

# Analyzing Text Data

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Agenda

- Text Preprocessing
- Text Vectorization
- Sentiment Analysis

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Let's begin the discussion by answering a few questions on Text Preprocessing

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.  
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

Which of the following tasks are performed during text cleaning?

A

Stemming

B

Lowercasing

C

Removal of Special Characters

D

Adding extra white spaces

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

Which of the following tasks are performed during text cleaning?

A

Stemming

B

Lowercasing

C

Removal of Special Characters

D

Adding extra white spaces

This file is meant for personal use by ankur.agarwal02@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing

Process of preparing and refining raw text data by removing noise, such as special characters, punctuation, stop words, and irrelevant symbols, to standardize text data

Improves the suitability of the data for analysis and modeling

## Stopword Removal

Removes common words like “and”, “the”, “is”, etc which often appears frequently and generally do not add ‘contextual value’ to the text

## Stemming

Converts the word into its root form, reducing it to its base or stem, to capture the core meaning

## Lowercasing

Converts all the words into lower case letters

## Remove special characters

Removes special characters like “, ”, !, @, etc

## Strip extra white spaces

Removes extra spaces between the words

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing - Example

Text Preprocessing is key in natural language processing (NLP). It involves refining raw text for machine understanding and analysis. This process includes eliminating redundant words, reducing words to their root form, standardizing text to lowercase, excluding symbols (!@#\$), ensuring clean spaces, and filtering out numerical characters 1234.

Text  
Cleaning



text preprocess key natur languag  
process nlp involv refin raw text  
machin understand analysi process  
includ elimin redund word reduc word  
root form standard text lowercas  
exclud symbol ensur clean space filter  
numer charact 1234

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

Consider the following three sentences:

"The cat jumped"  
"The dog barked"  
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying the bag-of-words (BoW) model to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A

6

B

9

C

12

D

15

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.



# Text Preprocessing Quiz

Consider the following three sentences:

"The cat jumped"  
"The dog barked"  
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying the bag-of-words (BoW) model to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A

6

B

9

C

12

D

15

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Bag of Words (BoW) model

Represents text by **counting the frequency of unique words** in a document **without considering the order or structure** of the words

Creates a "bag" (or set) of words in a text corpus, ignoring grammar and word order

## Example:

Sentence	the	cat	jumped	dog	barked	chased
The cat jumped	1	1	1	0	0	0
The dog barked	1	0	0	1	1	0
The cat chased the dog	1	1	0	1	0	1

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

Words with high frequency are always considered to be stop words.

A

True

B

False

This file is meant for personal use by ankur.agarwal02@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

Words with high frequency are always considered to be stop words.

A

True

B

False

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Stopwords

Words in any language which do not add much meaning to a sentence.

Can safely be ignored without sacrificing the meaning of the sentence.

## Example:

Consider you are analyzing a collection of documents related to cooking recipes.

The word "salt" might appear very frequently across these documents due to its significance in cooking.

However, "salt" will not be considered a stop word in this context because it carries essential meaning within the domain.

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

The main characteristic of a lexicon-based approach in sentiment analysis is that it utilizes a predefined set of words with assigned sentiment scores

A

True

B

False

This file is meant for personal use by ankur.agarwal02@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Text Preprocessing Quiz

The main characteristic of a lexicon-based approach in sentiment analysis is that it utilizes a predefined set of words with assigned sentiment scores

A

True

B

False

This file is meant for personal use by ankur.agarwal02@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

# Sentiment Analysis

The process of analyzing a piece of text and categorizing it based on the context and emotions expressed within the text

In general, categorization is done as positive, negative, or neutral

Used extensively to analyze end-user feedback, gain insights on the sentiment of the user towards a product/service, and identify areas of improvement

## Example:

“The movie was fun, brisk, and imaginative” =>

**Positive**

This file is meant for personal use by ankur.agarwal02@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.



# Sentiment Analysis - Lexicon-based Approach

Uses predefined dictionaries (or lexicons) containing words and their associated sentiment labels

Each word in new, unseen text is compared against the lexicons to determine its sentiment label, and then the values are combined to get the prediction

This file is meant for personal use by [ankur.agarwal02@gmail.com](mailto:ankur.agarwal02@gmail.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.



# Happy Learning !

