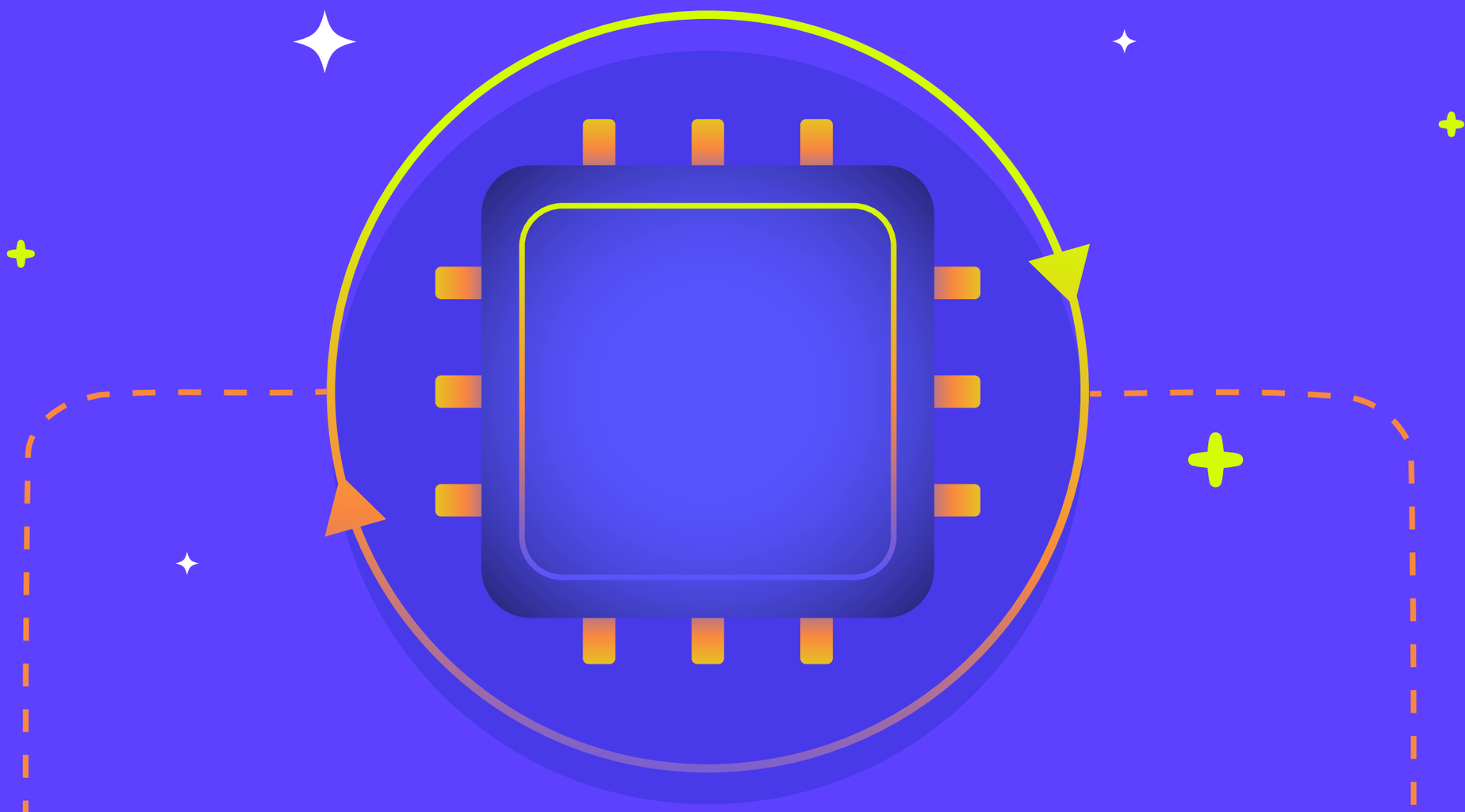
educative



7 key steps in NLP Preprocessing

educative

TOKENIZATION

Break down the text into “tokens.”

Example: The cat sat on the bed.

Tokens: *The, cat, sat, on, the, bed*

```
5   # Creating token of words
6   print("Creating token of words:")
7   from nltk.tokenize import word_tokenize
8   text="My name is Adithya Challa I wrote this shot!"
9   tokenize_word=word_tokenize(text)
10  print(tokenize_word)
11  print("\n")
```

educative

STEMMING

Remove the prefixes and suffixes to obtain the root word.

Example:
List of words: Affection, Affects, Affecting, Affected, Affecting

Root word: Affect

```
13  # Stemming
14  print("Stemming:")
15  from nltk.stem import PorterStemmer
16  words=["light","lighting","lights"]
17  ps=PorterStemmer()
18  for w in words:
19      rootword-ps.stem(w)
20      print(rootword)
21  print("\n")
```

educative

LEMMATIZATION

Group together different inflected forms of a word into a base word called *“lemma.”*

Example:
List of words: going, gone, went

Lemma: go

```
23  # Lemmatization:Converts allverb forms into root word
24  print("Lemmatization:Converts allverb forms into root word:")
25  from nltk.stem import WordNetLemmatizer
26  lem=WordNetLemmatizer()
27  print(lem.lemmatize("playing"))
28  print("\n")
```

educative

POS TAGGING

We identify the parts of speech (POS) for various tokens.

Example:
Sentence: The dog killed the bat.

Parts of speech: Definite article, noun, verb, definite article, noun.

```
30  # POS Tag
31  print("POS Tag:")
32  from nltk import word_tokenize,pos_tag
33  text="My name is Adithya Challa I wrote this shot!"
34  print(pos_tag(word_tokenize(text)))
```

educative

NAMED ENTITY RECOGNITION

Classify named entities mentioned in the text into categories such as “People,” “Locations,” “Organizations,” and so on.

Example:

Text: Google CEO Sundar Pichai resides in New York.

Named entity recognition:

Google – Organization
Sundar Pichai – Person
New York – Location

educative

CHUNKING/SEGMENTATION

Place individual pieces of information and group them. Chunking combines tokens into larger units, typically based on their grammatical roles.

Example:
Text: " quick brown fox jumps over the lazy dog."

Chunking Goal: Extract noun phrases.

Result:"The quick brown fox", "the lazy dog"

STOP WORDS REMOVAL

The goal is to remove commonly occurring words in segments of text that don’t add much information/value to the text.

Example:

“the,” “a,” and “an,” are common examples.

educative

DON'T STOP HERE!

Continue diving into NLP and AI to future-proof your career as we continue further into the new age of AI!

Happy learning!