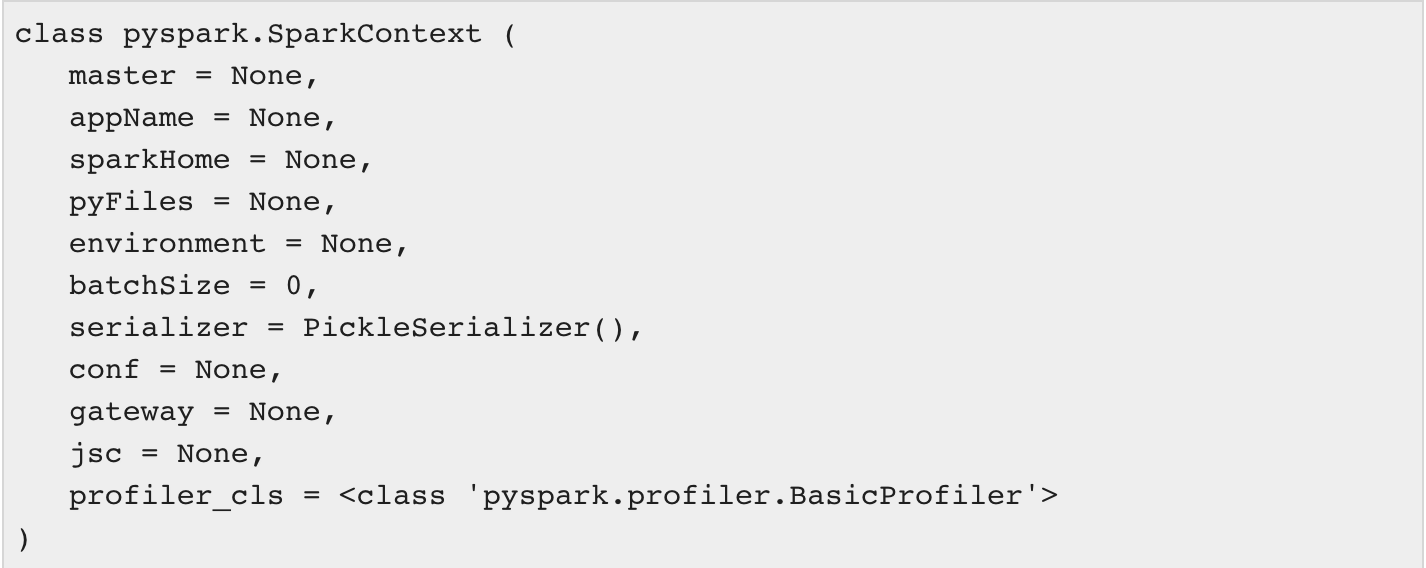
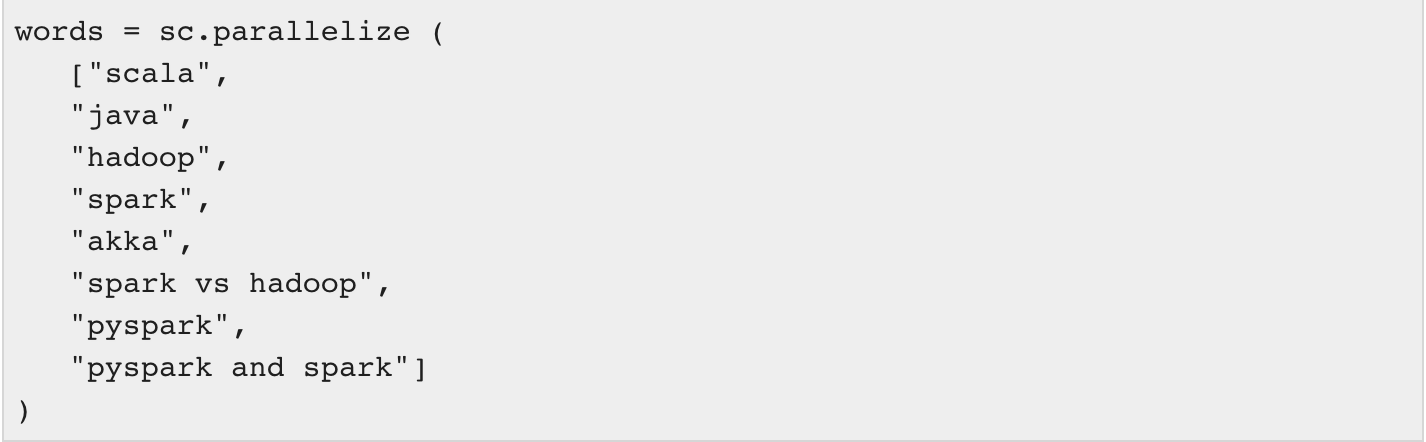
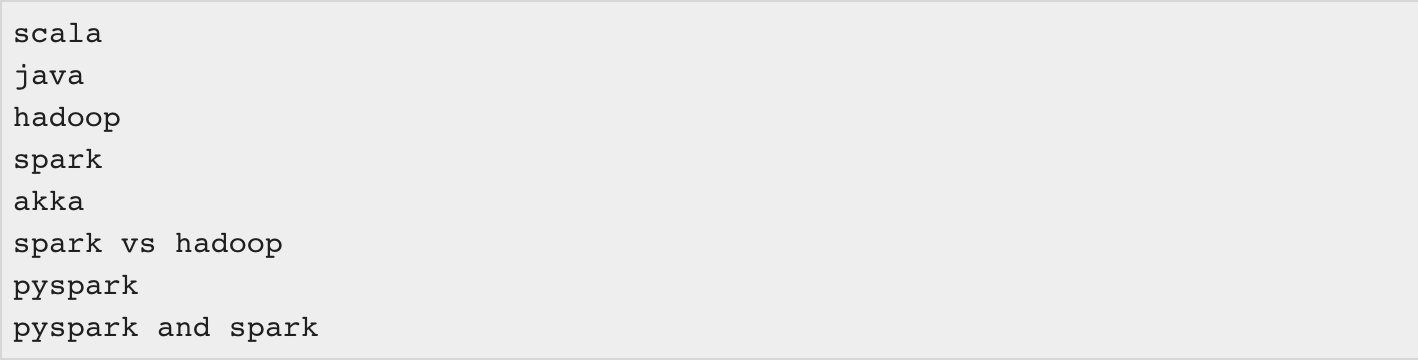
Tìm Hiểu PySpark

Spark Context

* **Master** – Tên URL kết nối.
* **appName** – Tên ứng dụng.
* **sparkHome** – Nơi lưu trữ spark.
* **pyFiles** − File .zip hoặc .py gửi đến cụm kết nối và được thêm vào PYTHONPATH.
* **Environment** – Các nút hoặc các biến môi trường
* **batchSize** – Số lươgnj đối tượng Python được biểu diễn dưới dạng 1 đối tượng Java duy nhất. Ghi 1 để tắt, 0 để tự động lựa chọn kích thước dựa trên kích thước đối tượng hoặc để giá trị -1 để sử dụng kích thước không giới hạn.
* **Serializer** – Tuần tự hoá RDD.
* **Conf** – đối tượng Spark Conf để cài đặt các thuộc tính của Spark.
* **Gateway** – Sử dụng gateway đã tồn tại và JVM hoặc tạo mới JVM.
* **JSC** − The JavaSparkContext instance.
* **profiler\_cls** – Lớp thông tin tuỳ chỉnh để tạo thông tin cá nhân (giá trị mặc định là pyspark.profiler.BasicProfiler).

Spark RDD (Resilient Distributed Dataset)

* Resilient Distributed Dataset là tập tài liệu phân phối có khả năng phục hồi
* Câu lệnh trên tạo ra 1 mảng các string và được chia thành các phần bằng nhau dựa vào khai báo của Spark Context
* Kết quả sau khi chạy lệnh foreach
* Collect ()
* Sau khi chạy lệnh collect()

