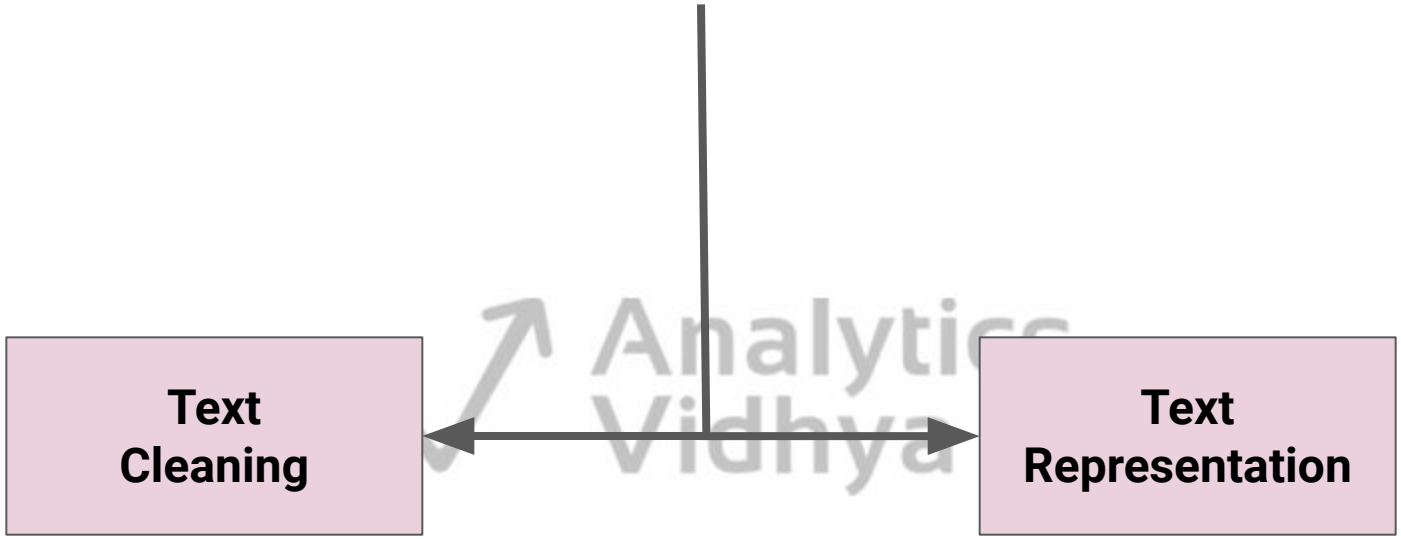
Text Preprocessing

Text Preprocessing



Text Cleaning

I can't believe ho%^&w many people are read&ing my book.

unwanted symbols

Text Cleaning

I can’t believe how many people are reading my book.

unwanted symbols

Text Cleaning

I can’t believe how many people are reading my book.

unwanted symbols

Deep learning is part of a broader family of machine learning (https://en.wikipedia.org/wiki/Machine_learning) methods

URL Links

Text Cleaning

I can’t believe how many people are reading my book.

unwanted symbols

Deep learning is part of a broader family of machine learning methods

URL Links

Text Cleaning

I can't believe how many people are reading my book.

unwanted symbols

Deep learning is part of a broader family of machine learning methods

URL Links

* The data revolution has transformed the way businesses run and customers behave. *

HTML Tags

Text Cleaning

I can’t believe how many people are reading my book.

unwanted symbols

Deep learning is part of a broader family of machine learning methods

URL Links

The data revolution has transformed the way businesses run and customers behave.

HTML Tags

Text Noise

- Unwanted or useless information present in text data.

Text Noise

- Unwanted or useless information present in text data.
- Examples of noise:

Text Noise

- Unwanted or useless information present in text data.
- Examples of noise:
  - URLs, punctuation marks, numbers
    (https://en.wikipedia.org/wiki/, {. , : ! ? ;}

Text Noise

- Unwanted or useless information present in text data.
- Examples of noise:
  - URLs, punctuation marks, numbers
    (https://en.wikipedia.org/wiki/, {. , : ! ? ;}
  - Slangs: Non-dictionary words
    (swag, dope, bro, etc)

Text Noise

- Unwanted or useless information present in text data.
- Examples of noise:
  - URLs, punctuation marks, numbers
    (https://en.wikipedia.org/wiki/, {. , : ! ? ;}
  - Slangs: Non-dictionary words
    (swag, dope, bro, etc)
  - Spelling mistakes
    (cntrol, playe, finaly, etc)

Steps for Text Cleaning

- Fix text encoding and casing:

Steps for Text Cleaning

- Fix text encoding and casing:
    - Different encodings for different languages such as ascii (English), West Europe (Latin), Big5 (Chinese), etc.

Steps for Text Cleaning

- Fix text encoding and casing:
    - Different encodings for different languages such as ascii (English), West Europe (Latin), Big5 (Chinese), etc.

Letter	ASCII Code
a	097
b	098
c	099
d	100
e	101
f	102

Steps for Text Cleaning

- Fix text encoding and casing:
    - Different encodings for different languages such as ascii (English), West Europe (Latin), Big5 (Chinese), etc.
    - Universal encoding: utf8

Steps for Text Cleaning

- Fix text encoding and casing:
  - Different encodings for different languages such as ascii (English), West Europe (Latin), Big5 (Chinese), etc.
  - Universal encoding: utf8
  - Lowercase entire text

Steps for Text Cleaning

- Fix text encoding and casing:
  - Different encodings for different languages such as ascii (English), West Europe (Latin), Big5 (Chinese), etc.
  - Universal encoding: utf8
  - Lowercase entire text
    “Cap” and “cap” are different terms for machine

Steps for Text Cleaning

- Noisy entities removal:
  - URLs, special characters, HTML tags

Steps for Text Cleaning

- Noisy entities removal:
    - URLs, special characters, HTML tags
- Punctuation marks removal:

    Example: “I went to New York!!”

Steps for Text Cleaning

- Noisy entities removal:
  - URLs, special characters, HTML tags
- Punctuation marks removal:

  Example: “I went to New York!!”
  Cleaned: “I went to New York”

What are Regular Expressions?

- Regular Expressions or Regex use patterns in the text

What are Regular Expressions?

- Regular Expressions or Regex use patterns in the text
- Patterns are special characters with an associated textual meaning.

What are Regular Expressions?

- Regular Expressions or Regex use patterns in the text
- Patterns are special characters with an associated textual meaning. For example,
  - “\d” is used to identify digits
  - “\w” is used to identify alphanumeric characters

Thank You