I choose the OpenNLP library because it is a simple enough library that has a lot of features in common with Spacy. It offers tokenization, sentence segmentation, PoS tagging, chunking, parsing, and perceptron-based machine learning.

The Spacy tokenizer works by splitting the words by whitespace and then checking for exceptions as “U.K.” or “don’t” and prefixes, suffixes and infixes (ex: $, km, --)

The OpenNLP WhitespaceTokenizer uses white spaces to tokenize the input text. Both gave similar results, even though Spacy could extract better information from complex structures.

For the second task, I created similar code in both the Python and Java solution. Both use RegEx to match the words in the dictionaries and then count how many are there. Both gave identical results.

The only problem occurred when trying to extract the VPs and NPs. Both Spacy and OpenNLP have their own machine learning models that have been trained on different training data. The models and the training data are different. This resulted in big variations between the two. This considered, Spacy has the superior machine learning model and training data.