MLSD: Spark Intro

March 10, 2021

1. Setup your VM/Docker Spark environment. You may install Spark locally if you prefer.[1]
2. Try the simple RDD operations listed in the slides and some of the examples in Spark homepage.[2]
3. Using the Spark shell, count the occurrences of the words in the example file (lusiadas.txt).
   You can use the code available in the elearning page.
4. Run spark-submit to execute the word count code

   ```
   $ spark-submit spark_word_count.py wordcount/lusiadas.txt wordcount/result
   ```

5. 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.
6. 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/docs/latest/
[2] https://spark.apache.org/docs/latest/quick-start.html