

MapReduce w Sparku

16 grudnia 2017

Opis pliku z zadaniami

Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu.
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).

1 Spark



Treść

Na zajęciach będziemy wykorzystywać Apache Spark, w którym będziemy stosować obliczenia oparte na paradygmacie MapReduce.

Do rozpoczęcia pracy należy wykonać następujące kroki:

1. Wszelkie informacje o Apache Spark:
`https://spark.apache.org/docs/latest/index.html`
2. Należy pobrać paczkę instalacyjną ze strony:
`http://spark.apache.org/downloads.html`
3. Pobraną paczkę należy rozpakować:
`tar xvfz spark-2.2.0-bin-hadoop2.7.tar`
4. Pracę z systemem można rozpocząć poprzez konsolę:
`./bin/spark-shell`

2 Proste zadania w Sparku

15p.◇

Treść

W tym zadaniu należy wykonać następujące zadania:

- **WordCount**: należy stworzyć plik zawierający listę słów oraz liczbę ich wystąpień w dziełach Szekspira (odpowiedni plik został załączony na stronie przedmiotu),
- **MatrixVectorMultiplication**: należy policzyć iloczyn macierzy i wektora (odpowiednie pliki zostały dołączone do strony przedmiotu),
- **ApproximatePI**: należy napisać program, który przybliży liczbę π .

W celu wykonania tych zadań skorzystaj z następujących poleceń:

- `textFile`,
- `FlatMap`,
- `map`,
- `reduceByKey`,
- `sortBy(_._2)`,
- `saveAsTextFile`,
- `take`,
- `collect`,
- `broadcast`,
- `toDF`,
- `orderBy`,
- `show`,
- `parallelize`,
- `1 to 10`,
- `math.random`,
- `if () else`.

3 Selekcja i grupowanie

10p.◇

Treść

W tym zadaniu należy wykonać następujące zadania:

- **Selection:** należy wybrać i wyświetlić wszystkie utwory zespołu *Queen*,
- **Group:** należy wypisać wszystkich wykonawców z liczbą utworów większą od 150.

W celu wykonania tych zadań skorzystaj z następujących poleceń:

- `read.csv`,
- `rdd`,
- `Some`,
- `None`,
- `r(1)`,
- `foreach`,
- `println`,
- `count`,
- `reduceByKey`,
- `r._1`,
- `sortBy`.