Examination Code 05

Input: Cho một tập hợp các tài liệu (gọi là ngữ liệu). Mỗi tài liệu được lưu trữ trong một tập tin văn bản nhiều dòng. Nội dung tài liệu chỉ chứa các chữ cái viết thường, khoảng trắng và ký tự dòng mới. Mã tài liệu được thể hiện ở dòng đầu tiên.

Sample input

Doc ID	doc01	doc02	doc03
Content	id01 full of water full of sweet	id02 dessert of sweet treat to eat	id03 treat to enjoy treat to sweet

Main requirement: Đối với mỗi 2-gram (cụm 2 từ liên tiếp nhau) trong ngữ liệu, hãy đếm số lần xuất hiện của nó trong mỗi tài liệu và xác định tài liệu mà 2-gram này có số lần xuất hiện cao nhất.

- Nếu số lần xuất hiện trong hai tài liệu bằng nhau, chọn tài liệu nào cũng được.

Technical requirement:

- Các định dạng đầu vào/đầu ra của Hadoop phải khớp với dữ liệu một cách hiệu quả.
- Lượng dữ liệu chuyển từ giai đoạn map sang giai đoạn redue phải được tối ưu hóa ở mức cao nhất có thể.

Output: Các bộ giá trị (*2-gram*, *docid*), trong đó *2-gram* là cặp hai từ liên tiếp nhau trong ngữ liệu, và *docid* là mã của tài liệu mà *2-gram* xuất hiện nhiều lần nhất. Không quan trọng thứ tự của các bộ giá trị.

Sample output

```
full of id01
of water id01
of sweet id01
dessert of id02
treat to id03
to eat id02
to enjoy id03
to sweet id03
```