Part of Speech (POS) tagging

- POS tagging is the task of assigning tags to each word in a sentence according to its part of speech.

The	quick	brown	fox	jumped	over	the	lazy	dog
DET	VERB	VERB	NOUN	VERB	PREP	DET	VERB	NOUN

- The set of tags, or *Parts,* varies by dataset, task etc., and can be much larger than the standard set of eight.

- Used in syntactic parsing, and for word sense disambiguation (WSD)

```
leaves -> leave
  VERB

leaves -> leaf
  NOUN
```

Tokenization

- Whitespace tokenization
  - Segment on whitespace, compute vocabulary from top-k ranked words, add extra token for OOV words.
- Character tokenization
  - Segments on characters, very simple, but often limits performance on downstream tasks
- Subword tokenization methods
  - Byte-Pair Encoding (BPE)
    - Recursive algorithm to compute the common unicode character sequences in a dataset. Recursion stops based on frequency hyperparameter.
  - WordPiece
    - Similar to BPE, but instead of adding a sequence based on frequency alone, it normalizes frequency by the frequency of its constituent unicode character(s) / pairs.
  - Unigram Language Model
    - Starts with a complete vocabulary, and progressively shrinks it by removing tokens that result in the bottom percentile of log likelihood loss when removed.
  - SentencePiece
    - Completely agnostic to whitespace by including “\s” in the set of characters it recognizes, and is thus the only language agnostic tokenizer. It uses BPE+Unigram tokenization for subword regularization.