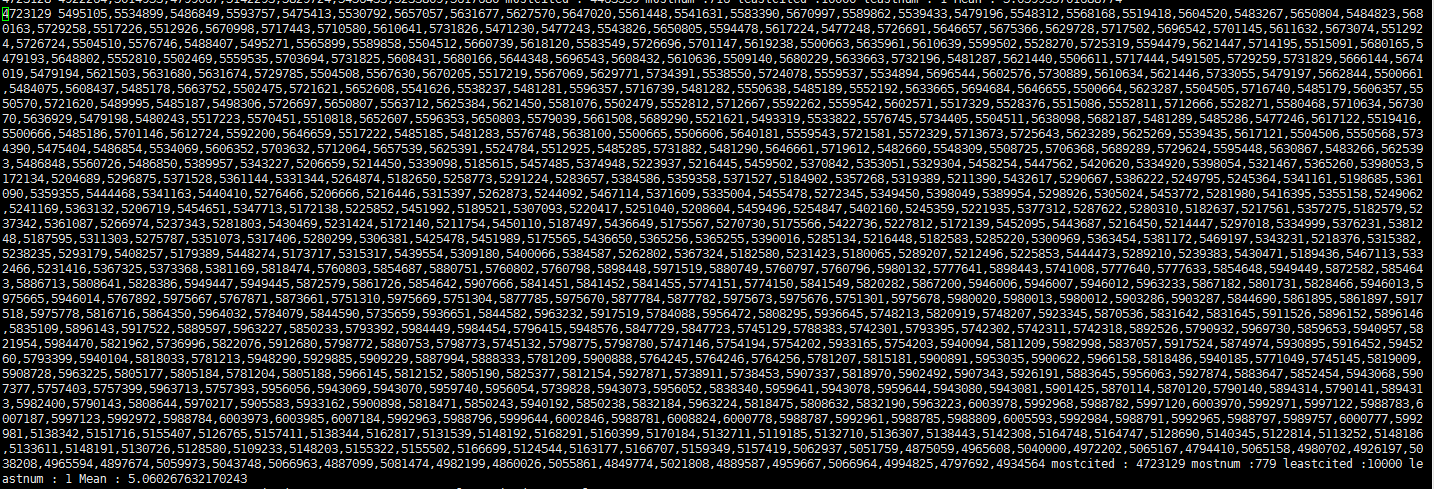
Data Science Assignment#1

리눅스의 프로세스 자동 관리 프로그램 추가 기능 구현

학부 : 소프트웨어 학부

학번 : 20150291

이름 : 하현수

최대 CITING 된 특허와 그 수  
  
맨 밑줄 조회 결과, most cited는 4723129번이고 인용수는 779라는 것을 알 수 있었습니다.  
  
vi 검색 기능으로 조회 결과 4723129번 특허가 실제 인용 횟수 779번으로 가장 많았다는 것을 알 수 있었습니다.

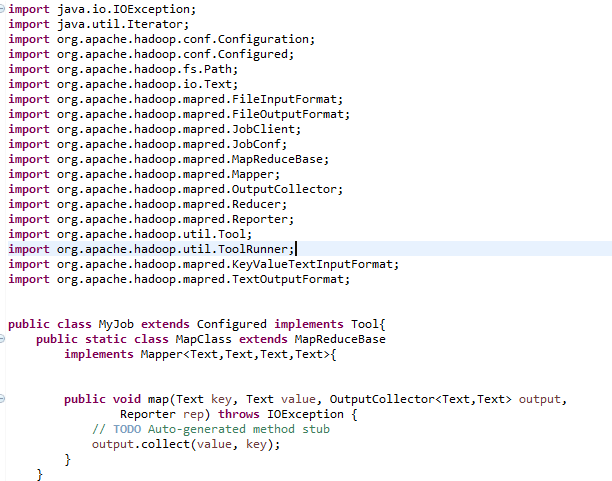
(실제 Xshell에 있는 제 output파일 part-00000에서 4723129를 조회한 결과입니다.)  
사용 하둡 버전 2.7.4  
citing과 cited가 세지는 문제가 있어 실행 시에는 맨 윗줄 citing cited를 제거하였습니다.

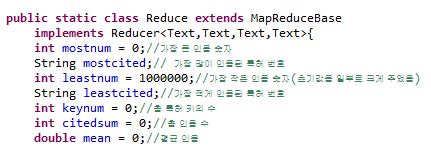
최소 인용 특허와 그 수

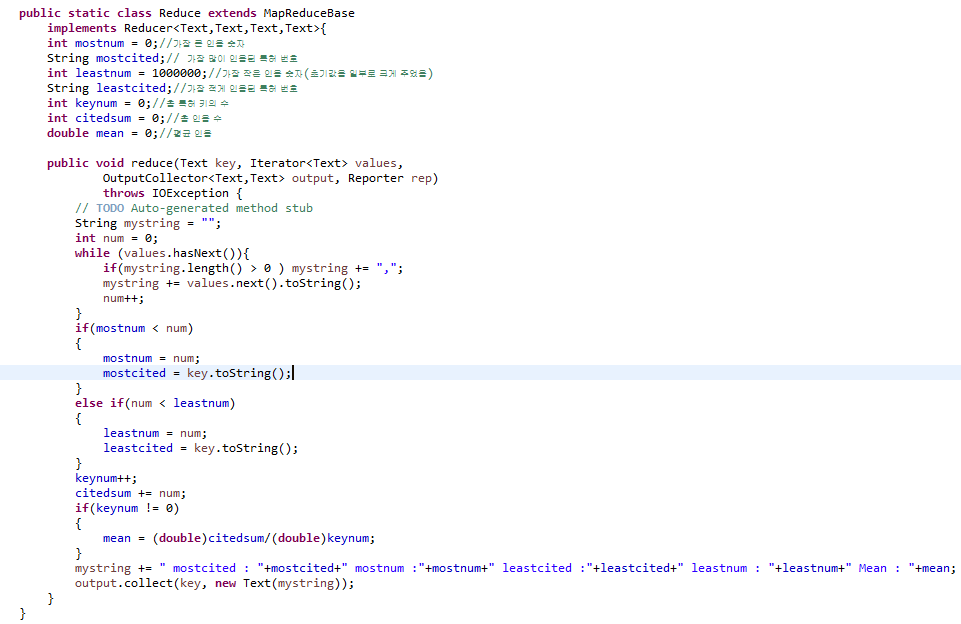
  
마지막줄에서 leastcited가 10000이라는 것이 확인 되었고, 가장 적은 인용수는 1이라는 것을 알 수 있었습니다.  
가장 적은 인용수인 1을 가지는 특허는 많았지만, 최초에 나온 키값이 등록되어 leastcited는 10000으로 등록되었습니다.  
  
4539112 특허만 리스트에 있는 10000번 특허의 모습입니다.

평균값  
  
최종값의 Mean값이 5.06981(소수5째자리까지)라고 표시하고 있습니다.

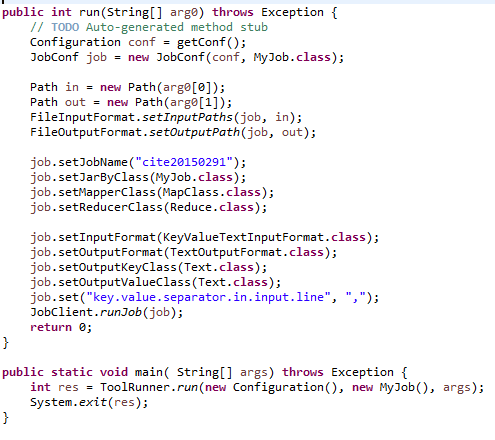
구현방식

  
메이븐을 이용하여 하둡에 필요한 코어 헤더파일을 추가하고, Myjob클래스를 선언하였습니다.  
그후 MapClass를 Mapper와 MapReduceBase를 이중상속받아 구현하였고,  
정상적인 map함수 작동을 위하여 Mapper<Text,Text,Text,Text>의 제네릭 형식을 지정해 주었습니다.(key와 value와 outputcollector의 인자 모두 Text이기 때문)

  
Reduce 클래스로 마찬가지로 MapReduceBase와 Reducer를 이중 상속합니다.  
또한 정상적인 map함수 작동을 위하여 제너릭 형식을 <Text,Text,Text,Text>로 지정합니다.  
(마찬가지로 key와 Iterator, OutputCollector의 타입이 모두 Text>이기 때문)  
그 후, 과제에서 요구하는 최대/최소 인용 숫자와 특허 번호, 그리고 평균 인용 수를 구하기 위한  
변수들을 선언합니다.  
mostnum은 가장 큰 인용 숫자, mostcited는 가장 많이 인용된 특허 번호를 뜻합니다.  
leastnum은 가장 작은 인용 숫자로 처음에 일부로 큰 값을 주어 최초 if문이 실행되도록 하였습니다. Leastcited는 가장 적게 인용된 특허 번호를 뜻합니다.  
keynum은 총 특허 키의 개수, citedsum은 총 인용 수가 대입됩니다.  
이를 통해 citedsum/keynum으로 평균값을 구할 수 있습니다.  
그 평균값은 mean에 기록됩니다.



Reduce함수입니다. 처음에 num을 0으로 두고 리듀스의 while문을 돌립니다. While문은 인용된 특허목록을 쉼표와 함께 문자열에 붙여주는 역할을 합니다. 즉 해당 key의 특허 인용 리스트를 쉼쉼표를 통해 이어 붙이는 역할을 합니다. 또한 돌 때 마다 num++을 이용하여 num값을 늘려줍니다. 이를 통해 해당key값이 얼마나 인용되었는지 num을 통해 세는 것이 가능해집니다. 이를 통해 최대,최소 특허 인용 횟수를 셀 수 있습니다.  
만약 mostnum < num 이라면, 현재 key에 해당하는 인용수가 가장 큰 것이므로, mostnum에 인용 횟수인 num을 대입하고, mostcited는 key값으로 합니다.(굳이 타입을 string으로 해서 key.tostring()을 사용한 이유는 원인은 정확히 알 수 없으나, Text로 선언된 key값을 직접 넣을 시 소팅이 제대로 되지 않거나, leastcited에 쓰레기 값이 들어가는 등의 문제가 발생하여 String으로 변경하였더니 정상 작동 하였습니다.)  
또한 만약 num < leastnum 이라면 인용횟수가 가장 작은 것이므로 leastnum에 해당 인용 횟수인 num을 대입합니다. 이때 leastnum의 최초 초기화값은 1000000이므로 최초 if문 진입시 무조건 첫번째 특허정보의 인용 횟수가 leastnum으로 삽입되도록 합니다.  
(leastnum의 초기화가 0으로 될 경우 0이 출력될 수 있다. 0이 아니더라도 적당히 작은 수는 잘못된 계산을 야기할 수 있기 때문)  
leastnum에 num을 대입한후에는 leastcited에 현재 특허의 key값을 넣어 줍니다.  
그리고 reduce함수가 호출될 때 마다 keynum++을 이용하여 총 reduce함수가 호출된 수, 즉  
key의 개수를 셀 수 있습니다.  
그후 citedsum이라는 변수에 += num을 해줌으로써, 총 인용횟수를 셀 수 있습니다.  
그 후 mean에 citedsum/keynum 값을 대입해주어 평균을 구합니다. (if문은 divide by zero에 대한 예외처리입니다.)  
그런 다음 출력하게 될 문자열 mystring에 각 변수의 값 mostcited, mostnum, leastcited, leastnum, mean을 출력시키도록 하였습니다.

  
run에서는 Job에 제 학번의 이름을 삽입하였고, 입출력 값을 지정해 주었습니다.  
또한 key value를 ,로 나눈다는 것을 선언해주었습니다.  
main에서는 맵리듀스를 실행시킵니다.