# PageRank in PySpark with HDFS

Experiments with default partitioning , external partitioning and persistence of RDDs while implementing PageRank using PySpark on a Hadoop cluster.

## Dataset

[Berkeley-Stanford web graph](https://snap.stanford.edu/data/web-BerkStan.html) was used in this project.

## Scripts

page\_rank.py : Spark application that implements the PageRank algorithm with default partitions, external partitions and persistence.

## Usage

**Page rank with default partition:**

**Please change the method name in sparkContext method to pageRank(spark, input\_file, iterations, partitions) at line 103**

>> spark-submit --verbose page\_rank.py -i web-BerkStan.txt -p <partitions> -itr <iterations>

**Page rank with external partition:**

**Please change the method name in sparkContext method to pageRank\_Partitions(spark, input\_file, iterations, partitions) at line 103**

>> spark-submit --verbose page\_rank.py -i web-BerkStan.txt -p <partitions> -itr <iterations>

**Page rank with external partition and persistence:**

**Please change the method name in sparkContext method to pageRank\_Cache(spark, input\_file, iterations, partitions) at line 103**

>> spark-submit --verbose page\_rank.py -i web-BerkStan.txt -p <partitions> -itr <iterations>