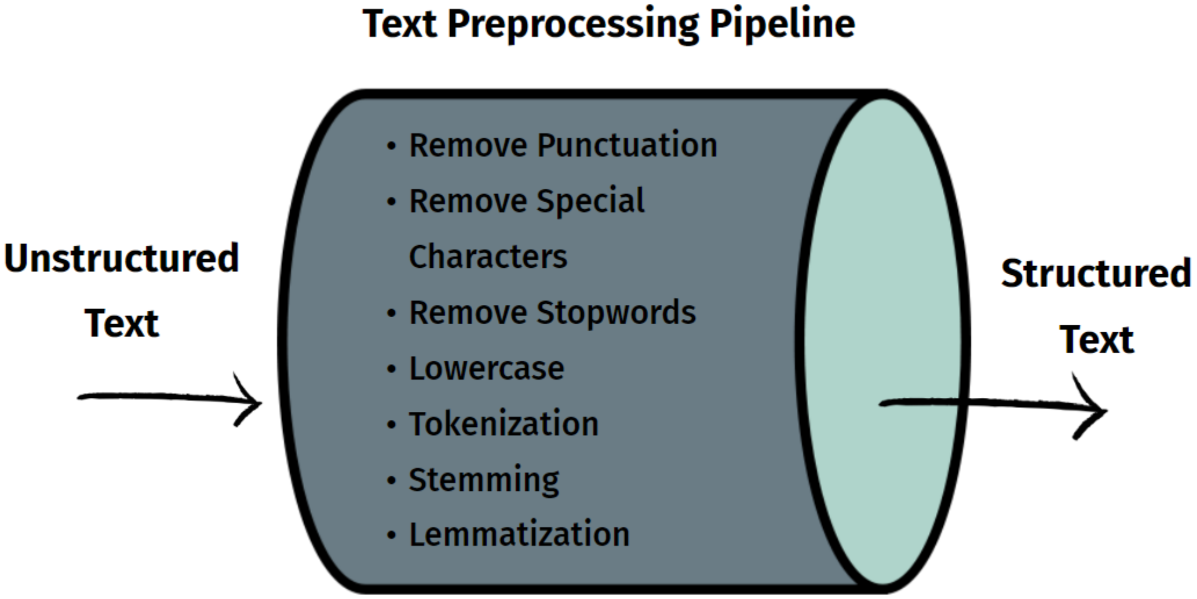
Text Preprocessing

The process of cleaning and transforming unstructured text data to prepare it for analysis.

What are the modules we are using?

a. **NLTK:** Stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human languages and respond in an appropriate manner.
   i. Using NLTK we can perform operations such as data cleaning, visualization, and vectorization that will help us in classifying our text.



Text Preprocessing Pipeline

- Remove Punctuation
- Remove Special Characters
- Remove Stopwords
- Lowercase
- Tokenization
- Stemming
- Lemmatization

Key Vocab:

- □ Stopwords
- □ NLTK
- □ Tokenize
- □ Tokens
- □ Stemming
- □ Lemmatization
- □ Porter Stemmer
- □ Wordnet

Exploratory Data Analysis

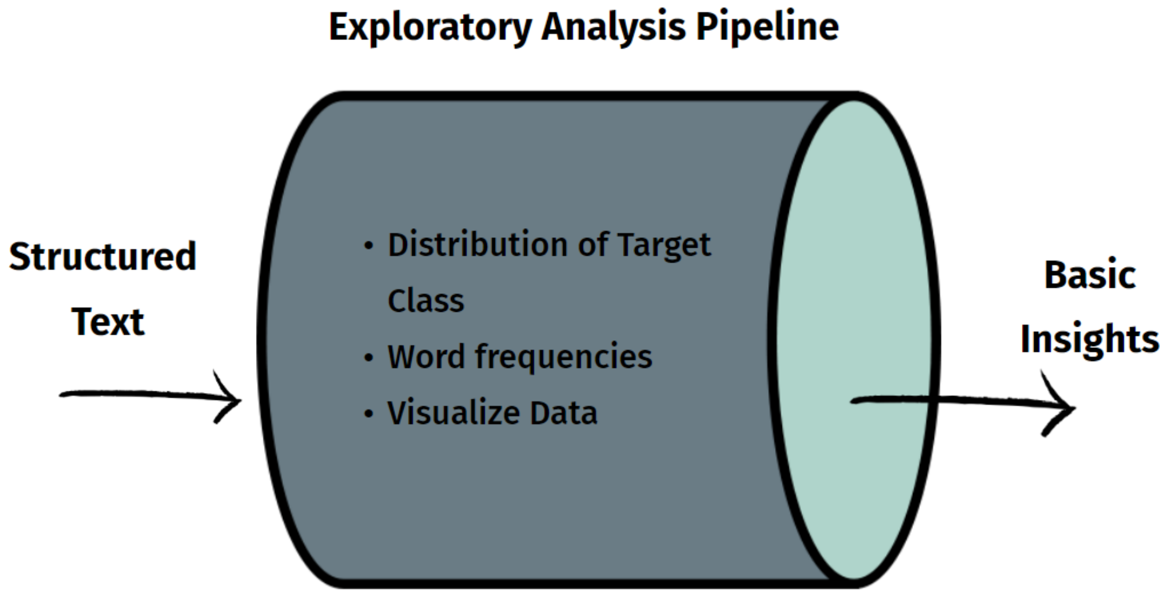
Exploratory Data Analysis is the process of exploring data, generating insights, testing hypotheses, checking assumptions and revealing underlying hidden patterns in the data.

What are the modules we are using?

b. **pandas:** It provides data structures and functions to efficiently work with structured data, such as tabular data or time series data.

   i. We will use pandas to create structured data frames (tables).

c. **Matplotlib**: Python library for creating static, animated, and interactive visualizations.



Key Vocab:

- □ Exploratory Data Analysis (EDA)
- □ Target Class
- □ Bigrams, Trigrams, Unigrams
- □ Word Frequencies