

Practical 7

AIM:

Extract sample document and apply following document preprocessing methods: tokenization, pos tagging, stop words removal, stemming and lemmatization.

REQUIREMENT:

AnacondaInstaller

Windows10OS

Linux

JupyterNotebook

THEORY:

Data Preprocessing in Python:

- Data preprocessing is the process of **cleaning and transforming raw data** before using it in algorithms.
 - Raw data is often **unstructured and not suitable** for analysis.
 - It helps in converting data into a **clean and usable format**.
-

Need of Data Preprocessing:

- Improves **accuracy of machine learning models**
 - Handles **missing values and inconsistencies**
 - Ensures data is in the **correct format for algorithms**
 - Helps in applying **multiple ML models on the same dataset**
-

NLP Techniques in Data Science:

1. Tokenization:

- Splitting text into **words or sentences**

2. Stop Words Removal:

- Removes common words like **“the”, “is”, “in”**
 - Helps reduce **unnecessary data and processing time**
-

Lemmatization:

- Converts words to their **base/root form with meaning**
- Example: *running* → *run*
- More **accurate than**

stemming Applications:

- Search engines
- Text analysis systems

Advantages:

- More accurate
- Maintains proper meaning

Disadvantages:

- Slower and more complex
-

Stemming:

- Reduces words to their **root form (may not be meaningful)**
- Example: *liked, likes* → *like*

Errors in Stemming:

- **Overstemming:** Different words become same root
- **Understemming:** Similar words remain different

.

LibrariesUsed:

NLTK:.

Sklearn

Pandas:

CONCLUSION:

In this experiment we have studied about data preprocessingusing natural language processing technique.