Tokenization

Let's try to also split by punctuation

- nltk.tokenize.WordPunctTokenizer

This	is	Andrew	'	s	text	,	isn	'	t	it	?

- Problem: "s", "isn", "t" are not very meaningful

We can come up with a set of rules

- nltk.tokenize.TreebankWordTokenizer

This	is	Andrew	's	text	,	is	n't	it	?

- "'s" and "n't" are more meaningful for processing