

빅데이터 최신기술

문장 생성 확률 계산

소프트웨어학부

20163162

차운성

1. 과제 수행 내용 설명

- 프로그램 개요

프로그램 실행 명령 옵션으로 대용량 말뭉치 파일을 입력하면, 대용량 말뭉치에서의 bigram 음절을 count 한 후 확률을 계산하려는 문장에서 시작음절과 끝 음절은 고려하지 않고, 한 음절 이후 등장하는 다음 음절에 대한 확률값을 각각 계산해 곱한다.

$$P(s1, s2, s3, ..., sn) = P(s2 | s1) \times P(s3 | s2) \times \dots \times P(sn | sn-1)$$

키보드로 입력받은 문장들의 확률을 모두 구했으면 “q”를 입력해 키보드 입력을 중지하면 프로그램이 생성한 문장들의 확률을 각각 구하고, 각 케이스별로 가장 높은 확률을 가지는 문장과 해당 문장의 생성확률을 출력한다.

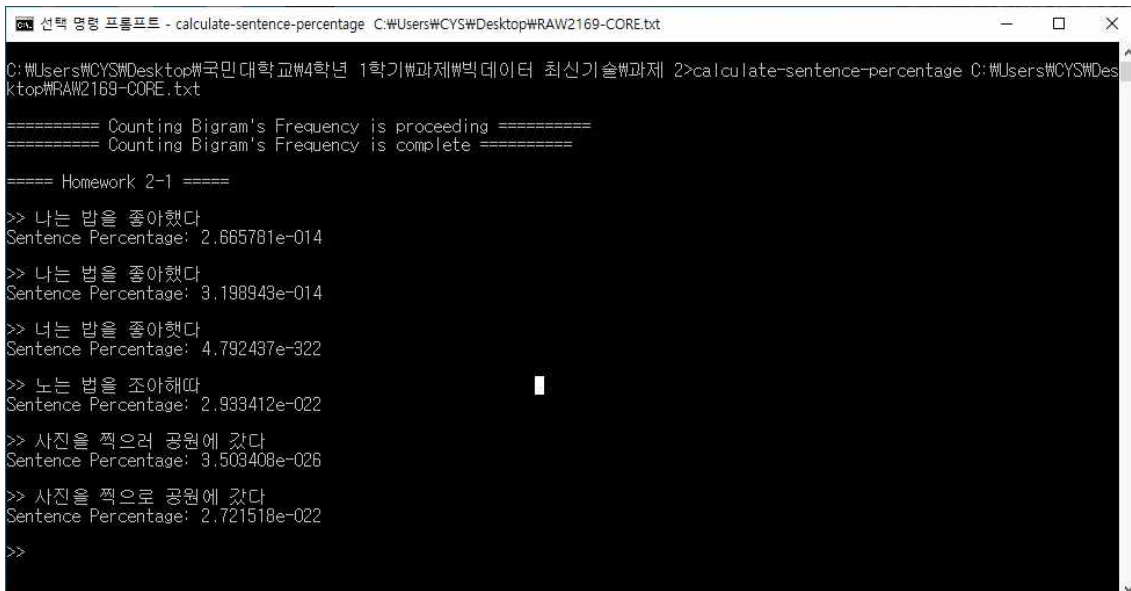
- 세부 구조

각 음절의 생성확률을 계속해서 곱해나가면 확률값이 상당히 작아지기 때문에 %f 포맷으로 확률값을 출력하면 0.0으로 수렴한 값이 출력된다. 때문에 %f 포맷이 아닌 %e 포맷을 사용해 확률값을 출력하였다. 또한 어절단위의 공백(' ')을 고려해야 하기 때문에 “가_”의 경우에는 “가”에 해당하는 인덱스로 cntBlank를 count 해주었고, “_가”의 경우에는 “가”에 해당하는 인덱스로 cntBlankStart를 count 해주었다. 이 외의 “가나”와 같이 한글 음절이 연속해서 나오는 경우에는 bigramCnt를 count 해주었다.

PerSentence 구조체를 정의해 과제 2에서 생성되는 문장들의 어절 별 index와 문장 생성 확률을 구조체 배열에 저장하고, 가장 낮은 확률인 DBL_MIN보다 큰 확률의 문장들만 높은 확률 순 내림차순으로 케이스 별로 출력하게 하였다.

2. 실행화면 스크린샷

1) RAW2169-CORE.txt



```
선택 명령 프롬프트 - calculate-sentence-percentage C:\Users\CYS\Desktop\RAW2169-CORE.txt
C:\Users\CYS\Desktop\국민대학교\4학년_1학기\빅데이터_최신기술\과제_2>calculate-sentence-percentage C:\Users\CYS\Desktop\RAW2169-CORE.txt

===== Counting Bigram's Frequency is proceeding =====
===== Counting Bigram's Frequency is complete =====

===== Homework 2-1 =====

>> 나는 밥을 좋아했다
Sentence Percentage: 2.665781e-014
>> 나는 밥을 좋아했다
Sentence Percentage: 3.198943e-014
>> 너는 밥을 좋아했다
Sentence Percentage: 4.792437e-022
>> 나는 밥을 좋아했다
Sentence Percentage: 2.939412e-022
>> 사진을 찍으러 공원에 갔다
Sentence Percentage: 3.503408e-026
>> 사진을 찍으러 공원에 갔다
Sentence Percentage: 2.721518e-022
>>
```

- 음절 출현 확률이 0.0인 경우엔 DBL_MIN으로 설정해 계산했지만, 최종 문장 생성 확률이 0.0인 경우에도 DBL_MIN으로 초기화하였다.

```

>> a
===== Homework 2-2 =====

----- 2-1 Highest Percentage Sentence -----
나.아.했.다. 3.198943e-014
나.아.했.다. 2.665781e-014
나.아.했.다. 1.847680e-014
나.아.했.다. 1.284467e-014
나.아.했.다. 1.070388e-014
나.아.했.다. 1.038981e-014
나.아.했.다. 1.011425e-014
나.아.했.다. 8.428523e-015
나.아.했.다. 7.418967e-015
나.아.했.다. 5.841897e-015
나.아.했.다. 4.888603e-015
나.아.했.다. 4.171806e-015
나.아.했.다. 3.493087e-015
나.아.했.다. 3.284993e-015
나.아.했.다. 3.269316e-015
나.아.했.다. 2.910901e-015
나.아.했.다. 2.724425e-015
나.아.했.다. 2.498616e-015
나.아.했.다. 2.173381e-015
나.아.했.다. 2.062176e-015
나.아.했.다. 2.017575e-015
나.아.했.다. 1.962914e-015
나.아.했.다. 1.888327e-015
나.아.했.다. 1.867000e-015
나.아.했.다. 1.555830e-015
나.아.했.다. 1.545652e-015

```

- 프로그램 내부에서 생성된 문장 중 '나는 밥을 좋아했다'와 '나는 법을 좋아했다',
가 가장 높은 생성 확률을 가지는 문장이 되었고, 이후 DBL_MIN보다 높은 생성 확률을
가지는 문장에 한해서만 내림차순으로 출력되게 하였다.

```

----- 2-2 Highest Percentage Sentence -----
사.전.을.찍.으.로.공.원.에.갔.다. 2.721518e-022
사.전.을.찍.으.로.공.원.에.갔.다. 9.842579e-023
사.전.을.찍.으.로.공.원.에.갔.다. 4.390632e-023
사.전.을.찍.으.로.공.원.에.갔.다. 2.295857e-023
사.전.을.찍.으.로.공.원.에.갔.다. 2.238579e-023
사.전.을.찍.으.로.공.원.에.갔.다. 1.848814e-023
사.전.을.찍.으.로.공.원.에.갔.다. 1.587906e-023
사.전.을.찍.으.로.공.원.에.갔.다. 8.302419e-024
사.전.을.찍.으.로.공.원.에.갔.다. 4.510949e-024
사.전.을.찍.으.로.공.원.에.갔.다. 3.703588e-024
사.전.을.찍.으.로.공.원.에.갔.다. 3.611505e-024
사.전.을.찍.으.로.공.원.에.갔.다. 2.982696e-024
사.전.을.찍.으.로.공.원.에.갔.다. 1.888288e-024
사.전.을.찍.으.로.공.원.에.갔.다. 1.631419e-024
사.전.을.찍.으.로.공.원.에.갔.다. 1.559513e-024
사.전.을.찍.으.로.공.원.에.갔.다. 1.339431e-024
사.전.을.찍.으.로.공.원.에.갔.다. 7.277524e-025
사.전.을.찍.으.로.공.원.에.갔.다. 3.805078e-025
사.전.을.찍.으.로.공.원.에.갔.다. 3.710471e-025
사.전.을.찍.으.로.공.원.에.갔.다. 3.064431e-025
사.전.을.찍.으.로.공.원.에.갔.다. 3.046379e-025
사.전.을.찍.으.로.공.원.에.갔.다. 2.631972e-025
사.전.을.찍.으.로.공.원.에.갔.다. 2.515966e-025
사.전.을.찍.으.로.공.원.에.갔.다. 1.581943e-025
사.전.을.찍.으.로.공.원.에.갔.다. 1.376136e-025
사.전.을.찍.으.로.공.원.에.갔.다. 1.256471e-025
사.전.을.찍.으.로.공.원.에.갔.다. 1.254762e-025
사.전.을.찍.으.로.공.원.에.갔.다. 6.138741e-026

```

- 마찬가지로 '사진을 찍으로 공원에 갔다'와 '사전을 찍으로 공원에 갔다'가 가장 높은
생성 확률을 가지는 문장이 되었고, 이후 DBL_MIN보다 높은 생성 확률을 가지는 문장에
한해서만 내림차순으로 출력되게 하였다.

2) gtle.txt

```
명령 프롬프트 - calculate-sentence-percentage C:\Users\CYS\Desktop\gtlee.txt
C:\Users\CYS\Desktop>calculate-sentence-percentage C:\Users\CYS\Desktop\gtlee.txt

===== Counting Bigram's Frequency is proceeding =====
===== Counting Bigram's Frequency is complete =====

===== Homework 2-1 =====

>> 나는 밥을 좋아했다
Sentence Percentage: 2.005390e-014
>> 나는 밥을 좋아했다
Sentence Percentage: 9.621981e-015
>> 너는 밥을 좋아했다
Sentence Percentage: 2.225074e-308
>> 노는 밥을 좋아했다
Sentence Percentage: 2.225074e-308
>> 사진을 찍으러 공원에 갔다
Sentence Percentage: 1.272075e-026
>> 사진을 찍으로 공원에 갔다
Sentence Percentage: 1.383778e-022
>>
```

- RAW2169-CORE.txt의 bigram 음절을 count 했을 때와 달리 문장 별로 생성 확률이 차이가 있음을 확인할 수 있다.

```
명령 프롬프트
>> q
===== Homework 2-2 =====

----- 2-1 Highest Percentage Sentence -----
내 밥을 좋아했다 3.075939e-014
내 밥을 좋아했다 2.005390e-014
내 밥을 좋아했다 1.475853e-014
내 밥을 좋아했다 1.408514e-014
내 밥을 좋아했다 9.621981e-015
내 밥을 좋아했다 8.466864e-015
내 밥을 좋아했다 7.767681e-015
내 밥을 좋아했다 6.762932e-015
내 밥을 좋아했다 5.520061e-015
내 밥을 좋아했다 5.356802e-015
내 밥을 좋아했다 4.891059e-015
내 밥을 좋아했다 3.879845e-015
내 밥을 좋아했다 3.726979e-015
내 밥을 좋아했다 3.492423e-015
내 밥을 좋아했다 3.381342e-015
내 밥을 좋아했다 2.478442e-015
내 밥을 좋아했다 2.454694e-015
내 밥을 좋아했다 2.346759e-015
내 밥을 좋아했다 2.138141e-015
내 밥을 좋아했다 1.622388e-015
내 밥을 좋아했다 1.615846e-015
내 밥을 좋아했다 1.352755e-015
내 밥을 좋아했다 1.346319e-015
내 밥을 좋아했다 1.228503e-015
내 밥을 좋아했다 1.135718e-015
내 밥을 좋아했다 9.307520e-016
```

- RAW2169-CORE.txt의 bigram 음절을 count 했을 때와 문장 생성 확률이 달라 가장 높은 생성 확률을 가지는 문장이 차이가 있음을 확인할 수 있다.

2-2 Highest Percentage Sentence		
사	진	3.899565e-024
사	진	1.383778e-022
사	진	2.116063e-023
사	진	1.346662e-024
사	진	2.059305e-025
사	진	1.269443e-025
사	진	1.941223e-026
사	진	7.368898e-023
사	진	1.126847e-023
사	진	7.171246e-025
사	진	1.096622e-025
사	진	3.686473e-026
사	진	2.410731e-025
사	진	3.788078e-024
사	진	6.774057e-027
사	진	2.477175e-023
사	진	6.525737e-027
사	진	4.267439e-026
사	진	6.922689e-026
사	진	4.527022e-025
사	진	7.113490e-024
사	진	4.651794e-023
사	진	4.276286e-027
사	진	7.322647e-024
사	진	1.119805e-024
사	진	6.023185e-027
사	진	3.938800e-026
사	진	7.126431e-026
사	진	6.189194e-025

- 마찬가지로 RAW2169-CORE.txt의 bigram 음절을 count 했을 때와 문장 생성 확률이 달라 가장 높은 생성 확률을 가지는 문장이 차이가 있음을 확인 할 수 있다.