Finnegans Wake – aggregation

1. Download the text of Finnegans Wake from https://gitlab.com/opstar/share20/-/raw/master/fwake.txt?inline=false

2. Load the entire text to HDFS.

3. Use Spark RDD functions to split the text and:

   a. Count the total number of words
   b. Count the number of distinct words (total vocabulary)

4. Find the 10 longest words (there are some very long words in the text!)
   hint: *w.sortBy(lambda x: len(x), False).take(10)*

5. Retrieve a list of all the words that occur more than 2000 times and how many times they each occur. Expected result:

```
[('the', 11250), ('and', 7366), ('of', 6975), ('to', 4293), ('a', 4236),
('in', 3411), ('his', 2853), ('for', 2190)]
```