Tokenization

In [1]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Xiaowei\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\wordnet.zip.
```

Out[1]:
```
True
```

In [2]:
```
text = "This is Andrew's text, isn't it?"
```

In [3]:
```
tokenizer = nltk.tokenize.WhitespaceTokenizer()
tokenizer.tokenize(text)
```

Out[3]:
```
['This', 'is', "Andrew's", 'text,', "isn't", 'it?']
```

In [4]:
```
tokenizer = nltk.tokenize.TreebankWordTokenizer()
tokenizer.tokenize(text)
```

Out[4]:
```
['This', 'is', 'Andrew', "'s", 'text', ',', 'is', "n't", 'it', '?']
```

In [5]:
```
tokenizer = nltk.tokenize.WordPunctTokenizer()
tokenizer.tokenize(text)
```

Out[5]:
```
['This', 'is', 'Andrew', "'", 's', 'text', ',', 'isn', "'", 't', 'it', '?']
```

Stemming (further in the video)

In [8]:
```
text = "feet wolves cats talked"
tokenizer = nltk.tokenize.TreebankWordTokenizer()
tokens = tokenizer.tokenize(text)
```

In [9]:
```
stemmer = nltk.stem.PorterStemmer()
" ".join(stemmer.stem(token) for token in tokens)
```

Out[9]:
```
'feet wolv cat talk'
```

In [10]:
```
stemmer = nltk.stem.WordNetLemmatizer()
" ".join(stemmer.lemmatize(token) for token in tokens)
```

Out[10]:
```
'foot wolf cat talked'
```

In [ ]:
```

```