

CSE3081/AIE3051 (1반): 알고리즘 설계와 분석

[숙제 4] Dynamic Programming 연습

담당 교수: 임 인 성

2025년 11월 18일

마감: 11월 29일 토요일 오후 8시 정각

제출물 및 제출 방법: 정확한 제출물과 방법은 조교가 사이버 캠퍼스에 공지할 예정임.

목표: (1) 알고리즘 설계 기법 중의 하나인 Dynamic Programming 방법에 대한 이해도를 높이도록 한다. (2) 주어진 문제로부터 재귀적인 구조를 유추하고, 이를 테이블을 사용하여 계산하는 과정에 대하여 연습하여 본다.

1. 다음 글을 읽은 후 아래에 기술하는 gapped alignment 문제를 풀어보자 (출처: Wikipedia).

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.

Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

두 개의 문자열 (string)에 대해 임의로 indel이라고 하는 gap -을 적절히 삽입할 수 있으나, 그러한 삽입을 가급적 방지하기 위하여 매번 $-p$ 점 만큼의 감점을 부과한다. 만약 indel이 아닌 두 대응되는 문자가 일치하면 s 점 만큼의 점수를, 아닐 경우 $-f$ 점 만큼의 감점을 부과한다. 만약 $s = 2$, $f = 1$, $p = 2$ 라고 가정할 경우, $X = \text{ATCGGATCT}$ 와 $Y = \text{ACGGACT}$ 에 대해 $X = \text{ATCGGAT-CT}$ 와 $Y = \text{A-C-GG-ACT}$ 와 같이 gap이 삽입되었다면 전체 유사성 점수는 1점, 그리고 $X = \text{ATCGGATCT}$ 와 $Y = \text{A-CGG-ACT}$ 와 같이 gap이 삽입되었다면 전체 유사성 점수는 7점이 된다.

이제 임의로 주어진 문자열 $X = x_1x_2 \cdots x_m$ 과 $Y = y_1y_2 \cdots y_n$ 에 대하여, $O(mn)$ 시간 복잡도를 가지는 dynamic programming 알고리즘을 사용하여, 전체 유사성 점수를 최대로 해주는 gap 삽입 방법을 출력해주는 프로그램을 작성하여 보자. 여러분의 프로그램은 다음과 같은 입출력 요구사항을 만족해야 한다.

입력 형식

프로그램이 수행되면 이름이 `input.txt`인 텍스트 파일에서 다음과 같은 형식으로 저장되어 있는 정보를 읽어 들여야 한다.

```
twostrings.bin
s f p
```

이 파일의 첫 번째 줄에는 두 문자열 데이터를 저장하고 있는 텍스트 파일의 이름이 저장되어 있다. 이 파일에는 이진 형식 (binary format)으로 데이터가 저장되어 있는데, 첫 4 바이트에는 X 의 길이 m , 그리고 다음 4 바이트에는 Y 의 길이 n 이 각각 int 타입으로 저장이 되어 있고, 다음 m 바이트에는 X 가, 그리고 그 다음 n 바이트에는 Y 가 char 타입으로 저장이 되어 있다. 다음 세 개의 양의 정수 s, f, p 가 (의미는 분명) int 타입으로 저장되어 있다.

출력 형식

프로그램 수행 후 이름이 `output.txt`인 텍스트 파일에 다음과 같은 방식으로 계산 결과를 출력하라.

```
1
10
1
8
3
2
4
7
```

여기서 첫 줄에는 자신이 구한 최대 전체 유사성 점수, 두 번째 줄에는 gap을 포함한 전체 문자열의 길이가 저장이 되어야 한다. 다음 X 에 삽입된 gap의 개수와 그 개수 만큼 gap이 삽입된 위치를 저장되어야 하며, 그 다음 마찬가지로 Y 에 대한 gap 정보가 저장되어야 한다.

- 회문(回文)이란 'madam'이나 '소주만병만주소'처럼 앞에서부터 읽으나 뒤에서부터 읽으나 동일한 단어나 구를 뜻한다. 이 문제에서는 '0', '1', '2', …, '9' 등 10개의 문자로 구성된 문자열을 생각하자. 임의의 문자열은 몇 개의 회문으로 구성되어 있다고 할 수 있는데 (한 개의 문자로 구성된 문자열은 그 자체가 회문임), 예를 들어, '37352259889'의 경우,

- '373' + '5225' + '9889',
- '373' + '5' + '22' + '5' + '9889',
- '3' + '7' + '3' + '5225' + '9' + '8' + '8' + '8' + '9'

등과 같이 회문으로 구성되어 있다고 할 수 있다.

지금 임의로 주어진 문자열 $X = x_1x_2 \dots x_n$ 을 dynamic programming 기법을 사용하여 최소 개수의 회문으로 나누어 주는 프로그램을 작성하여 보자. 여러분의 프로그램은 다음과 같은 입출력 요구사항을 만족해야 한다.

입력 형식

입력 데이터는 이름이 `input.txt`인 텍스트 파일에 저장되어 있다. 첫 째 줄에는 입력 문자열의 개수가 주어져 있고, 다음 그 개수만큼 각 문자열의 길이와 문자열 쌍이 순서대로 저장되어 있다.

```
2
11
37352259889
14
57134317562326
```

출력 형식

계산 결과는 이름이 `output.txt`인 텍스트 파일 (ASCII format)에 저장되어야 한다. 첫 째 줄에는 문제를 해결한 입력 문자열의 개수가 저장되어야 하며, 다음 그 개수만큼 각 문자열에 대해 최소 회문 개수와 해당 회문이 저장되어야 한다.

```

2
3
373
5225
9889
2
571343175
62326

```

3. 다음의 문제를 dynamic programming 기법을 사용하여 해결하자.

Given a set of n positive integers $A = \{a_1, a_2, \dots, a_n\}$, partition A into two subsets A_1 and A_2 (i.e. $A = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$) such that $|S_1 - S_2|$ is minimized, where S_i is the sum of elements in A_i ($i = 1, 2$).

이제 주어진 집합 A 에 대하여 (i) 위 문제에서 찾고자 하는 최소 차이와 (ii) 그러한 최소 차이에 해당하는 두 집합 A_1 과 A_2 를 구하는 프로그램을 작성하라.

입력 형식

입력 데이터는 이름이 `input.txt`인 텍스트 파일 (ASCII format)에 저장되어 있다. 입력 파일의 첫 줄에는 테스트 케이스의 개수가 정수 타입으로 주어져 있고, 이후 각 케이스에 대한 입력 데이터가 주어진다. 각 케이스에 대해서는 먼저 첫 줄에 A 의 원소의 개수 n 이 정수 타입으로 저장되어 있고 (원소의 개수는 최대 1024개까지 처리 가능하여야 함), 이후 그 개수에 해당하는 줄 각각에 A 의 원소가 정수 타입으로 저장되어 있다.

```

2
4
1
2
3
4
7
3
8
17
2
6
1
4

```

출력 형식

계산 결과는 이름이 `output.txt`인 텍스트 파일에 저장되어야 한다. 출력 파일의 첫 줄에는 테스트 케이스의 개수를 저장해야 하며 (입력 데이터의 첫 줄과 동일), 다음 각 테스트 케이스에 대한 답이 순서대로 저장되어야 한다. 각 테스트 케이스의 경우 먼저 첫 줄에 최소 차이, A 의

원소의 개수 n , 그리고 첫 번째 부분 집합 A_1 의 원소의 개수 n_1 등 세 개의 정수를 저장한 후, 이후 두 부분 집합의 원소를 한 줄에 한 개씩 순차적으로 저장해야 한다. 이때 첫 n_1 개의 원소는 A_1 의 원소로, 그리고 나머지 $n - n_1$ 개의 원소는 A_2 의 원소로 간주한다. 만약 동일한 최소 비용을 가지는 집합 분할이 두 개 이상 존재할 경우, 어떠한 결과를 출력해도 문제를 올바로 푼 것으로 간주한다. 다음은 앞의 입력 데이터에 대한 문제 풀이 결과를 보여주고 있다.

```
2
0 4 2
1
4
2
3
1 7 4
3
8
6
4
17
2
1
```

- [주의]

1. 조교는 자신의 명령어 파일과 입력 데이터를 사용하여 여러분의 프로그램이 정확한 값을 계산하는지 확인할 예정이니, 자신이 생성한 데이터 (필요 시 적절히 압축하여)와 원시 코드를 조교가 수행하기 편리한 형태로 제출할 것.
2. 숙제 제출 기간 동안 조교가 숙제와 관련하여 중요한 공지 사항을 게시판에 올릴 수 있으니 항상 수업 게시판을 확인하기 바람.
3. 제출 파일에서 바이러스 발견 시 본인 점수 **X (-1)**이고, 다른 사람의 숙제를 복사할 경우 관련된 사람 모두에 대하여 만점 X (-10)임.