Big Data Programming Assignment3

**WordCount**

2017101319

국제학과 한동훈

**What I’ve learned**

>> WordCount는 input을 단어 별로 구분해서 그 단어의 개수를 세어야 한다. Map function에서는 input을 단어 별로 구분하기 위해서 StringTokenizer을 사용한다. Input을 모두 tokenize 한 후에 이를 반복문을 돌면서 (단어, 1)을 intermediate pair를 만든다. Reduce function에서는 이렇게 구한 intermediate pair를 반복문을 통해 돌면서 단어가 나올 때마다 sum에 1을 더해준다.

덧붙이자면 해당 Wordcount 파일은 띄어쓰기를 단위로 tokenize 과정을 거치기 때문에, “(Amendment”나 ”(Amendments” 같은 것들은 서로 다른 토큰으로 인식하며, “(” 같은 기호을 구분하지 못한다는 특징을 가진다.

**Input & Output**

>>Input은 지난 번 wordcount와 동일하게 미국 헌법이다. 해당 input을 단어 별로 나누어서 단어의 개수를 셀 것이다.

스크린샷, 컴퓨터, 노트북, 모니터이(가) 표시된 사진

자동 생성된 설명

<사진1. US constitution(input)>

스크린샷이(가) 표시된 사진

자동 생성된 설명

<사진2. 실행 화면>

컴퓨터, 모니터, 노트북, 스크린샷이(가) 표시된 사진

자동 생성된 설명

<사진3. Output>

**Command Line**

>> bin/hadoop com.sun.tools.javac.Main WordCount.java //컴파일

>> jar cf wc.jar WordCount\*.class. //Jar file 만들기

>> bin/hadoop jar wc.jar WordCount user/dhk1349/wordcount/input user/dhk1349/wordcount/output //jar 파일을 실행