Game of Thrones

1 Context

Aemon the Castle Black’s maester is a close advisor of Jeor Mormont, the Lord Commander. He is also the Castle Black’s library director. This library has hundreds of thousands of books and some of them are so rare you can not even find them in the Citadel.

Aemon always had this puzzle in his mind to find out those words that get commonly used across the books available at the library. He decided to ask Sam, his favorite trainee for some help to find a way to easily find that for him. Then Sam asked his friend Jon Snow who always liked information search and data engineering.

2 Problem statement

We want you to write a map reduce program (in any programming language that you are comfortable with) to calculate the Nth frequently occurring words across the input files.

Test case:

Input

File 0
How much wood would a woodchuck chuck If a woodchuck could chuck wood

File 1
He would chuck, he would, as much as he could, And chuck as much as a woodchuck would If a woodchuck could chuck wood

Nth frequently occurring words: 3

Output
much, wood, could, he

We want a solution that works on a massive dataset. Our algorithm has to be able to run on a distributed system so we are not limited by the amount of memory or CPU of a single machine

You will find the dataset of documents for which your program has to work in the dataset folder of the repository.