Preprocessing

First, I stripped out the speech marks from the text replacing them with spaces. Then I removed all apostrophes without inserting a space. This is to above my regex tokenizer splitting don’t into don t. However this has the effect of contracting some words thought shouldn’t be the prime example being we’re to were. But were, we and are are removed in my stop word removal so this does not add false information to the lexicon in this example.