Exercise 18

*[(('Sperm', 'Whale'), 119), (('Moby', 'Dick'), 84), (('old', 'man'), 80), (('White', 'Whale'), 74), (('Captain', 'Ahab'), 62), (('sperm', 'whale'), 56), (('mast', 'head'), 45), (('Right', 'Whale'), 39), (('mast', 'heads'), 36), (('ye', 'see'), 36), (('whale', 'ship'), 33), (('cried', 'Ahab'), 33), (('Captain', 'Peleg'), 32), (('white', 'whale'), 31), (('Aye', 'aye'), 31), (('Mr', 'Starbuck'), 29), (('one', 'hand'), 28), (('quarter', 'deck'), 28), (('whale', 'boat'), 26), (('?"', 'said'), 26), (('cried', 'Stubb'), 23), (('one', 'side'), 23), (('every', 'one'), 21), (('never', 'mind'), 20), (('let', 'us'), 20), (('ye', 'ye'), 19), (('chief', 'mate'), 19), (('said', 'Stubb'), 18), (('New', 'Bedford'), 18), (('years', 'ago'), 18), (('?"', 'cried'), 17), (('cried', 'Starbuck'), 17), (('old', 'Ahab'), 16), (('Cape', 'Horn'), 16), (('lower', 'jaw'), 16), (('something', 'like'), 16), (('would', 'seem'), 16), (('try', 'works'), 16), (('Look', 'ye'), 16), ((".'", '"\''), 16), (('thee', 'thou'), 16), (('round', 'round'), 16), (('ivory', 'leg'), 15), (('boat', 'crew'), 15), (('would', 'fain'), 15), (('Dough', 'Boy'), 15), (('well', 'known'), 15), (('Fast', 'Fish'), 15), (('three', 'years'), 15), (('Loose', 'Fish'), 14)]*

The task for Exercise 18 was to find the 50 most common bigrams in a text while excluding stop words and punctuation. A bigram is a pair of words that appear frequently together. To achieve this task, we had to create a list of words in which the stop words, which we got through the NLTK corpus, and punctuation, which we created our own list for, were removed. After doing so, we simply used the NLTK bigrams method and a frequency distribution to find the most common bigrams. I found that the most common bigram in text1, *Moby Dick*, was (‘Sperm’, ‘Whale’), which makes sense because the main villain in Moby Dick is a Sperm Whale.