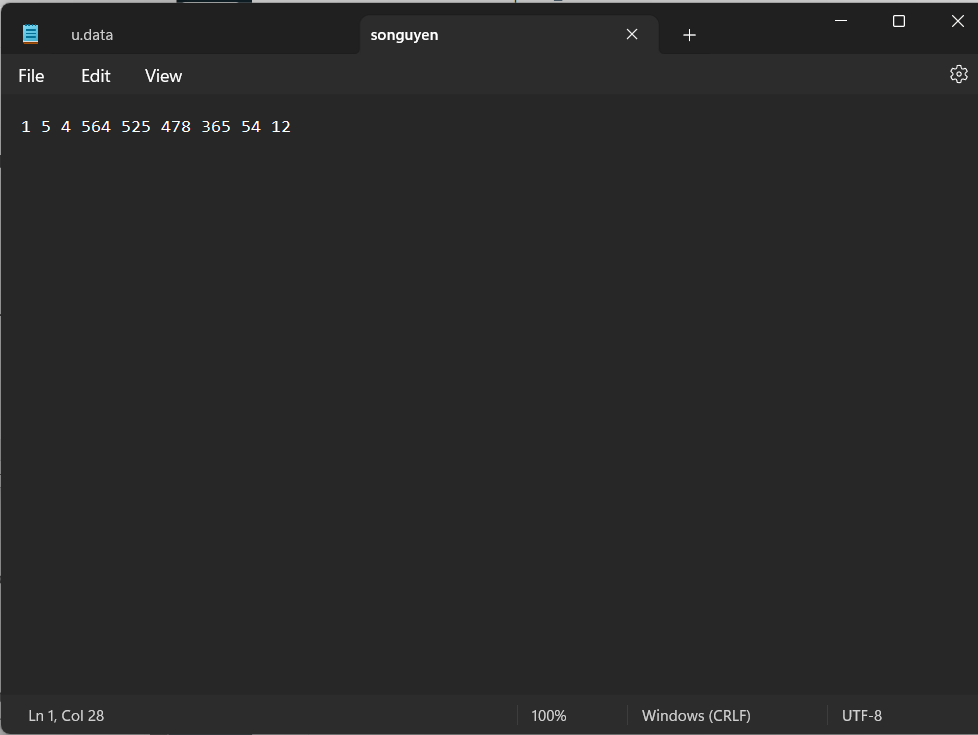
**Bài thực hành số 03. Lập trình MapReduce trên Spark**

Bài 1. Tìm số lớn nhất

Cho tệp songuyen.txt chứa các số nguyên, giữa các số cách nhau bởi khoảng

trống. Viết chương trình tìm số lớn nhất trong tệp.



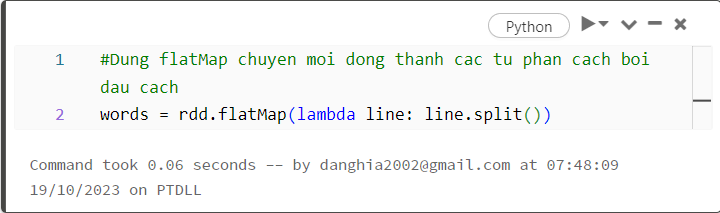
- Đọc file



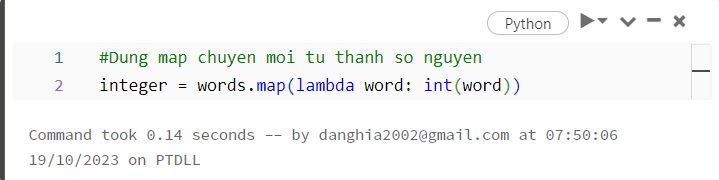
- Đọc tệp thành RDD



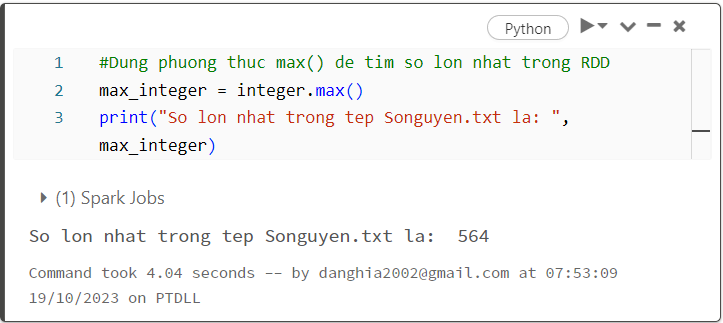
- Dùng flatMap chuyển mỗi dòng thành các từ phân cách bởi dấu cách



- Dùng map chuyển mỗi từ thành số nguyên

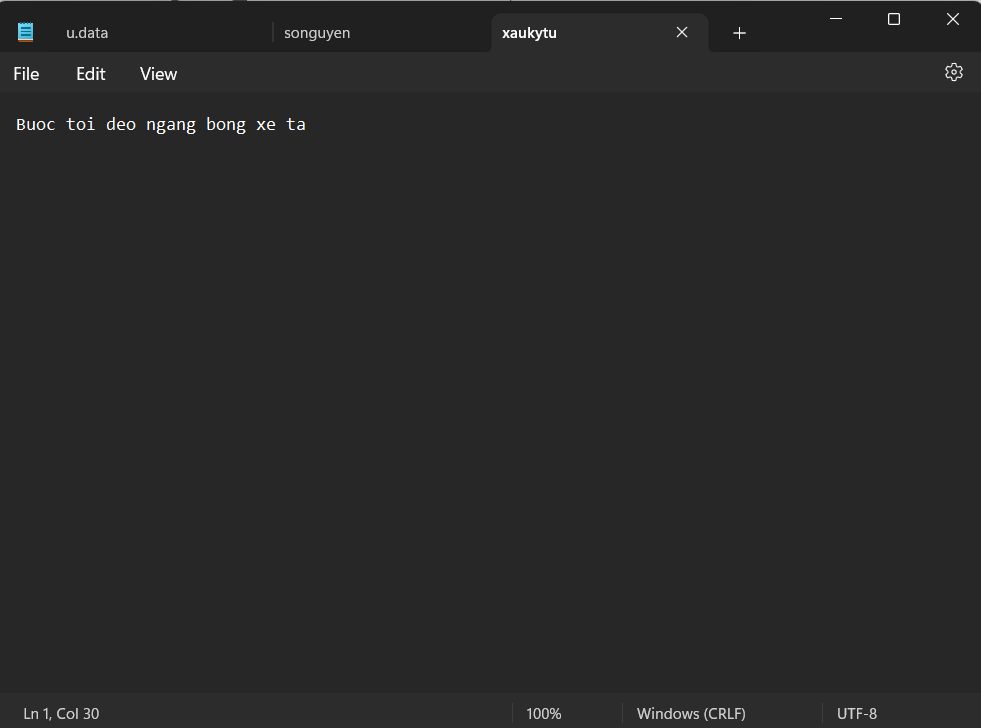


- Dùng phương thức max() để tìm số lớn nhất trong RDD.



Bài 2. Đếm số lấn xuất hiện một xâu

Viết chương trình đếm số lần xuất hiện của xâu x trong một tệp.



- Đọc file

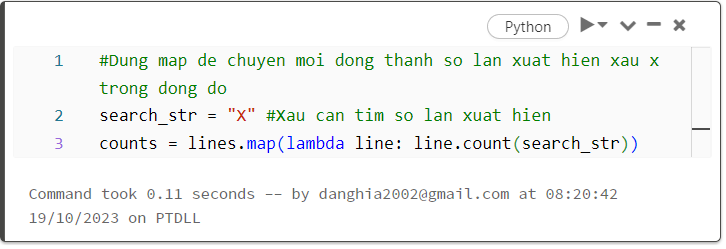


- Đọc tệp vào RDD



- Dùng map chuyển mỗi dòng thành số lần xuất hiện xâu x trong dòng đó (dùng

str.count(substr))



- Dùng reduce cộng các số trong RDD lại

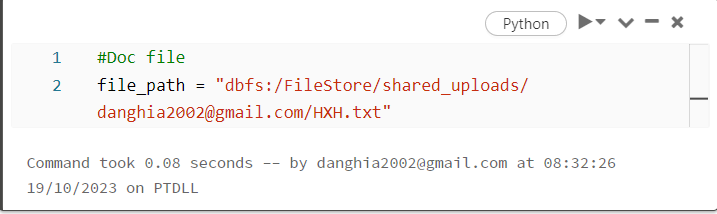


Bài 3. Bigram

Cho trước một tệp văn bản HXH.txt, hãy thống kê số lần xuất hiện từng cặp từ

trong văn bản, sắp thứ tự giảm theo số lần xuất hiện rồi ghi ra thư mục HXH.

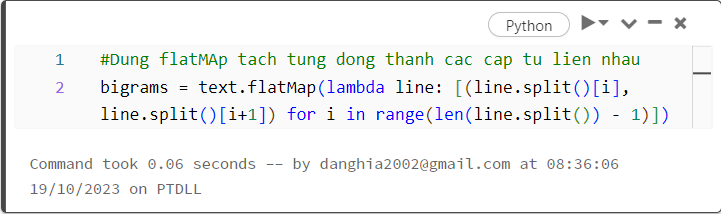
- Đọc file



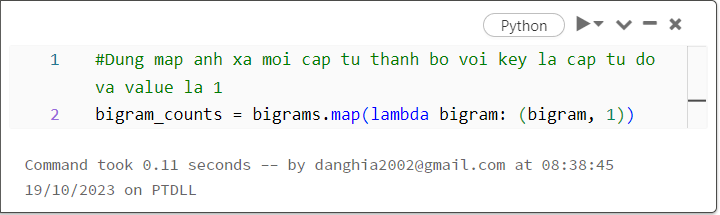
- Đọc tệp vào RDD



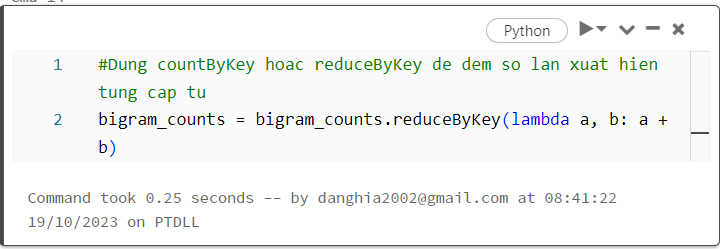
- Dùng flatMap tách từng dòng thành các cặp từ liền nhau



- Dùng map ánh xạ mỗi cặp từ thành bộ với key là cặp từ đó và value là 1



- Dùng countByKey hoặc reduceByKey để đếm số lần xuất hiện từng cặp từ



- Ghi kết quả ra thư mục HXH

