Übung Big-Data-Technologien
Prof. Dr. Christoph Schmitz

Übungsblatt 5

Hinweise

Sie finden Beispielcode für Spark in Java, Python und Scala in Stud.IP in `bdt-uebung-5-spark.zip`. Darüberhinaus bietet http://spark.apache.org/examples.html etliche Beispiele in allen drei Sprachen.

Sie können Ihre Programme auf `infbdtXYZ` starten mit

```
spark-submit --master (local|yarn) mein-programm-jar-with-dependencies.jar <argumente>
spark-submit --master (local|yarn) mein-programm.py <argumente>
```

Scala ist komplizierter – mehr dazu in der Übung.
Wir verwenden Spark 1.6.2 – achten Sie darauf, wenn Sie z. B. API-Dokumentation lesen.

1 Spark: RDDs

Schreiben Sie mit Spark ein Programm, das die *unterschiedlichen* Wörter in einer Textsammlung zählt, die:

- mit „p“ beginnen
- mit „p“ enden
- mit „p“ beginnen *und* enden (also die Schnittmenge der vorgenannten)

„Unterschiedlich“ heißt, dass mehrfache Vorkommen des gleichen Wortes nur als ein Wort gezählt werden sollen.

Berechnen Sie alle drei Zahlen in *einem* Datenfluss, und vermeiden Sie unnötige Durchläufe über die Daten. Geben Sie die drei Anzahlen sowie die Wörter der dritten Menge aus.

Auf den Shakespeare-Beispieldaten sollte das Resultat sein:

- 2017 Wörter mit „p“ am Anfang, 181 mit „p“ am Ende
- 11 mit „p“ am Anfang und Ende: phillip, peep, prop, poop, pap, pump, protectorship, pomp, philip, plump, pip

2 Spark: Top-k

Schreiben Sie mit Spark ein Programm, das die k häufigsten Wörter aus einer Textsammlung berechnet und mit ihren Häufigkeiten ausgibt. Dabei sollen Stoppwörter (aus der bekannten Liste) nicht mitgezählt werden.

Erproben Sie Ihr Programm wieder am Shakespeare-Textkorpus. Das Ergebnis für k = 5 sollte sein: thy (3807), lord (3271), thee (3108), sir (2982), king (2964)

Hinweis: Sie können die Stoppwort-Liste über eine Broadcast-Variable verteilen oder als RDD über einen Outer Join mit den Wörtern aus dem Text verknüpfen. Die Variante mit der Broadcast-Variablen ist deutlich einfacher.