6 steps_ NLP_WorkFlow

Acquisition

- In Natural Language Processing (NLP), this data is referred to as a corpus.
- A corpus is a structured collection of text data, typically compiled for research or analysis within a specific domain
- corpus can be sourced from diverse file formats, including .txt, .csv, and .xml.
- Common data sources encompass speech transcriptions, web-scraped content, and text-based reviews.

Inspection and exploration

- understanding the data's structure, identifying noise, and determining how to clean the data for machine learning applications
- This stage is also known as Exploratory Data Analysis (EDA), which involves examining independent variables, dependent variables, variable types, and variable data types to understand the characteristics of the data and its inherent structural relationships.

Preprocessing and cleansing

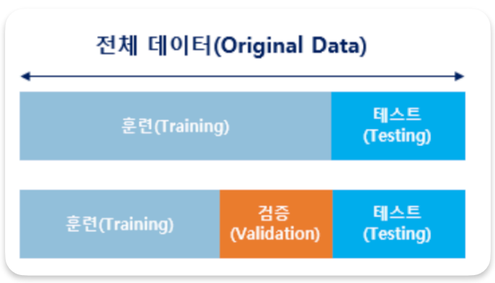
- This stage involves many steps, including tokenization, cleaning, normalization, and stop word removal in the case of natural language processing

Modeling and Training

- selected a suitable machine learning algorithm, you train the machine with the preprocessed data.
- After the machine has learned from the data and the training is successful, it can perform natural language processing tasks such as machine translation, speech recognition, and text classification.
- you shouldn't train the machine with all of the data.
- some of the data should be reserved for testing, and only the training data should be used for training->allows you to measure the current performance using the test data after the machine has learned and prevent **overfitting**.

Ideally, you should divide the data into three sets: training, validation, and testing, and only use the training data for training.

Validation Data Vs Test Data



Validation data is used to assess the current performance of the model, i.e., how well the machine has learned from the training data

Test data, on the other hand, is used to evaluate the final performance of the model.It's not used for improving the model but rather to quantify and evaluate the model's performance.

Evaluation

Once training is complete, performance is evaluated using the test data. The evaluation method measures how closely the data predicted by the machine matches the actual correct answers in the test data.

Deployment