문제해결프로젝트 프로젝트 V

자카드 유사도 (5점)

자카드 유사도는 집합 간의 유사도를 검사하는 여러 방법 중의 하나로 알려져 있다. 두 집합 A, B 사이의 자카드 유사도 J(A, B)는 두 집합의 교집합 크기를 두 집합의 합집합 크기로 나눈 값으로 정의된다.

예를 들어 집합 A = {1, 2, 3}, 집합 B = {2, 3, 4}라고 할 때, 교집합 A ∩ B = {2, 3}, 합집합 A ∪ B = {1, 2, 3, 4}이 되므로, 집합 A, B 사이의 자카드 유사도 J(A, B) = 2/4 = 0.5가 된다. 집합 A와 집합 B가 모두 공집합일 경우에는 나눗셈이 정의되지 않으니 따로 J(A, B) = 1로 정의한다.

자카드 유사도는 원소의 중복을 허용하는 다중집합에 대해서 확장할 수 있다. 다중집합 A는 원소 "1"을 3개 가지고 있고, 다중집합 B는 원소 "1"을 5개 가지고 있다고 하자. 이 다중집합의 교집합 A ∩ B는 원소 "1"을 min(3, 5)인 3개, 합집합 A ∪ B는 원소 "1"을 max(3, 5)인 5개 가지게 된다. 다중집합 A = {1, 1, 2, 2, 3}, 다중집합 B = {1, 2, 2, 4, 5}라고 하면, 교집합 A ∩ B = {1, 2, 2}, 합집합 A ∪ B = {1, 1, 2, 2, 3, 4, 5}가 되므로, 자카드 유사도 J(A, B) = 3/7, 약 0.42가 된다.

이를 이용하여 문자열 사이의 유사도를 계산하는데 이용할 수 있다. 문자열 "FRANCE"와 "FRENCH"가 주어졌을 때, 이를 두 글자씩 끊어서 다중집합을 만들 수 있다. 각각 {FR, RA, AN, NC, CE}, {FR, RE, EN, NC, CH}가 되며, 교집합은 {FR, NC}, 합집합은 {FR, RA, AN, NC, CE, RE, EN, CH}가 되므로, 두 문자열 사이의 자카드 유사도 J("FRANCE", "FRENCH") = 2/8 = 0.25가 된다.

이때 영문자로 된 글자 쌍만 유효하고, 기타 공백이나 숫자, 특수 문자가 들어있는 경우는 그 글자 쌍을 버린다. 예를 들어 "ab+"가 입력으로 들어오면, "ab"만 다중집합의 원소로 삼고, "b+"는 버린다.

다중집합 원소 사이를 비교할 때, 대문자와 소문자의 차이는 무시한다. "AB"와 "Ab", "ab"는 같은 원소로 취급한다.

두 문자열이 주어졌을 때 자카드 유사도를 계산하는 프로그램을 작성하시오. (실행시간 0.2초 이하 @i7 3.8GHz CPU, 입력 받는 시간 제외)

<입력 조건>
- 첫 번째 줄과 두 번째 줄에 두 문자열이 주어진다. 각 문자열의 길이는 2 이상 1000 이하이다.

<출력 조건>
- 첫 번째 줄에 두 문자열의 자카드 유사도를 출력한다. 유사도 값을 0부터 1 사이의 실수이므로, 이를 다루기 쉽도록 65536을 곱한 후 소수점 아래를 버리고 정수부만 출력한다.
- 두 번째 줄에 입력 받은 이후부터 결과 출력까지 걸린 실행시간을 초 단위로 출력한다.

<예제 입력>

```
FRANCE
french
```

<예제 출력>

```
16384
실행시간: xxx초
```

<예제 입력>

```
handshake
shake hands
```

<예제 출력>

```
65536
실행시간: xxx초
```

<예제 입력>

```
aa1+aa2
AAAA12
```

<예제 출력>

```
43690
실행시간: xxx초
```

<주의사항>

- Cpp 파일 1개(자카드유사도.cpp)를 과제 게시판에 업로드할 것 (그 외 파일은 허용 안됨, 압축파일 형태로 제출하지 말 것)
- Cpp 파일의 코드에 주석을 상세히 기입할 것
- 필요한 모든 헤더 파일 및 함수를 cpp 파일에 포함시킬 것
- 띄어쓰기나 줄 바꿈에 주의할 것
- 수강생들간의 Copy 발견 시 모두 0점 처리함
- GNU Compiler Collection (g++ 9.2, clang++ 10.0) 컴파일러에서 에러 없이 실행되어야 함.
- 실행시간 측정 방법, 컴파일러 설치 및 설정 방법은 첫 번째 프로젝트 첨부파일 참조

<평가기준>

- 다양한 입력 테스트케이스에 대해서 프로그램 실행 시 출력 값이 모두 맞고, 실행시간 조건을 만족할 경우 5점 만점 처리함
- 다양한 테스트케이스에 대해서 프로그램 실행 시 출력 값이 모두 맞고, 실행시간 조건을 위반하거나 실행시간 출력이 없는 경우 4점 처리함
- 다양한 테스트케이스에 대해서 프로그램 실행 시 출력 값이 한 번이라도 틀린 경우 3점 처리함
- 컴파일 에러, 런타임 에러 등으로 인해 프로그램 실행이 안 될 경우 2점 처리함
- 주석 설명이 없거나 불충분하면 2점 처리함
- 핵심 구현 내용 부재 시 1점 처리함 (예시: 입력만 받고 처리에 대한 구현이 없는 경우)
- 기한 내 미제출하거나 Copy 발견시 0점 처리함