Remote programming exercise

This project should be well tested and code should be checked into GitHub along with any output produced from running your program. Write your solution in the Java programming language, using only the standard libraries. Do not use any third-party libraries for natural language processing. Describe any assumptions you make in your implementation. What are the limitations of your approach? What other approaches might be possible? At the end of the assignment there should be a commit/push for each of the following features.

1. Write a program that identifies sentence boundaries and tokenizes the text in the file "*nlp_data.txt*" into words. It should correctly process all symbols, including punctuation and whitespace. Every word must fall into a sentence. Create data structures that efficiently express the data you have processed. When your program runs it should output an XML representation of your Java object model.

2. Modify your program from #1 to add rudimentary recognition of proper nouns ("named entities") in the input, and print a list of recognized named entities when it runs. The list of named entities is in the file "*NER.txt*". Enhance your data structures and output schema to store information about which portions of the text represent named entities.

3. Modify your program from #2 to use "*nlp_data.zip*" as its input. Use a thread pool to parallelize the processing of the text files contained in the zip. Aggregate the results and modify the output schema accordingly.