# WordCount mit Spark

Der enthaltene Spark Job ermöglicht ein WordCount ähnlich dem MapReduce Beispiel – allerdings in einem verfeinerten Zustand.

Als neues Feature wird neben dem standardmäßigen Wordcount auch eine gefilterte Variante angeboten, die nur Wörter herausschreibt, welche eine bestimmte Anzahl gezählt wurden. Diese Anzahl kann als drittes Argument neben der Inputdatei (Argument 1) und der Outputdatei (Argument 2) angegeben werden. Die gefilterte Version wird in einer separaten Outputdatei mit dem Präfix „Filtered“ ausgegeben.

Als weiteres Feature werden beide Listen nach Aufkommen der Wörter sortiert.

Der Job kann wie folgt ausgeführt werden:

*/usr/lib/spark/bin/spark-submit --class SparkWordCount \* *bigdata-sparkwordcount-1.0.0-SNAPSHOT.jar \pg14591.txt WordCountResult.txt 5*