최종 발표 자료

Word Counting Distributed System

팀명 : 1석 4조

발표자: 윤세준

프로젝트 최종 발표 2022.12.12

팀원 설명

윤세준
- 프로그램 설계
- 클라이언트 side 프로그램 작성
- 테스트 환경 및 개발환경 구축

김태연
- 필터링 코드작성
- 발표자료 작성

이민지
- 웹 UI
- 결과값 출력코드 작성

이찬영
- 크롤링
- 보고서 작성
- DB관리 및 쿼리문 작성

임정규
- 서버 side 프로그램 작성
- 워드카운팅 코드 작성

문종성
- 분산저장기능 코드 작성
- 프로그램 테스트

목 차

Contents

1 프로그램 설명

1. 프로그램 설명
2. 프로그램 설계
3. 주요 설계 사항

프로그램 설명

- 대용량 파일 워드카운팅 하는 프로그램
- 여러 장비로 나누어 처리하는 것이 더 빠르다는 것을 보이기 위한 목적으로 시스템을 구상
- 크롤링으로 모은 기사 데이터를 DB에 모아 자바 RMI를 활용한 분산 처리로 필터링 및 카운팅
- 서버의 개수 조절 가능 (4개 이상 그 이하도 가능)

프로그램 설명

기본 기능

분산 저장 가능

데이터가 client에 저장
client에 연결된 n개의 server에
데이터를 분할 저장

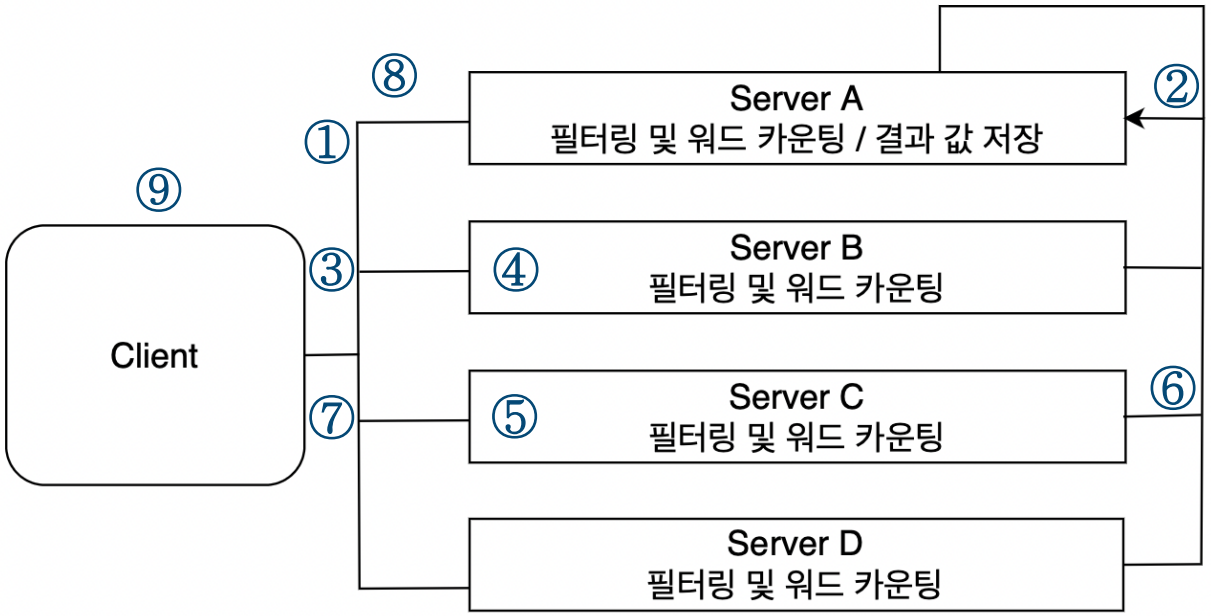
필터링

워드 카운팅 전 필터링
태그, 특수문자 제거
대소문자 획일화
의미를 갖지 않는 단어 제거
형태소분석 오픈라이브러리 사용

워드 카운팅

배열 생성
Key값을 문자로
Key에 해당하는 내용을 횟수
DB저장

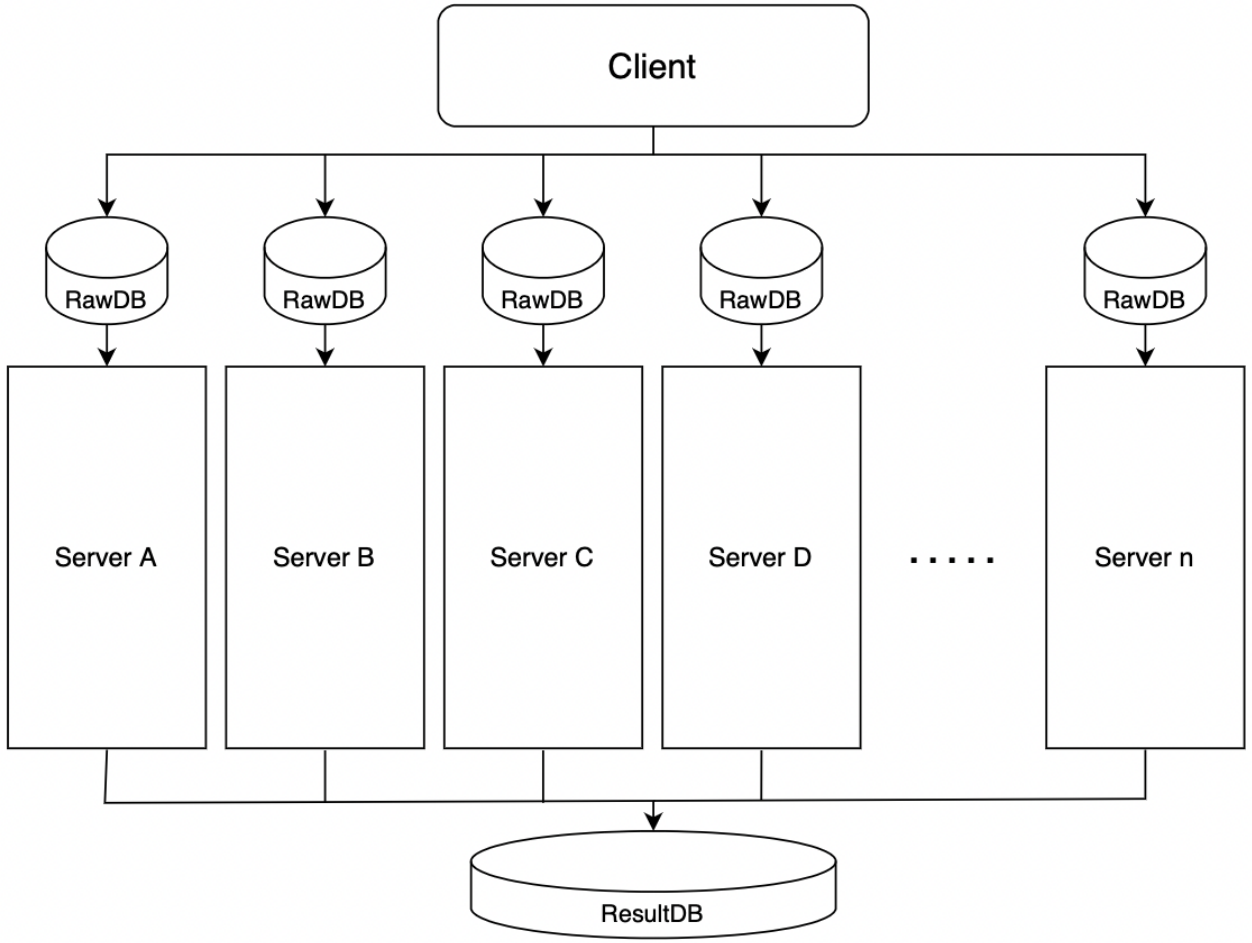
프로그램 설계



- Client : 서버에 연결, 데이터를 분산 저장, 필터링 및 워드카운팅 명령을 내리는 디바이스
- Server : DB에 저장된 데이터로 필터링, 워드카운팅 결과는 서버 A로 보냄

주요 설계 사항

주요 설계 사항 1 : 서버개수 제한이 없는 프로그램



미리 준비된 서버정보를 토대로 분산처리가 가능한 프로그램

1) 서버정보를 불러와 그 개수만큼 컨트롤러를 생성하고

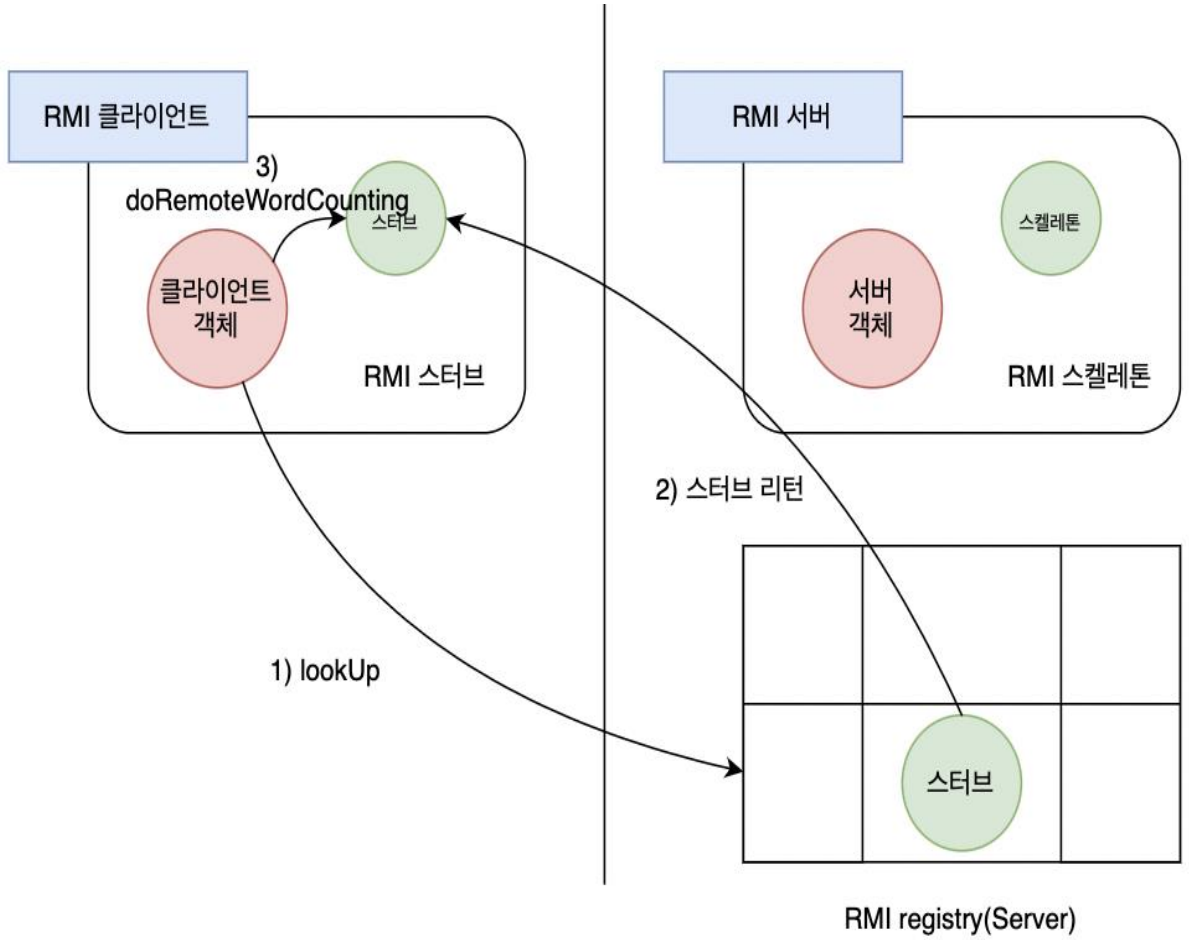
2) 각 컨트롤러의 conn이 null이 아니면 DB상태가 양호하다고 판단

3) DB DELETE를 수행하고 DB AUTO_INCREMENT를 1로 초기화

4) rawData 전체 라인 수를 읽어 txt파일에서 한 줄씩 읽으며 insert 수행

주요 설계 사항

주요 설계 사항 2 : 원격 메소드 호출

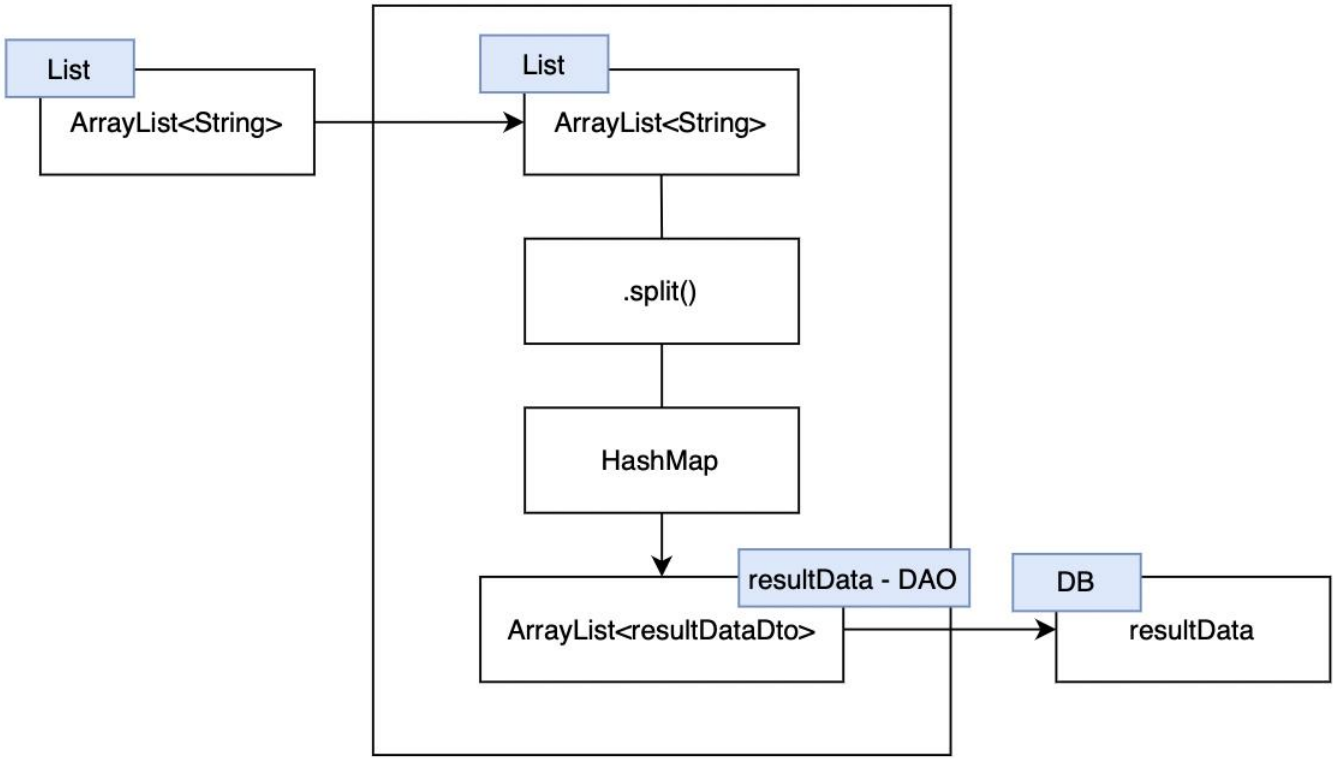


자바 RMI

- 원격 메소드 호출
- RMI를 통하여 클라이언트가 서버에게 워드카운팅을 하라는 명령을 전달
- 명령을 전달받은 서버는 워드카운팅을 수행한 후 결과값을 DB에 저장

주요 설계 사항

주요 설계 사항 3 : 워드 카운팅 분산 처리



스레드를 활용하여 워드카운팅을 분산하여 진행

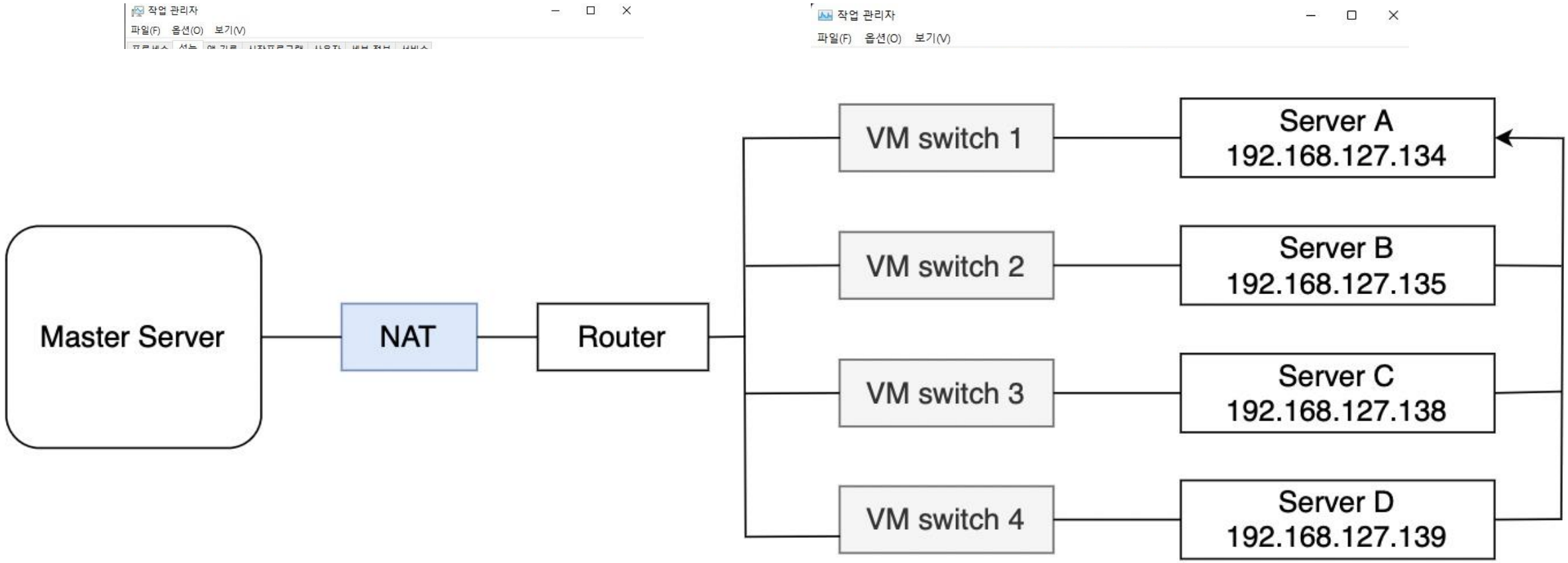
1. RMI를 통해 워드카운팅 명령을 내리더라도 서버의 수행이 끝날 때까지 대기하다가 다음 서버가 연산을 처리하게 됨.

2. 이를 해결하기 위해 서버개수만큼 스레드를 생성하여 모든 서버가 동시에 연산을 수행 할 수 있도록 프로그램 설계

2 실습 환경 구성

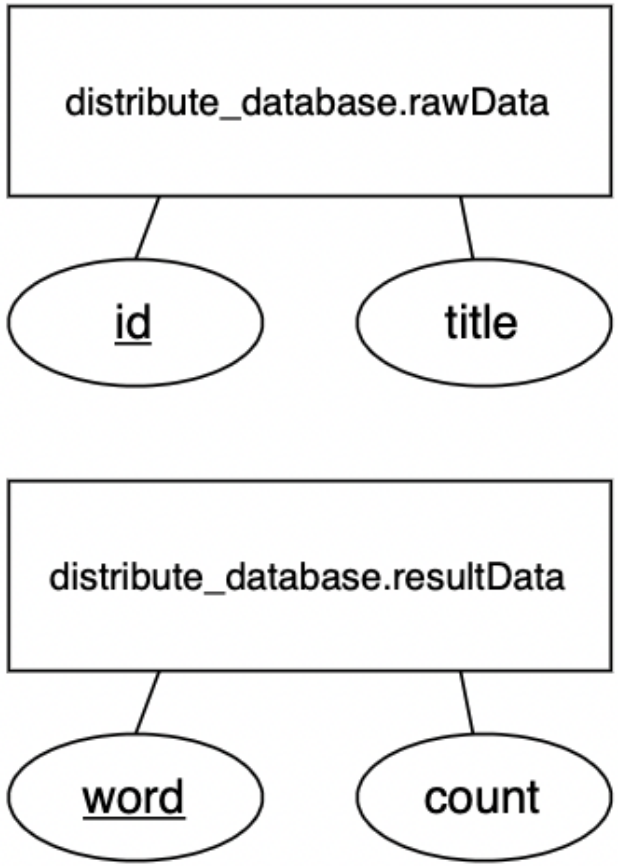
1. VM 구성
2. 데이터 구조

VM 구성



- 모든 가상머신이 자원을 100퍼센트를 사용하더라도 성능이 저하되는 문제 X
- 즉, 실험변수를 적절하게 통제했음을 알 수 있음.

데이터 구조

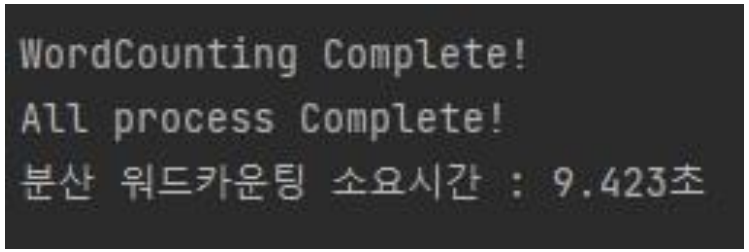


- rawData 속성으로 key값 id와 title
- resultData 속성으로 key값 word와 count
- 1500개의 백악관 성명발표문을 사용

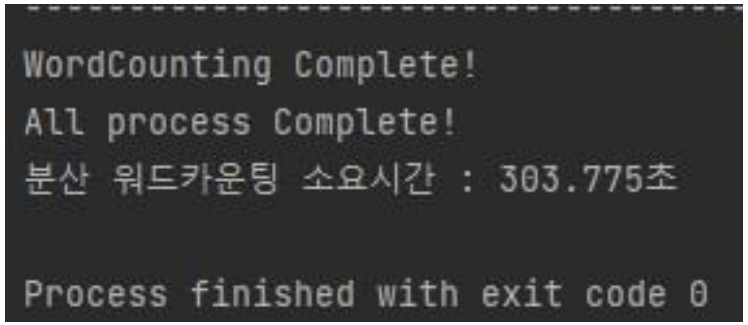
3 실행 결과

1. VM 환경 테스트
2. 인터넷 환경 테스트

VM 환경 테스트

	nonDistribute	Distribute
적은 데이터 량	 워드 카운팅 소요시간 : 9.423초	 워드 카운팅 소요시간 : 29.191초

- 분산 처리할 때 더 많은 시간 소요
- 데이터분산 저장 / 원격 메소드 호출 등으로 프로그램 동작 시간이 늘어남

많은 데이터 량	 워드 카운팅 소요시간 : 303.775초	 워드 카운팅 소요시간 : 181.324초

- 분산처리를 했을 때 분산처리를 하지 않았을 때보다 동작 시간이 줄어듦

인터넷 환경 테스트

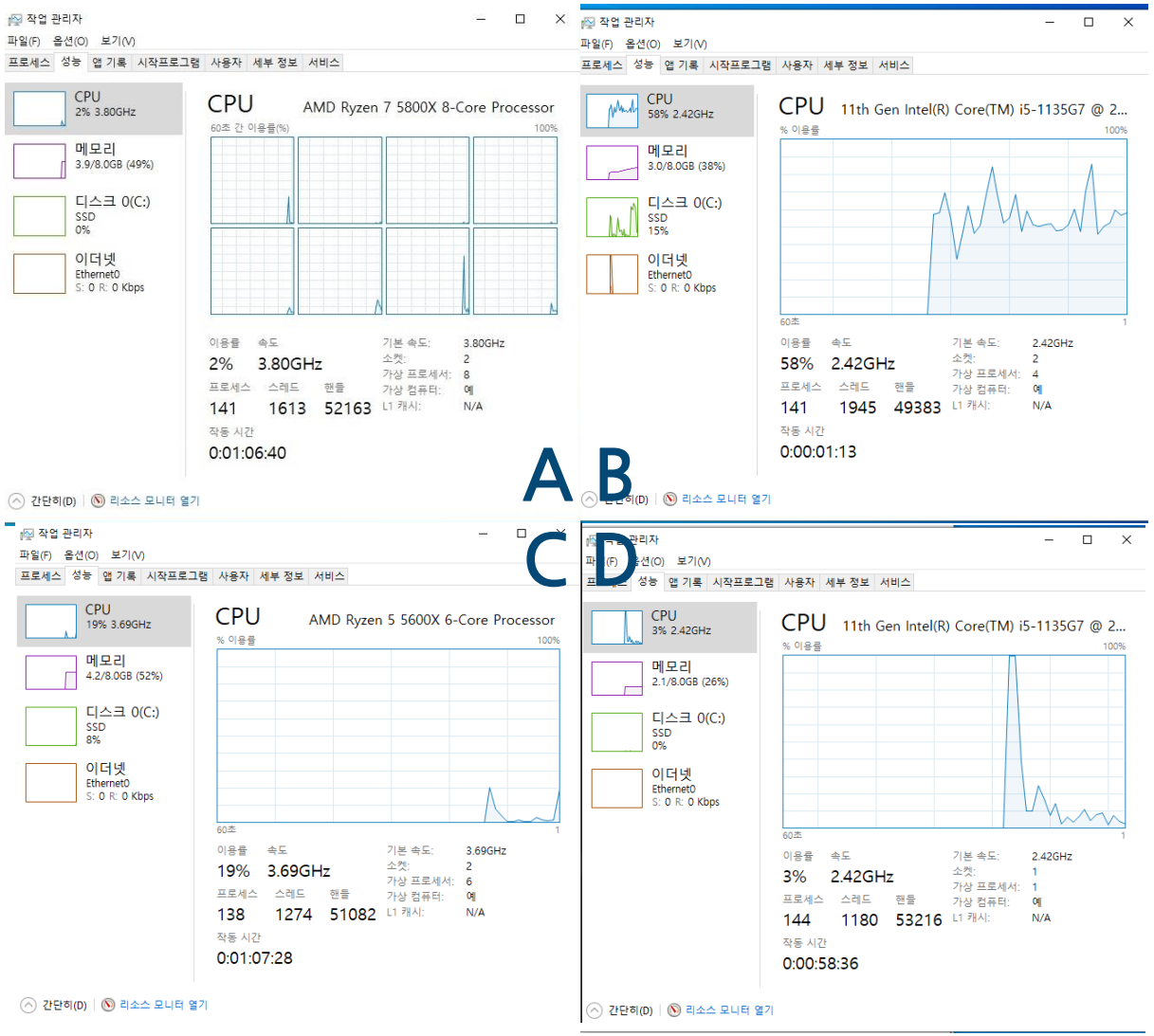
VPN 서버설정

PPTP 서버	◉ 실행	○ 중단	암호화(MPPE) 사용 ☐
L2TP 서버	○ 실행	◉ 중단	비밀키

VPN 접속 계정 관리

VPN 접속 계정	할당 될 IP 주소	연결 상태
serverA	192.168.0.21	접속됨(PPTP)
serverB	192.168.0.22	접속됨(PPTP)
serverC	192.168.0.23	접속됨(PPTP)
webclient	192.168.0.20	접속됨(PPTP)
serverD	192.168.0.24	접속됨(PPTP)

서버4대와 클라이언트를 VPN에 접속



A - C - B - D
순으로 성능이 좋지만
네트워크 환경과 같은 부가적인 요소로 인해서
실질적인 워드카운팅 속도는
A - D - C -B
순으로 빠르다.

인터넷 환경 테스트

VPN을 연결하여 기사 1500줄 분석
VPN과 네트워크상의 오버헤드가 발생 → 많은 시간

성능에 따라서 분산 X / 750 씩 균등하게 분산

성능이 좋은 A컴퓨터에 많은 부하
나머지 컴퓨터에 적은 부하

성능에 따라서 분산 / 750개 250개 250개 250개씩 분산

성능이 좋은 A컴퓨터에 많은 부하
성능이 좋지 못한 B컴퓨터에 부하를 줄이고 나머지 PC도 줄임

성능에 따라서 분산 / 1000개 100개 200개 200개씩 분산

이를 통해
각 머신들의 성능을 고려하여 연산을 분산하는것도 중요하다는 것을 알 수 있었다.