MLSD: Spark Intro

March 13th, 2025

1. Setup your Spark environment. You may install Pyspark using pip, conda[1, 2], or use Google Colab.
2. Create a Spark implementation to count the occurrences of distinct words in the supplied text file “lusiadas.txt”.
3. Adapt the code to find the most common biwords, that is the most common sequences of two words. Ignore words with less than 3 letters.
4. Create a Spark application that calculates the number of unique words that start with each letter of the alphabet. The counting should be case-insensitive (convert to lowercase) and should ignore words with less than 3 letters.

[1]https://spark.apache.org
[2]https://spark.apache.org/docs/latest/quick-start.html