Text normalization

- Contraction expansion

```
aren’t  -> are not
isn’t   -> is not
they’ll -> they will
```

- Punctuation & whitespace stripping

```
… had, for various reasons, went broke.  … -> had for various reasons went broke …
```

- Capitalization

```
Sara is gregarious -> sara is gregarious
```

- Stop word removal (applies to phrases, too)

Determiners: `For, an, nor, but`
Conjunctions: `the, a, an, another`
Prepositions: `in, under, towards, before`

- These normalizations are often performed using regular expressions

Stemming

[1] Porter, 1980

- 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]