1 WordCount for Named Entities

In this part, you will compute the word frequency for named entities in a large file. You are free to use any NLP library that works with Spark and Scala or PySpark. A good choice is this one: https://github.com/JohnSnowLabs/spark-nlp-workshop

The steps of the assignment would be as follows:

1. Find a large text file from the Gutenberg project: https://www.gutenberg.org and upload it to your Databricks cluster.
2. Write code for a mapreduce program in Scala/PySpark which reads in the file, and then extracts only the named entities. A good resource for this is the Spark NLP library of John Snow labs: https://nlp.johnsnowlabs.com
   You are free to use any other library also.
3. The output from the map task should be in the form of (key, Value) where key is the named entity, and value is its count (i.e. once every time it occurs)
4. The output from the reducer should be sorted in descending order of count. That is, the named entity that is most frequent should appear at the top.