

빅데이터 최신기술

음절 Bigram 확률 계산을 이용한
문장 생성

소프트웨어학부

20163162

차운성

1. 과제 수행 내용 설명

- 프로그램 개요

프로그램 실행 명령 옵션으로 n과 대용량 말뭉치 파일을 입력하면, 대용량 말뭉치에서 가장 빈도수가 높은 음절 중 상위 n개 리스트에서 무작위로 하나를 뽑아 시작 음절(현재 음절)로 선택하고, 현재 음절에서 다음으로 나오는 음절 중 빈도수가 가장 높은 3개 중 무작위로 하나 뽑아 다음 음절로 선택한다. 해당 과정을 반복 후 문장을 이루는 음절의 개수가 10개 이상이 되고 문장을 구성하는 마지막 음절이 '다'일 때 문장 구성을 완료하고 프로그램을 종료한다.(n: 3~5 사이의 값)

- 예외 처리

선택된 현재 음절로부터 다음 음절을 선택할 때 출현 빈도가 1인 이상인 음절이 없는 경우 음절 unigram 빈도가 높은 m개의 음절 중에서 임의로 1개의 음절을 선택한다.

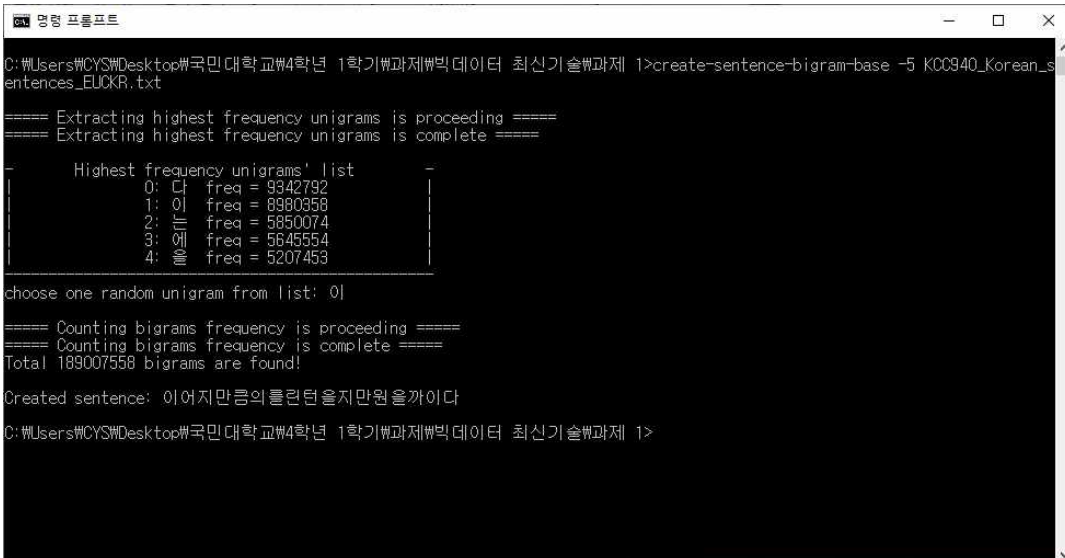
처음 시작음절 선택 시 unigram 추출 과정에서 각 음절들의 빈도수를 배열로 저장하도록 했는데, 다음 음절의 출현 빈도를 나타내는 구조체 배열에 있는 모든 음절의 빈도수가 0인 경우 전체 unigram 인덱스 중 random하게 추출해 다음 음절로 선택하도록 작성하였다.

- 세부 구조

상위 n개의 unigram을 뽑기 위해 먼저 상위바이트 인덱스, 하위바이트 인덱스, 해당 음절의 빈도수를 가지는 구조체를 만들고, 해당 구조체 배열을 만든다. 구조체 배열에 unigram 중 상위 n개의 unigram을 담고 n개의 unigram 중 하나의 unigram을 무작위로 추출해 시작음절로 설정한다. 이후 bigram에 대한 cnt를 계산한 후 시작음절에 대해 다음음절의 빈도수가 높은 3개의 음절을 구조체 배열에 담고, 이 중 무작위로 하나 추출해 그 다음 음절로 설정한다. 해당 과정을 반복해 문장의 음절이 10개 이상이 되고 다음 음절이 '다'일 경우 문장 생성을 종료한 후 프로그램을 종료한다.

2. 실행화면 스크린샷

1) n=5일 때의 실행 화면



```
C:\Users\CYS\Desktop\국민대학교\4학년_1학기\부과제\부과제_데이터_최신기술\과제_1>create-sentence-bigram-base -5 K00940_Korean_s
entences_EUCKR.txt

===== Extracting highest frequency unigrams is proceeding =====
===== Extracting highest frequency unigrams is complete =====

Highest frequency unigrams' list
-----
0: 다 freq = 9342792
1: 이 freq = 8980358
2: 는 freq = 5850074
3: 에 freq = 5645554
4: 을 freq = 5207453
-----

choose one random unigram from list: 0

===== Counting bigrams frequency is proceeding =====
===== Counting bigrams frequency is complete =====
Total 189007558 bigrams are found!

Created sentence: 이어지만큼의틀린턴을지만됨을까이다

C:\Users\CYS\Desktop\국민대학교\4학년_1학기\부과제\부과제_데이터_최신기술\과제_1>
```

2) n=4일 때의 실행 화면

```
명령 프롬프트
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 1>create-sentence-bigram-base -4 KCC940_Korean_s
entences_EUCKR.txt

==== Extracting highest frequency unigrams is proceeding ====
==== Extracting highest frequency unigrams is complete ====

Highest frequency unigrams' list
-
0: 다 freq = 9342792
1: 이 freq = 8980358
2: 는 freq = 5850074
3: 예 freq = 5645554
-

choose one random unigram from list: 예

==== Counting bigrams frequency is proceeding ====
==== Counting bigrams frequency is complete ====
Total 189007558 bigrams are found!

Created sentence: 예는다른바람들이다고위해서울시작가능하는다
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 1>
```

3) n=3일 때의 실행 화면

```
명령 프롬프트
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 1>create-sentence-bigram-base -3 KCC940_Korean_s
entences_EUCKR.txt

==== Extracting highest frequency unigrams is proceeding ====
==== Extracting highest frequency unigrams is complete ====

Highest frequency unigrams' list
-
0: 다 freq = 9342792
1: 이 freq = 8980358
2: 는 freq = 5850074
-

choose one random unigram from list: 다

==== Counting bigrams frequency is proceeding ====
==== Counting bigrams frequency is complete ====
Total 189007558 bigrams are found!

Created sentence: 다고객이라며 칠레이라며 칠성이다
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 1>
```