

## Tokenization

*tokenization* is a particular kind of document segmentation. *Segmentation* breaks up text into smaller chunks or segments, with more focused information content. **Segmentation can include breaking a document into paragraphs, paragraphs into sentences, sentences into phrases, or phrases into tokens (usually words) and punctuation.**

Tokenization- It’s the process of segmenting text into *tokens* .

A tokenizer breaks unstructured data, natural language text, into chunks of information that can be counted as discrete elements

#### Sentence Tokenization

Sentence tokenizer breaks text paragraph into sentences.

#### Word Tokenization

Word tokenizer breaks text paragraph into words.

## Stopwords

Stopwords considered as noise in the text. Text may contain stop words such as is, am, are, this, a, an, the, etc.

**Stop Words @Raghav is tuning a great AI model, at //raghavAI.com and teaching AI**

**and**

**is**

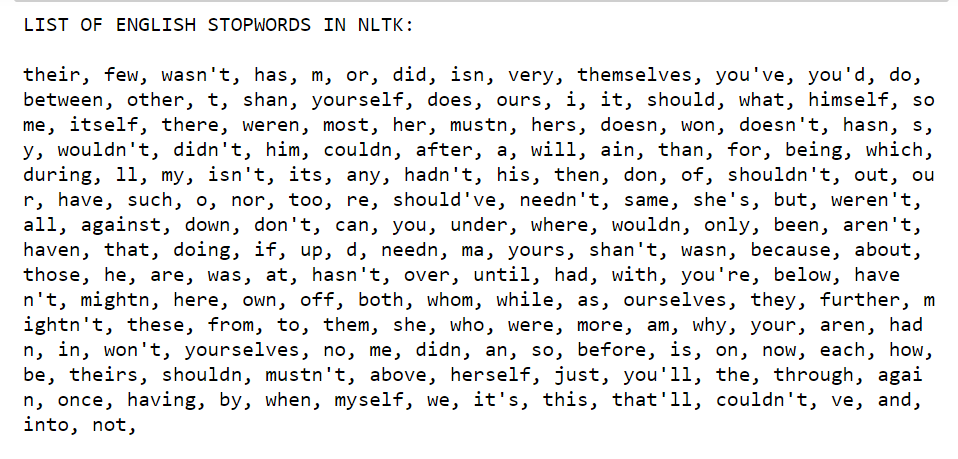
**a**

**at**

**has**

**for**

**of**



## Removing punctuations, URLs etc

## Lexicon Normalization

Lexicon normalization considers another type of noise in the text. For example, connection, connected, connecting word reduce to a common word "connect". It reduces derivationally related forms of a word to a common root word.

#### Stemming

Stemming is a process of linguistic normalization, which reduces words to their word root word or chops off the derivational affixes. For example, connection, connected, connecting word reduce to a common word "connect"

Transforming a word to its Base stem

Tuning great AI Model

Tuning

Tun

**Tuned Tune Tuning**

*Porter's algorithm- Porter Stemming- 1980*

[http://www.tartarus.org/~martin/PorterStemmer/](http://www.tartarus.org/%3CTT%3E~%3C/TT%3Emartin/PorterStemmer/)

\begin{example}
\begin{tabular}[t]{lll@{\hspace{1in}}lll}
\multicolumn{3}{l}{\te...
... & $\rightarrow$\ & & cats & $\rightarrow$\ & cat \\
\end{tabular}\end{example}.

**Lemmatization**

wordforms to a common root- **lemma**.

Wordforms- am', 'are', and 'is

Lemma- be

'dinner' and 'dinners- Dinner

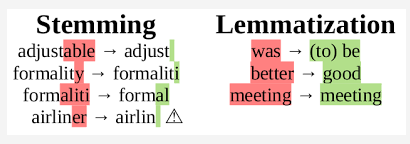
Lemmatization is a lot more powerful. It looks beyond word reduction and considers a language’s full vocabulary to apply a **morphological analysis** to words, aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the **lemma**.

Original Word ---> Root Word (lemma) Feature

meeting ---> meet (core-word extraction)

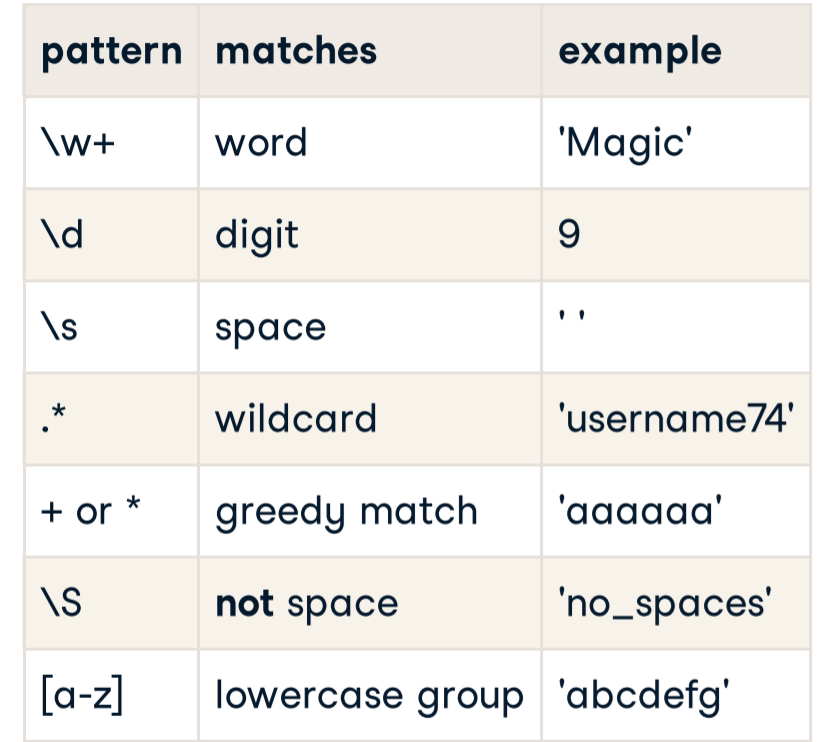
was ---> be (tense conversion to present tense)

mice ---> mouse (plural to singular)



## Advanced Processing

Regular Expressions- Creating simple String patterns and use them for pater matching for searching/replacing text



parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article

POS tagging, as identifying proper nouns can be useful in identifying person and organization names

e.g book can be a verb (book that flight) or a noun (hand me that book)

