

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.

$$s_1, s_2, s_3, \dots, s_n = P(s_2|s_1) \times P(s_3|s_2) \times \dots \times P(s_n|s_{n-1})$$

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.

해당 과정에서 대용량 말뭉치 중 “가?”\r\n”의 경우는 처리해주지 못해 KCC940_Korean_sentences_EUCKR.txt 말뭉치와 gtleee.txt 말뭉치는 정상적으로 bigram 음절을 count 해주었지만 해당 문자열이 포함된 RAW2169-CORE.txt의 경우는 counting 과정에서 runtime error가 발생해 프로그램이 종료되는 문제가 발생하였다.

2. 실행화면 스크린샷

1) KCC940_Korean_sentences_EUCKR.txt

```
명령 프롬프트 - calculate-sentence-percentage KCC940_Korean_sentences_EUCKR.txt
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 2>calculate-sentence-percentage KCC940_Korean_sentences_EUCKR.txt

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

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

>> 나는 밥을 좋아했다
Sentence Percentage: 5.235087e-016
>> 나는 밥을 좋아했다
Sentence Percentage: 2.945077e-015
>> 너는 밥을 좋아했다
Sentence Percentage: 0.000000e+000
>> 노는 밥을 조아해따
Sentence Percentage: 0.000000e+000
>> 사진을 찍으러 공원에 갔다
Sentence Percentage: 2.673730e-026
>> 사진을 찍으로 공원에 갔다
Sentence Percentage: 1.132754e-021
>>
```

- ‘너는 밥을 좋아했다’와 ‘노는 밥을 조아해따’의 경우 각각 ‘아햇’&‘했다’와 ‘해따’의 출현 확률이 0.0이기 때문에 생성확률이 0.0인 것이 확인되었다.

```
명령 프롬프트
>> 노는 법을 조아해따
Sentence Percentage: 0.000000e+000
>> 사진을 찍으려 공원에 갔다
Sentence Percentage: 2.673730e-026
>> 사진을 찍으로 공원에 갔다
Sentence Percentage: 1.132754e-021
>> q
===== Homework 2-2 =====
나는 밥을 좋아했다$      5.235087e-016
나는 밥을 좋아했다$      2.945077e-015
너는 밥을 좋아했다$      3.251097e-015
----- 2-1 Highest Percentage Sentence -----
>> 너는 밥을 좋아했다$, 3.251097e-015
사진을 찍으려 공원에 갔다$      2.673730e-026
사진을 찍으로 공원에 갔다$      1.132754e-021
----- 2-2 Highest Percentage Sentence -----
>> 사진을 찍으로 공원에 갔다$, 1.132754e-021
C:\Users\CYS\Desktop\국민대학교\4학년_1학기\과제\빅데이터_최신기술\과제_2>
```

- 프로그램 내부에서 생성된 문장 중 '나는 밥을 좋아했다'와 '나는 밥을 좋아했다', '너는 밥을 좋아했다'가 각각 가장 높은 확률의 문장으로 대체되었지만 이 중 '너는 밥을 좋아했다'가 가장 높은 확률의 문장이 되었다.
- 마찬가지로 '사진을 찍으려 공원에 갔다'와 '사진을 찍으로 공원에 갔다'가 각각 가장 높은 확률의 문장으로 대체되었지만 이 중 '사진을 찍으로 공원에 갔다'가 가장 높은 확률의 문장이 되었다.

2) gtleee.txt

```
명령 프롬프트 - calculate-sentence-percentage gtleee.txt
C:\Users\CYS\Desktop\국민대학교\4학년_1학기\과제\빅데이터_최신기술\과제_2>gcc -o calculate-sentence-percentage calculate-sentence-percentage.c
C:\Users\CYS\Desktop\국민대학교\4학년_1학기\과제\빅데이터_최신기술\과제_2>calculate-sentence-percentage gtleee.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: 0.000000e+000
>> 노는 법을 조아해따
Sentence Percentage: 0.000000e+000
>> 사진을 찍으려 공원에 갔다
Sentence Percentage: 1.272075e-026
>> 사진을 찍으로 공원에 갔다
Sentence Percentage: 1.383778e-022
>>
```

- '너는 밥을 좋아했다'와 '노는 법을 조아해따'의 경우는 KCC940_Korean_sentences_EUCKR.txt와 같은 이치로 생성 확률이 0.0이 되었고, 각 문장의 생성 확률이 KCC940_Korean_sentences_EUCKR.txt와는 차이가 있음을 확인할 수 있었다.

```
명령 프롬프트
Sentence Percentage: 0.000000e+000
>> 노는 법을 좋아해따
Sentence Percentage: 0.000000e+000
>> 사진을 찍으려 공원에 갔다
Sentence Percentage: 1.272075e-026
>> 사진을 찍으로 공원에 갔다
Sentence Percentage: 1.383778e-022
>> q
===== Homework 2-2 =====
나는 밥을 좋아했다$      2.005390e-014
나는 밥을 좋아했다$      3.075939e-014
----- 2-1 Highest Percentage Sentence -----
>> 나는 밥을 좋아했다$, 3.075939e-014
사진을 찍으려 공원에 갔다$      1.272075e-026
사진을 찍으로 공원에 갔다$      1.383778e-022
----- 2-2 Highest Percentage Sentence -----
>> 사진을 찍으로 공원에 갔다$, 1.383778e-022
C:\Users\CYS\Desktop\국민대학교\4학년 1학기\과제\빅데이터 최신기술\과제 2>
```

- 첫 번째 문장과 두 번째 문장 모두 가장 높은 생성 확률을 가지는 문장이 KCC940_Korean_sentences_EUCKR.txt와는 다른 문장으로 결정되는 것을 확인할 수 있었다.