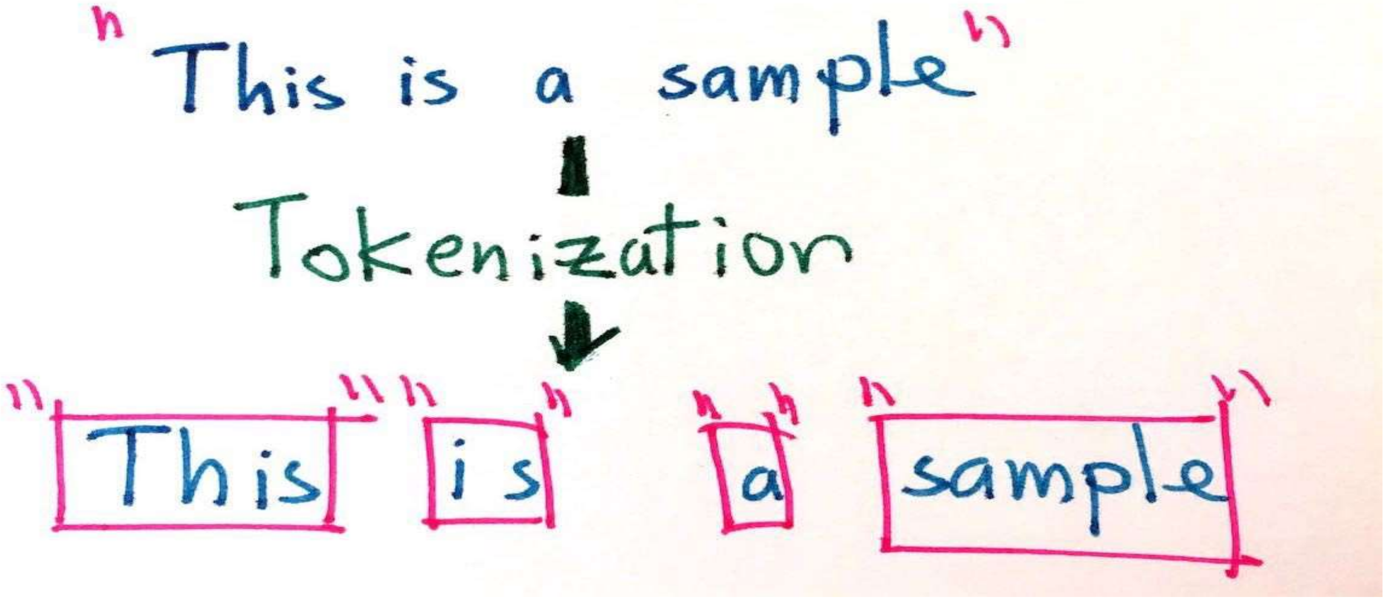
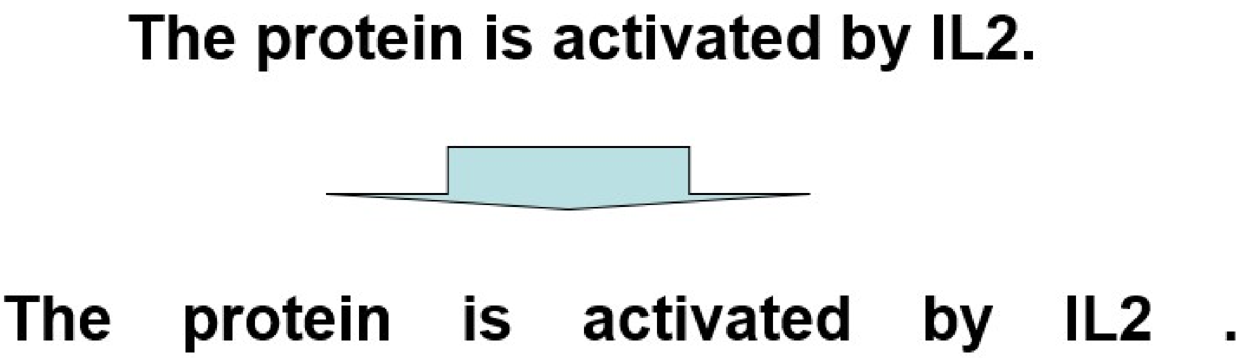
Tokenization

Divides the text into smallest units (usually words), removing punctuation.





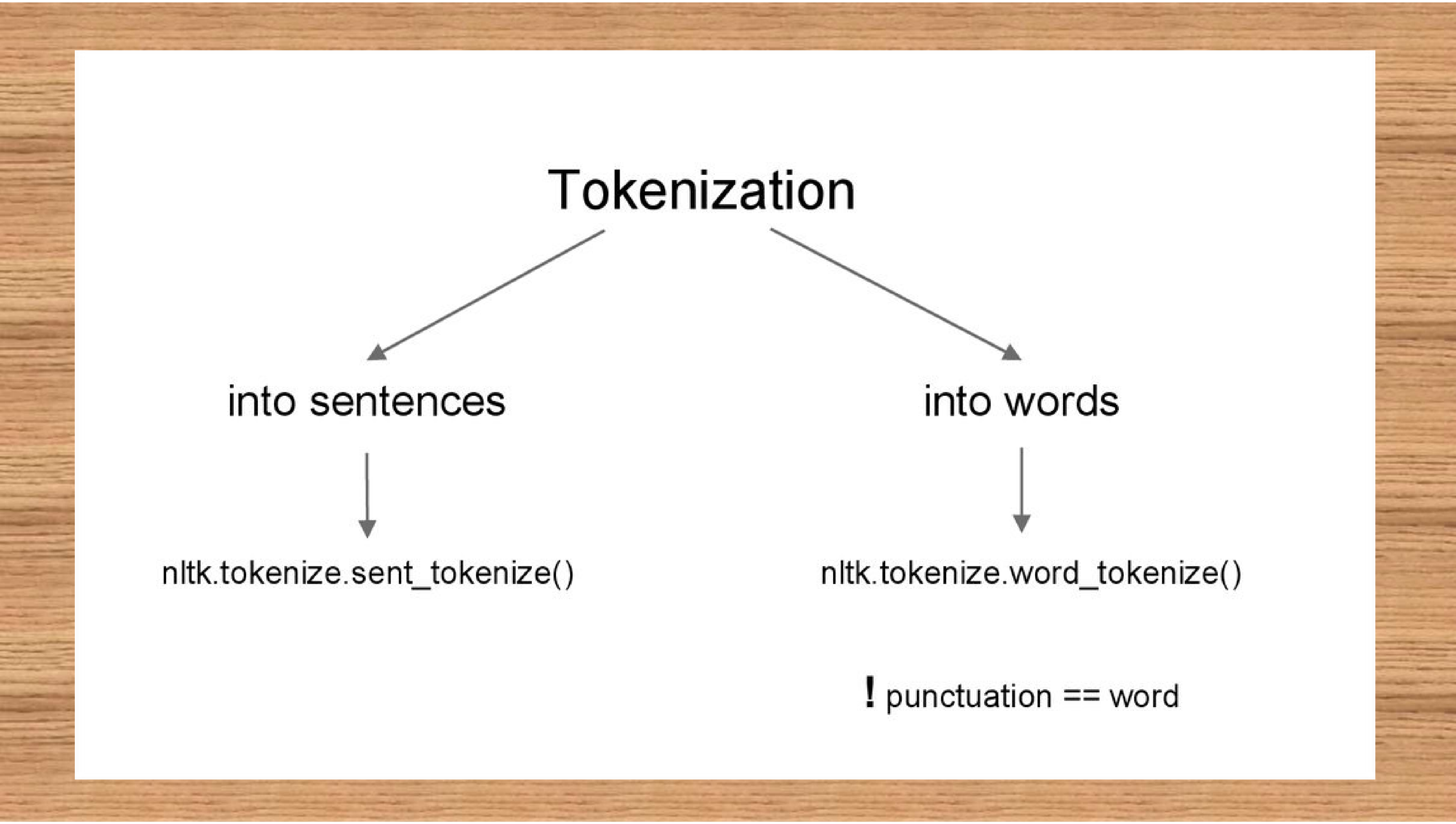
The protein is activated by IL2.

The protein is activated by IL2 .

Convert a sentence into a sequence of *tokens*

Why do we tokenize?

Because we do not want to treat a sentence as a sequence of *characters*!



Tokenization

into sentences

nltk.tokenize.sent_tokenize()

into words

nltk.tokenize.word_tokenize()

! punctuation == word

Tokenisation issues

separate possessive endings or abbreviated forms from preceding words:

- Mary’s → Mary ‘s
  Mary’s → Mary is
  Mary’s → Mary has

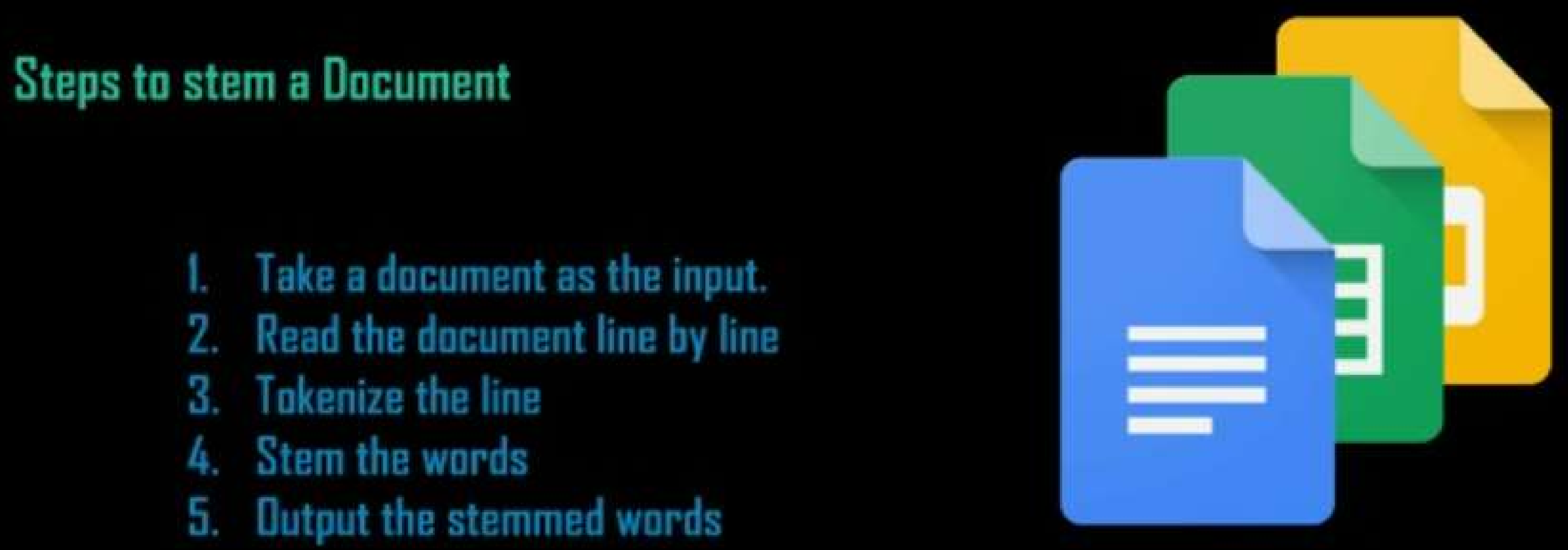
separate punctuation marks and quotes from words :

- Mary. → Mary .
- “new” → “ new “

Stemming

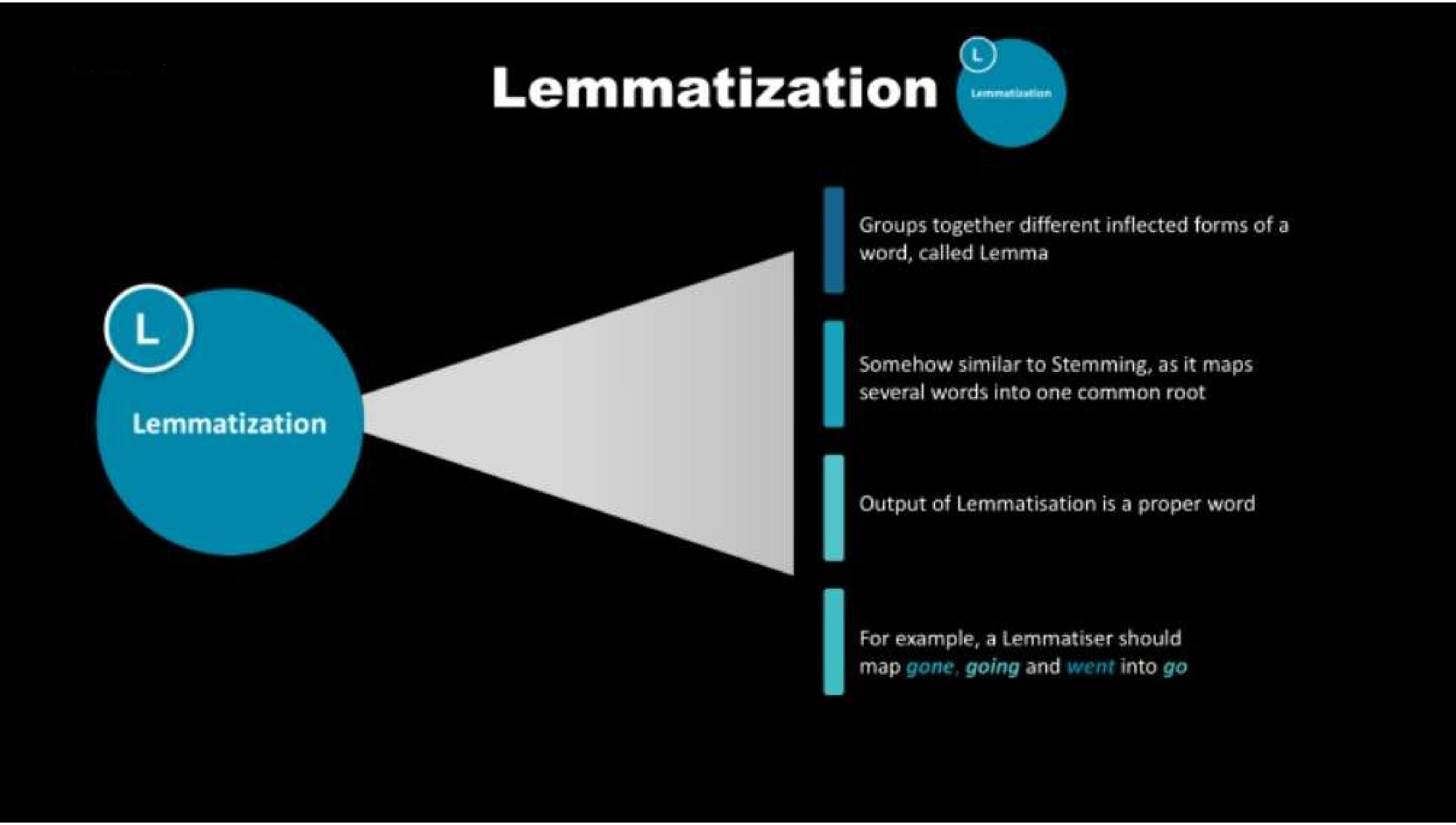
Stemming is the process of reducing inflection in words to their "root" forms such as mapping a group of words to the same Stem

Steps for Stemming



Steps to stem a Document

1. Take a document as the input.
2. Read the document line by line
3. Tokenize the line
4. Stem the words
5. Output the stemmed words



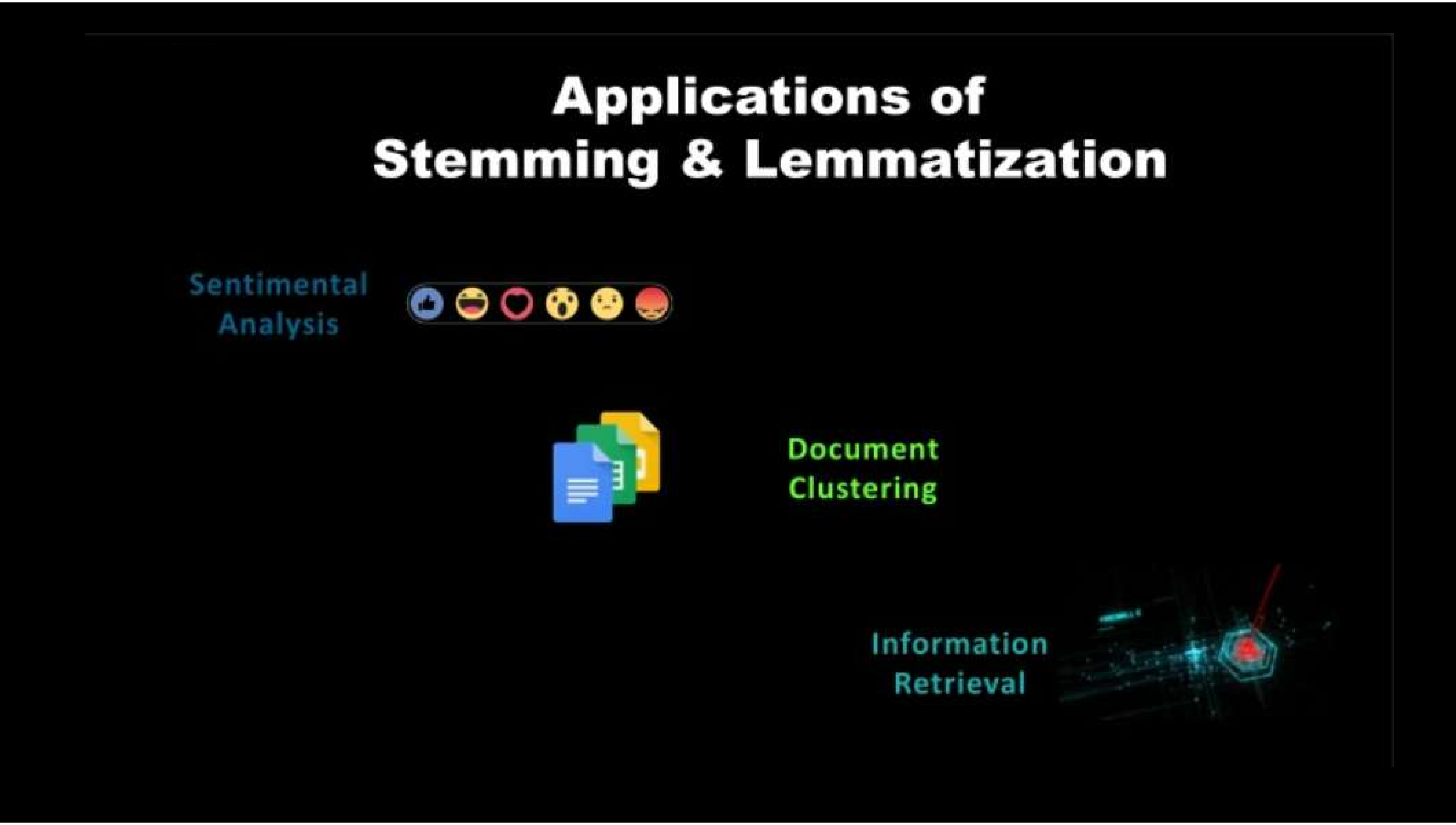
Lemmatization

Groups together different inflected forms of a word, called Lemma

Somehow similar to Stemming, as it maps several words into one common root

Output of Lemmatisation is a proper word

For example, a Lemmatiser should map ***gone***, ***going*** and ***went*** into ***go***

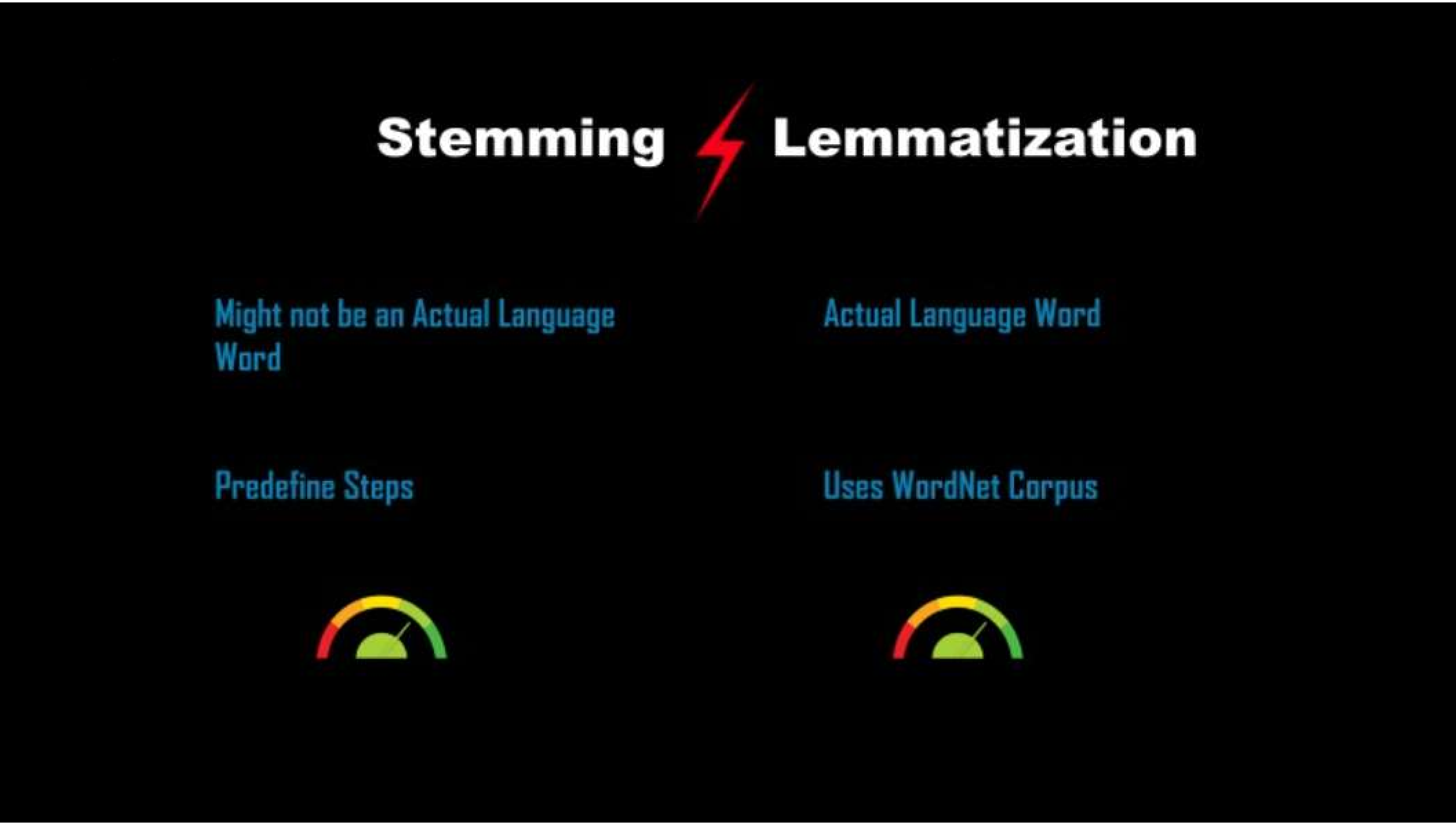


Applications of Stemming & Lemmatization

Sentimental Analysis

Document Clustering

Information Retrieval



Stemming ⚡ Lemmatization

Stemming	Lemmatization
Might not be an Actual Language Word	Actual Language Word
Predefine Steps	Uses WordNet Corpus

Tokenization

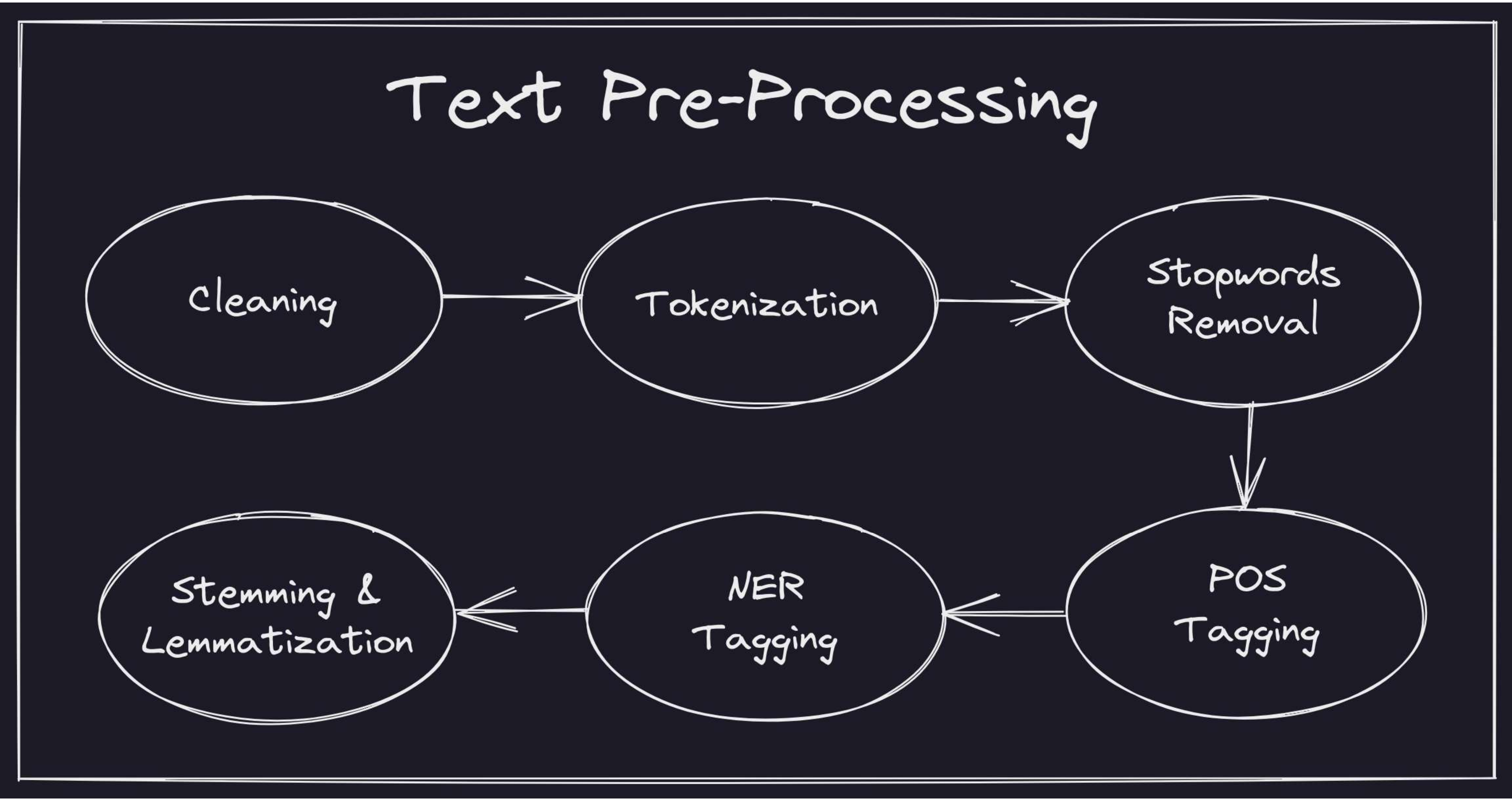
Stemming

Lemmatization

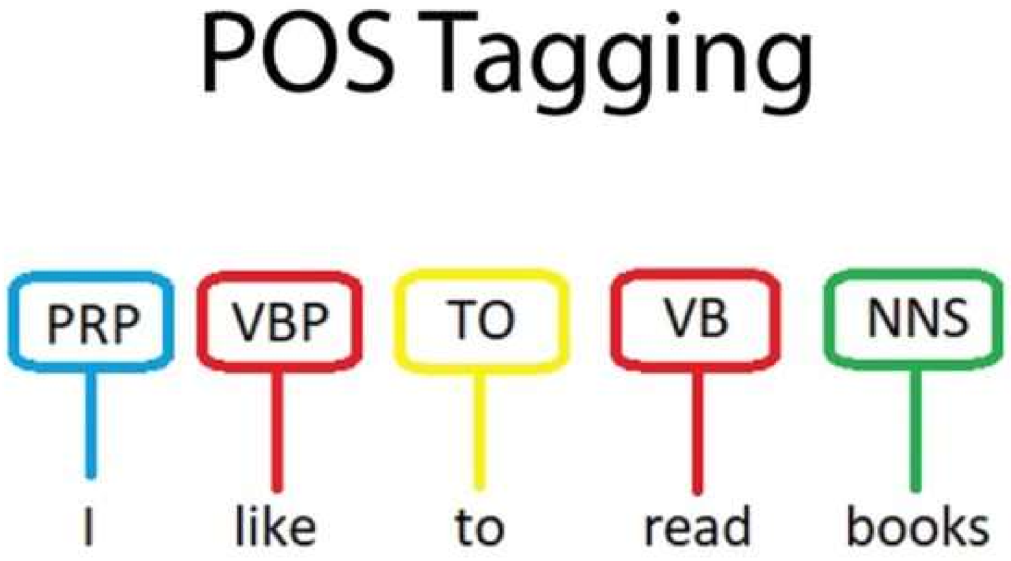
POS Tags

Named Entity Recognition

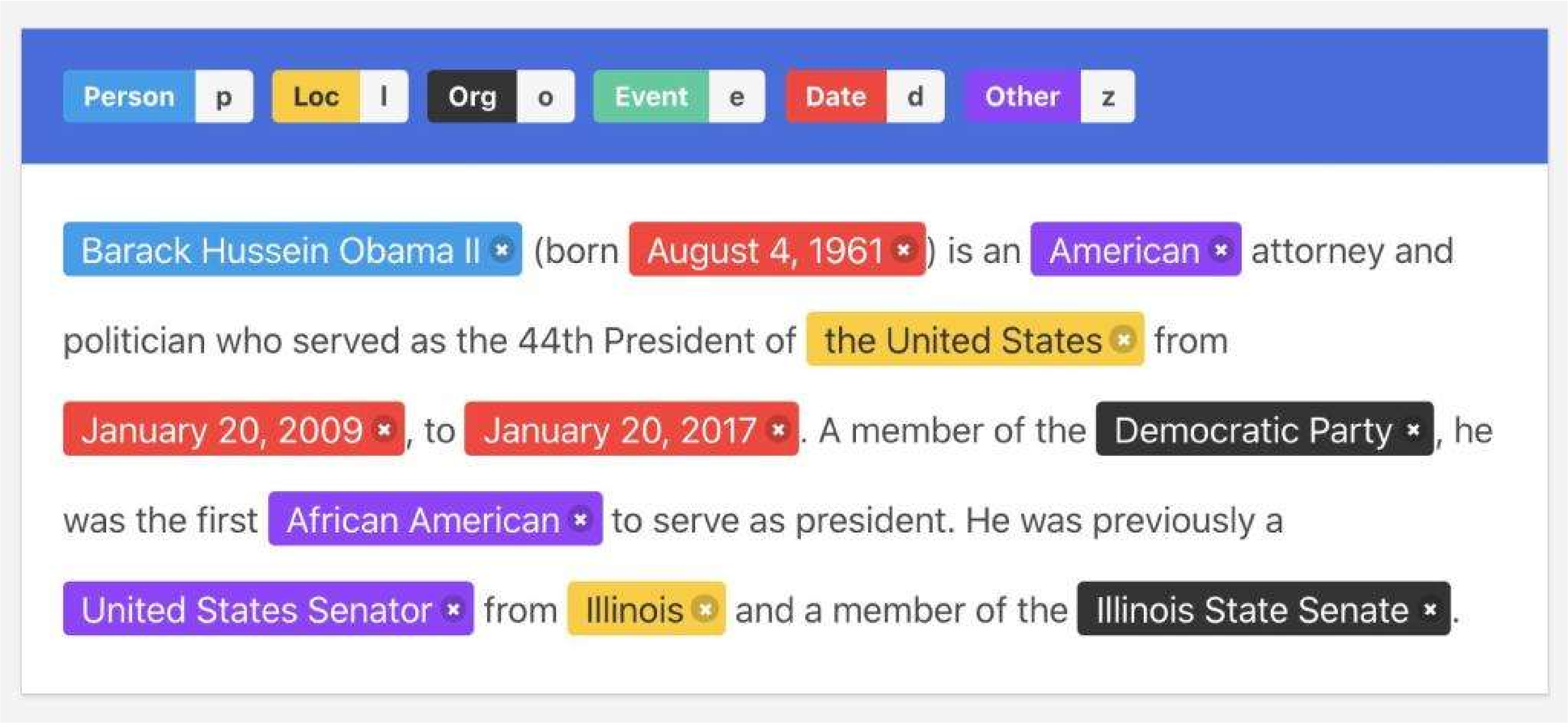
Chunking


Text Pre-Processing
Cleaning
Tokenization
Stopwords Removal
POS Tagging
NER Tagging
Stemming & Lemmatization

POS Tagging



NER Tagging

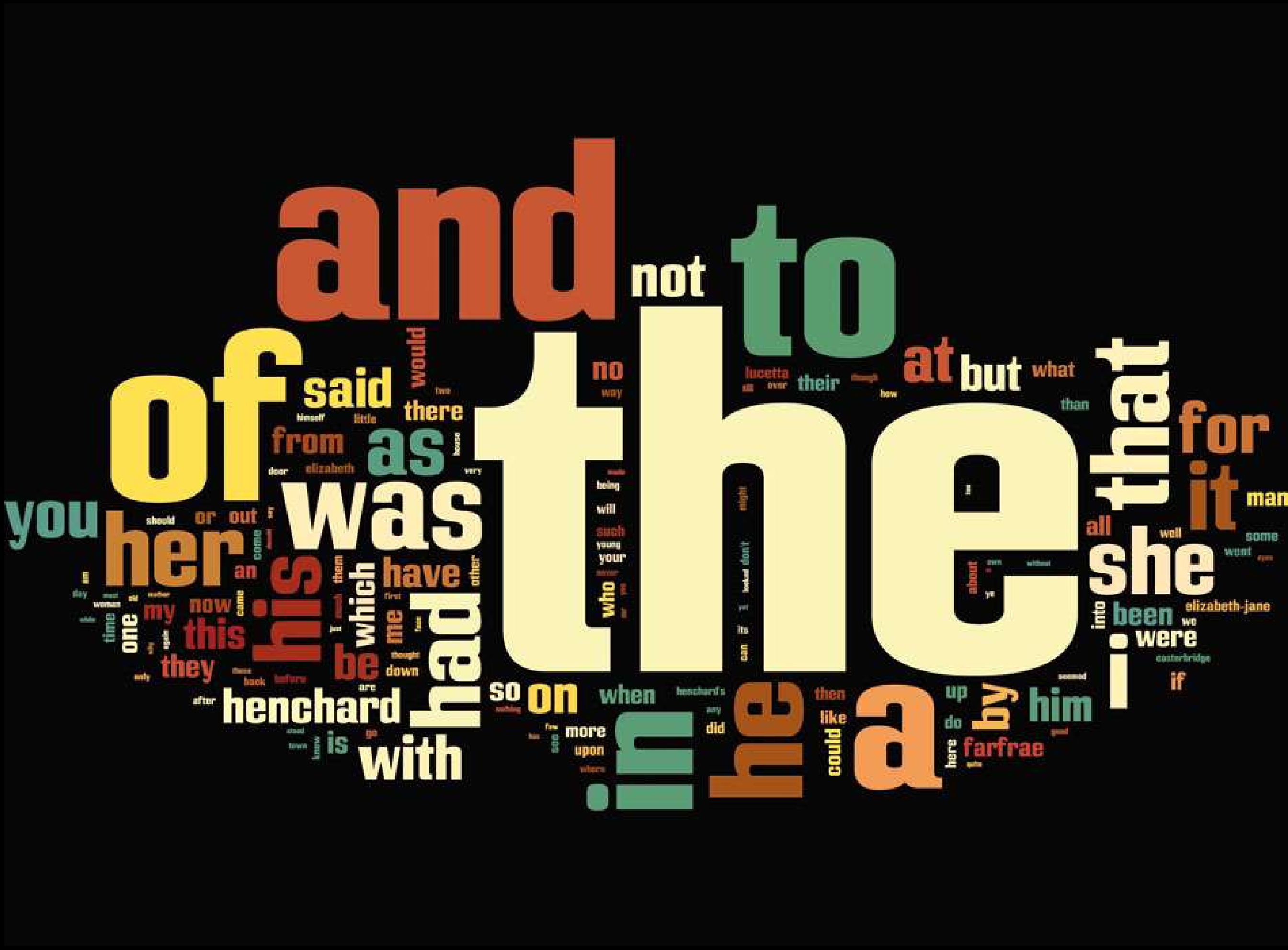


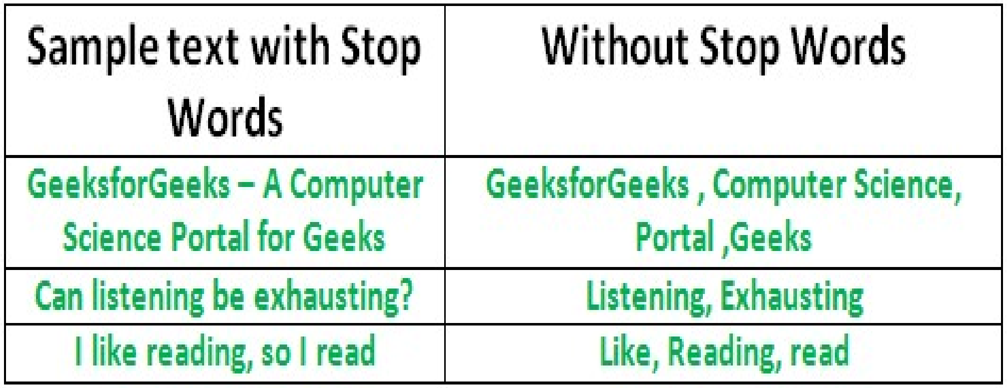
Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

Stopwords

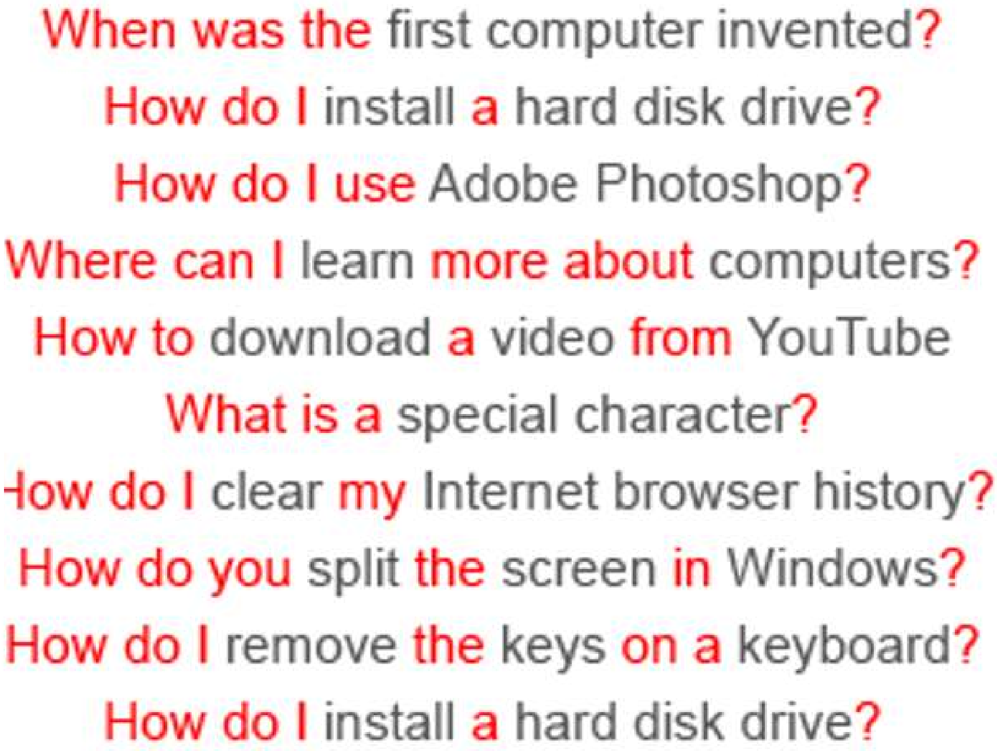
- **Words filtered out before processing a natural language** are called stop words.

- The most common words in any language, like articles, prepositions, pronouns, conjunctions, etc.





Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read



When was the first computer invented?

How do I install a hard disk drive?

How do I use Adobe Photoshop?

Where can I learn more about computers?

How to download a video from YouTube

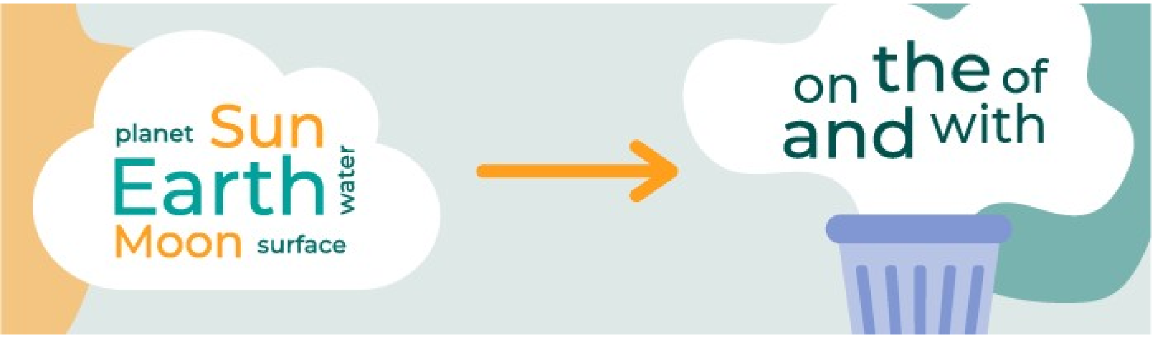
What is a special character?

How do I clear my Internet browser history?

How do you split the screen in Windows?

How do I remove the keys on a keyboard?

How do I install a hard disk drive?



Data Pre-processing

- Cleaning
- Tokenization
- Stemming
- Lemmatization
- Stopwords
- “Hello!!!. This Dr. Ghazaala Yasmin, Welcome to NLP class in SCSET, Bennet university…..”

Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a "."
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning