

자료 구조 숙제 #06 맵

2분반 마감 시간: 5월 17일 오후 11시 59분
1분반 마감 시간: 5월 20일 오후 11시 59분

2018/05/11

컴퓨터과학과
민경하

문제

- 텍스트 파일로부터 사용된 빈도순으로 단어를 계산하는 프로그램을 BST를 이용한 map과 heap을 이용해서 작성하시오.
- 단, 출력은 상위 10개의 단어와 그 빈도수만 하도록 한다.

예제

- 다음과 같은 텍스트 파일에 대해서는 다음과 같은 결과를 출력한다.

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th President of the United States from January 20, 2009 to January 20, 2017. The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008. Before that, he served in the Illinois State Senate from 1997 until 2004.

```
[5] the
[4] from
[3] to
[2] January
[2] Illinois
[2] United
[2] States
[2] American
[2] he
[2] served
```

예제

- 단, 알파벳 이외의 다른 문자가 섞인 단어는 모두 배제한다.

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th President of the United States from January 20, 2009 to January 20, 2017. The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008. Before that, he served in the Illinois State Senate from 1997 until 2004.

- 배제되는 단어들의 예)
 - 4
 - 44th
 - 1961)
 - 20,
 - 2017.
- 다음의 단어들은 처음과 마지막의 문자를 제거하고 고려할 것
 - (born → born
 - presidency, → presidency
 - that, → that

해결 방법 (map)

- 텍스트 파일을 읽으면서 단어들을 추출해서 BST에 삽입한다.
- BST의 node는 단어와 함께 단어의 개수를 세는 변수를 저장한다.
- 삽입 시에 동일한 단어를 가진 노드를 만나면 단어의 개수를 1 증가한다.

해결 방법 (heap)

- Map에서 단어들을 하나씩 제거하고 그 개수를 이용해서 max heap에 저장한다.
- 저장이 끝나면 max heap의 top을 10회 제거하고 그 값을 출력한다.

평가

- 10개의 단어 당 10점씩

벌점

- 1시간 늦으면 10% 감점
- Copy 또는 copied는
 - 1회 적발은 해당 숙제 0점 처리
 - 2회 적발은 이 과목 F 처리
- 본인이 직접 숙제 하지 않은 경우는 copy 또는 copied와 동일하게 처벌함
 - 반드시 코드를 설명할 수 있을 것