challenges an NLP application must address.

Q4 Which preprocessing step deletes low-value words such as "this" and "is"?

Q5 Which two techniques collapse different forms of a word (run, runner, ran, running) into one common form?

Q6 Which feature model represents a sentence as a vector of word-frequency counts and ignores word order?

Q7 In TF-IDF weighting, do words that appear across many documents get a higher or a lower weight?

Q8 What is the process of extracting entities such as "Titanic" or "North Atlantic" from text called?

Q9 Which AWS service turns speech in an audio file into text, and which turns written text into lifelike speech?

Q10 Which AWS service uses NLP to extract entities, key phrases, language, and sentiment from a document?

ANSWER KEY (rotate the page)

Q1 NLP — it predates machine learning. **Q2** Words are the lowest level; groups of words form phrases, phrases form sentences, and sentences convey ideas. **Q3** Discovering the structure of text, labeling data, representing context, and applying grammar. **Q4** Removing stop words. **Q5** Stemming and lemmatization (normalization). **Q6** Bag of words. **Q7** A lower weight. **Q8** Named entity recognition (NER). **Q9** Amazon Transcribe (speech → text); Amazon Polly (text → speech). **Q10** Amazon Comprehend.