Tokenization

Tokenization is a process that splits an input sequence into so-called tokens

- You can think of a token as a useful unit for semantic processing
- Can be a word, sentence, paragraph, etc.

An example of simple whitespace tokenizer

- nltk.tokenize.WhitespaceTokenizer

This	is	Andrew’s	text,	isn’t	it?

- Problem: “it” and “it?” are different tokens with same meaning