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

- Very common approach: segment on the space character

raw text: “The quick brown fox jumped over the lazy dog”

segmented text: [“The”, “quick”, “brown”, “fox”, “jumped”, “over”, “the”, “lazy”, “dog”]

- Once a tokenization scheme is decided on, we typically construct a lookup table mapping each token (or key) to an index that references the feature representation for that token. Here is an example of the BOW representation of the above text:

```
>>> raw_text = "the quick brown fox jumped over the lazy dog"
>>> segmented_text = raw_text.split(" ")
>>> bow = [0 for _ in range(len(set(segmented_text)))]
>>> lookup = {token: index for index, token in enumerate(set(segmented_text))}
>>> for token in segmented_text:
...     index = lookup[token]
...     bow[index] += 1
...
>>> bow
[1, 1, 1, 1, 1, 2, 1, 1]
>>> lookup
{'jumped': 0, 'over': 1, 'fox': 2, 'dog': 3, 'quick': 4, 'the': 5, 'lazy': 6, 'brown': 7}
```

- There are data-driven methods to do this that don’t rely on naively splitting on whitespace, and other feature representations too!