Data structure Project

ECLAB 전자상거래 및 인터넷응용 연구실
Electronic Commerce and Internet Application Laboratory

2016.05.09 ~ 2016.05.27

EC Lab

한상용 교수님

과제 안내

- Project #1 Heap sort
  - 기간 : 2016년 5월 9일 ~ 2016년 5월 18일 오후 11시 59분 (10일)
  - 업로드 : e-class, 과제방
- Project #2 Huffman Coding
  - 기간 : 2016년 5월 19일 ~ 2016년 27일 오후 11시 59분 (10일)
  - 업로드 : e-class, 과제방

문의 : alex.k.yura@ec.cse.cau.ac.kr

전자상거래 연구실, 218호, 208관 (제 2공학관)

Contents

Heap Sort

Problem

- 주어진 dataset에 있는 단어들을 Heap sort로 정렬하여 각 단어들이 몇 개가 있는지 출력하는 프로그램

  제한 조건) 마침표(.), 쉼표(,), 기타 특수기호(“,”, -, (,), etc…), 숫자 들은 제외함

  오직 알파벳으로만 이루어진 단어만 취급함

  소문자 알파벳만 취급함 (대문자를 소문자로 바꾸어서 처리해야 함)

  결과는 파일로 만들어 내야 함 (filename : wordlist.txt)

- dataset : 영자 신문 기사
- 단어 : 문장에서의 각 어절(띄어쓰기로 구분)을 단어라고 표현함.

  ex) 문장 : I am a boy, I am not a girl

  단어 : I, am, a, boy, I, am, not, a, girl

In & Output

〈Input〉
filename : dataset.txt

```
SEOUL, South Korea — After three straight losses, a South Korean expert rallied
on Sunday for his first victory against a Google computer program playing Go, an
ancient board game known as the most complex ever invented.
Lee Se-dol, 33, a boyish South Korean Go master, was all smiles after a
brilliant move forced the Google program, AlphaGo, to surrender the match in the
middle of the contest.
Hundreds of local Go enthusiasts and reporters who were gathered at the Four
Seasons Hotel in downtown Seoul burst into applause over the human Go master's
dramatic comeback against the machine.
"You know, I have played many, many Go games, but I don't think I have ever
been as happy with one single victory as with this one," Mr. Lee said. "This
is priceless."
Demis Hassabis, the chief executive of Google DeepMind, Google's artificial
intelligence company, said Mr. Lee's victory was a reminder that AlphaGo still
had room for improvement. A "creative genius" like Mr. Lee tests the limits of
the machine, he said.
Go has been seen as the last great challenge in computer programmers' efforts
to create software that can outwit humans in board games. Go is such a complex
game, with an almost infinite possible sequence of moves, that artificial-
intelligence experts had predicted that computer programs needed more than 10
more years before they would be able to beat Go legends like Mr. Lee, who has 18
international titles.
Before the best-of-five series began on Wednesday, Mr. Lee had been upbeat. But
the mood quickly sank after he lost the first three matches.
After losing his third match — and $1 million in prize money — on Saturday,
Mr. Lee admitted that the psychological pressure he felt in facing a nonhuman
foe was a big handicap.
```

〈Output〉
filename : wordlist.txt

```
a, 14
after, 4
an, 2
and, 4
.
.
.
.
```

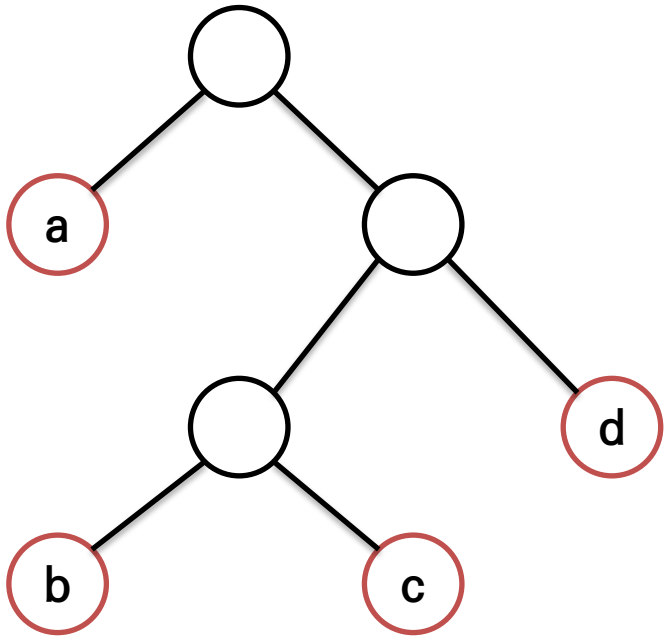
Huffman Coding

Huffman Coding

- Huffman Coding은 무손실 압축 기법(lossless data compression) 중 하나다

- Huffman Coding은 leaf node가 알파벳(character)으로 이루어진 트리를 만든다.

ex)



- 왼쪽으로 갈 때 0 을 할당, 오른쪽으로 갈 때 1 을 할당하면 알파벳을 다음과 같이 나타낼 수 있다.

- a 0
  b 100
  c 101
  d 11

Data compression

- 알파벳 한 글자는 8bit으로 이루어져 있는데, 문서에서 글자가 나타난 빈도에 따라 더 짧은 binary code를 할당함으로써 데이터를 압축할 수 있다.
- 예시)

alphabet	frequency	binary
a	10	0
b	2	100
c	3	101
d	7	10

original text = bcdaadadabaadccaddadaa

bit = $22 \times 8 = 176$

encoded text =

1001011000100100100001010110101001000

bit =

$$(10 \times 1) + (2 \times 3) + (3 \times 3) + (7 \times 2) = 39$$

Algorithm

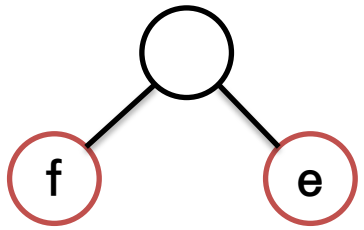
1. Initial data sorted by frequency (ascending power)
2. Combine the two lowest frequencies (make a sub-tree)

   Move it into its correct place
3. Repeat step 2 until remained data is noting

Example (1)

alphabet	a	b	c	d	e	f
frequency	45	13	12	16	9	5

1. combine lowest frequency : f, e

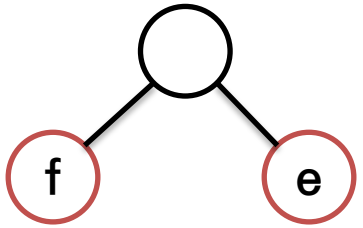


new1

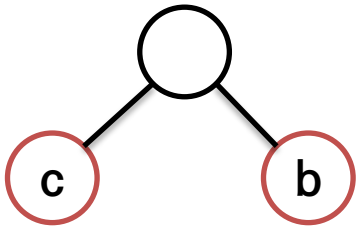
new frequency = f's + e's = 14

alphabet	a	b	c	d	new1
frequency	45	13	12	16	14

2. combine lowest frequency : c, b



new1



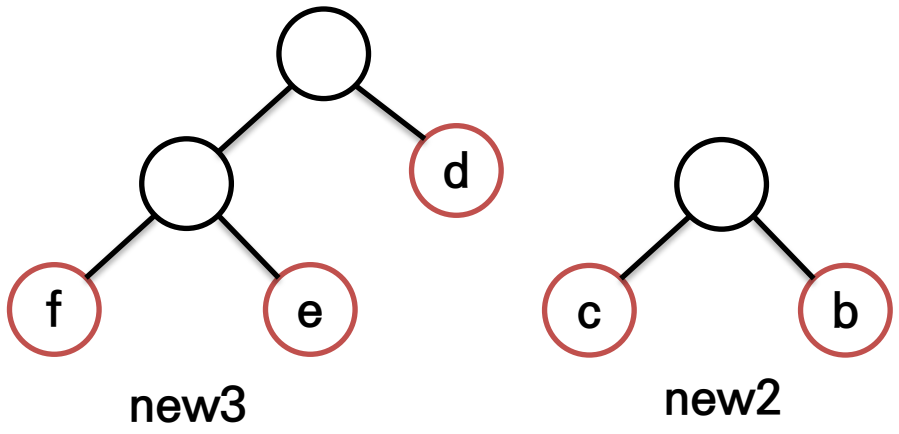
new2

new frequency = c's + b's = 25

Example (2)

alphabet	a	new2	d	new1
frequency	45	25	16	14

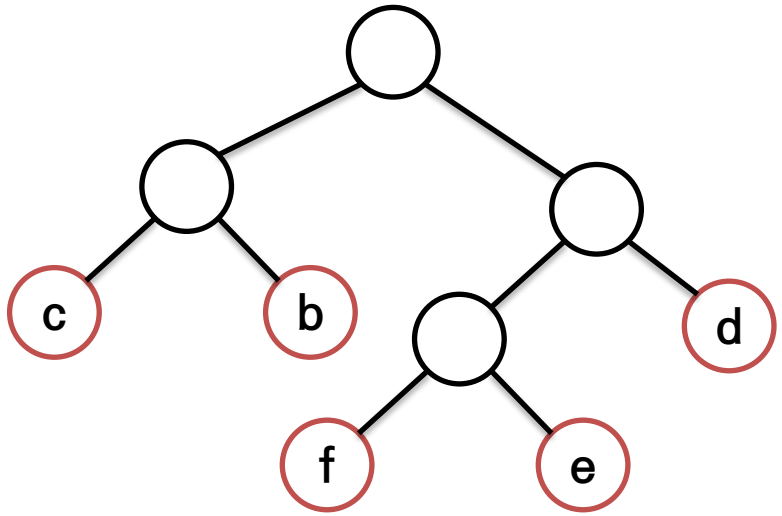
3. combine lowest frequency : new1, d



new frequency = new1's + d's = 30

alphabet	a	new2	new3
frequency	45	25	30

4. combine lowest frequency : new2, new3

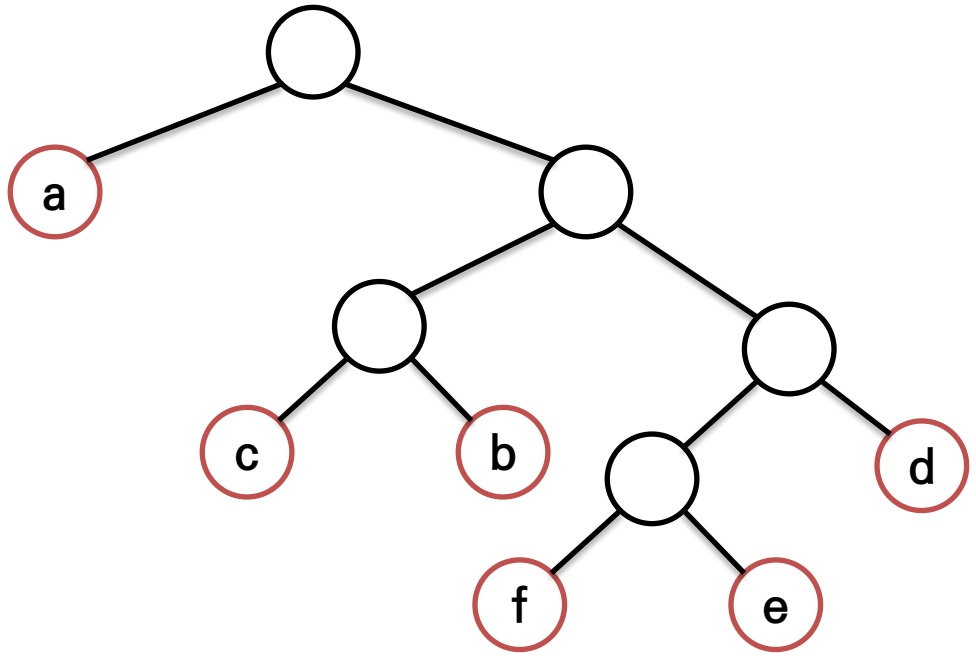


new frequency = new2's + new3's = 55

Example (2)

alphabet	a	new4
frequency	45	55

5. combine lowest frequency : a, new4



alphabet	binary
a	0
b	101
c	100
d	111
e	1101
f	1100

Evaluation

- Average bit rate

$$ABR = (F_1L_1 + F_2L_2 + \cdots + F_mL_m)/N$$

- F_i : The frequency of the i'th symbol
- L_i : The length of the code for the i'th symbol
- N : the length of the file

Project Problem

- 주어진 파일(dataset.txt)을 압축하기 위한 Huffman coding 프로그램 만들기
- 파일을 읽어 Huffman tree를 이용해 알파벳의 빈도에 따른 적절한 binary code 할당
- 실행 결과는 콘솔로 표시함
- Encoding할 문자는 알파벳 소문자 a-z만 사용 (특수문자는 무시, 대문자->소문자 변경)
- 기본적인 틀은 주어짐 (projectmain.txt -> projectmain.cpp 로 사용)
- /* write here */ 라는 주석이 달린 부분 및 파일 전처리 부분 코딩(표시는 따로 안되어 있음!)
- 사용 언어 : C, C++
- 제한 조건 : 외부 라이브러리 사용 금지
  gcc 사용 금지 (평가를 visual studio 2010으로 함)

Input & Predicted Output

〈Input〉
filename : dataset.txt

```
SEOUL, South Korea — After three straight losses, a South Korean expert rallied
on Sunday for his first victory against a Google computer program playing Go, an
ancient board game known as the most complex ever invented.
Lee Se-dol, 33, a boyish South Korean Go master, was all smiles after a
brilliant move forced the Google program, AlphaGo, to surrender the match in the
middle of the contest.
Hundreds of local Go enthusiasts and reporters who were gathered at the Four
Seasons Hotel in downtown Seoul burst into applause over the human Go master's
dramatic comeback against the machine.
"You know, I have played many, many Go games, but I don't think I have ever
been as happy with one single victory as with this one," Mr. Lee said. "This
is priceless."
Demis Hassabis, the chief executive of Google DeepMind, Google's artificial
intelligence company, said Mr. Lee's victory was a reminder that AlphaGo still
had room for improvement. A "creative genius" like Mr. Lee tests the limits of
the machine, he said.
Go has been seen as the last great challenge in computer programmers' efforts
to create software that can outwit humans in board games. Go is such a complex
game, with an almost infinite possible sequence of moves, that artificial-
intelligence experts had predicted that computer programs needed more than 10
more years before they would be able to beat Go legends like Mr. Lee, who has 18
international titles.
Before the best-of-five series began on Wednesday, Mr. Lee had been upbeat. But
the mood quickly sank after he lost the first three matches.
After losing his third match — and $1 million in prize money — on Saturday,
Mr. Lee admitted that the psychological pressure he felt in facing a nonhuman
foe was a big handicap.
```

〈Predicted Output〉

```
Character:l     Frequency:17    Bin: 11
Character:f     Frequency:8     Bin: 101
Character:s     Frequency:9     Bin: 100
Character:a     Frequency:5     Bin: 0111
Character:d     Frequency:5     Bin: 0110
Character:j     Frequency:12    Bin: 010
Character:k     Frequency:14    Bin: 001
Character:e     Frequency:7     Bin: 0001
Character:w     Frequency:2     Bin: 000011
Character:r     Frequency:1     Bin: 0000101
Character:t     Frequency:1     Bin: 0000100
Character:c     Frequency:1     Bin: 0000011
Character:i     Frequency:1     Bin: 0000010
Character:b     Frequency:1     Bin: 0000001
Character:q     Frequency:1     Bin: 0000000
```

Dataset

- 출처 : http://www.nytimes.com/2016/03/14/world/asia/south-korean-gets-priceless-victory-over-computer-in-go-match.html?_r=0

- 파일은 e-class에 업로드 되어 있음!

Thank you