The image shows a machine learning pipeline diagram with boxes: Data Acquisition → Text Cleaning → Pre-Processing → Feature Engineering, with "Improving the model" labeled, and Monitoring and Model Updating ← Deployment ← Evaluation ← Modeling.

Handwritten notes on the right: "Word → Vector Represent / Self"

Input

Chaplin wrote, directed, and composed the music for most of his films.

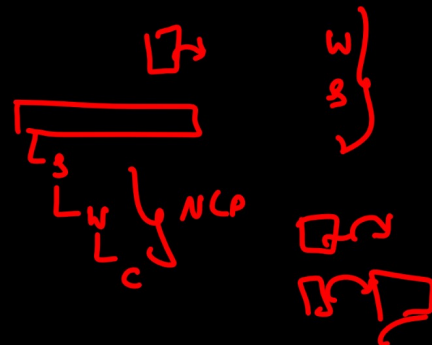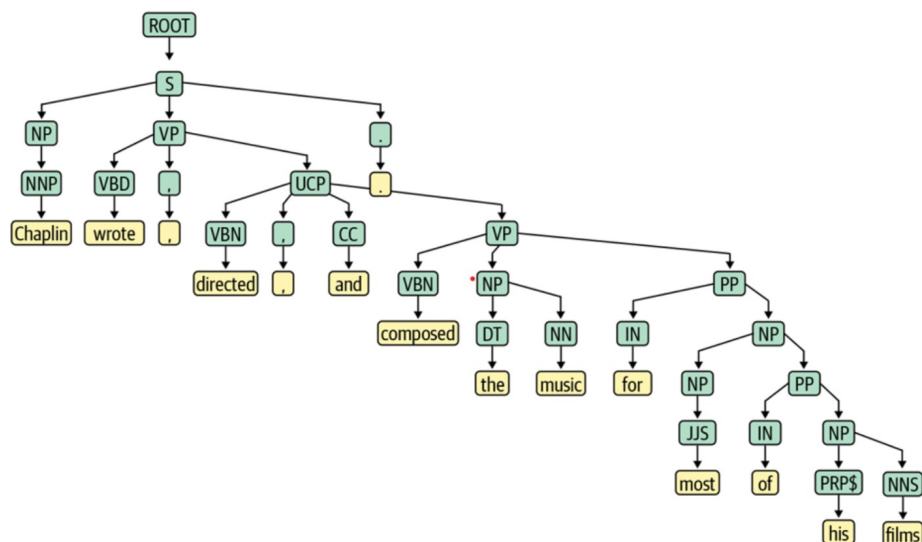## Tokenization with Lemmatization

Chaplin | write | . | direct | . | and | compose | the | music | for | most | of | he | film | .

Chaplin wrote, directed, and composed the music for most of his films.

POS Tagging

NNP | VBD | . | VBD | . | CC | VBN | DT | NN | IN | JJS | IN | PRP$ | NNS | .

Chaplin wrote, directed, and composed the music for most of his films.

Parse Tree

Classical NLP

Pre-processing

Documents → Tokenization → POS Tagging → Stopword Removal → ...

Modeling

Inference ← Modeling ← Feature Extraction

Output

- Sentiment
- Classification
- Entity Extraction
- Translation
- Topic Modeling

Deep Learning-Based NLP

Documents

Pre-processing

Dense Embeddings
obtained via word2vec,
doc2vec, GloVe, etc.

Hidden Layers

Output Units

Output
- Sentiment
- Classification
- Entity Extraction
- Translation
- Topic Modeling
- ...

Raw Text → Cleaning and Pre-processing → Tokenization - get words and other language units → Mathematically represent language units → Modeling → Evaluation