Stemming

- The process of removing word suffixes according to rules

```
connect    -> connect
connects   -> connect
connected  -> connect
connection -> connect
```

- Overstemming: When words with different canonical stems get reduced to the same root

```
universe   -> univers
universal  -> univers
university -> univers
```

- Understemming: When words with same canonical stem get reduced to different roots

```
alumnus -> alumnu
alumni  -> alumni
alumna  -> alumna
```

- For English this is often performed using Porter’s algorithm [1]

[1] Porter, 1980

Lemmatization

- Whereas stemming simply removes suffixes, lemmatization reduces inflected words to a semantically correct base word, called a *lemma*, that lies in the dictionary of valid words.

```
goose -> goose
geese -> goose
```

- Whereas stemmers are algorithmic in nature, lemmatizers require data (memory) in order to evaluate semantic relationships between words.

- Common lemmatized used the English language is the WordNet Lemmatizer.