

# AI개발 실무

## 13. 문장 생성

김 윤 기

교수



13<sub>week</sub>

A I   개 발   실 무   |   김 윤 기

# 문장 생성



- » 문장 생성이 무엇인지 알 수 있다.
- » 마르코프 체인의 개념을 이해할 수 있다.
- » 마르코프 체인으로 문장 생성을 구현할 수 있다.

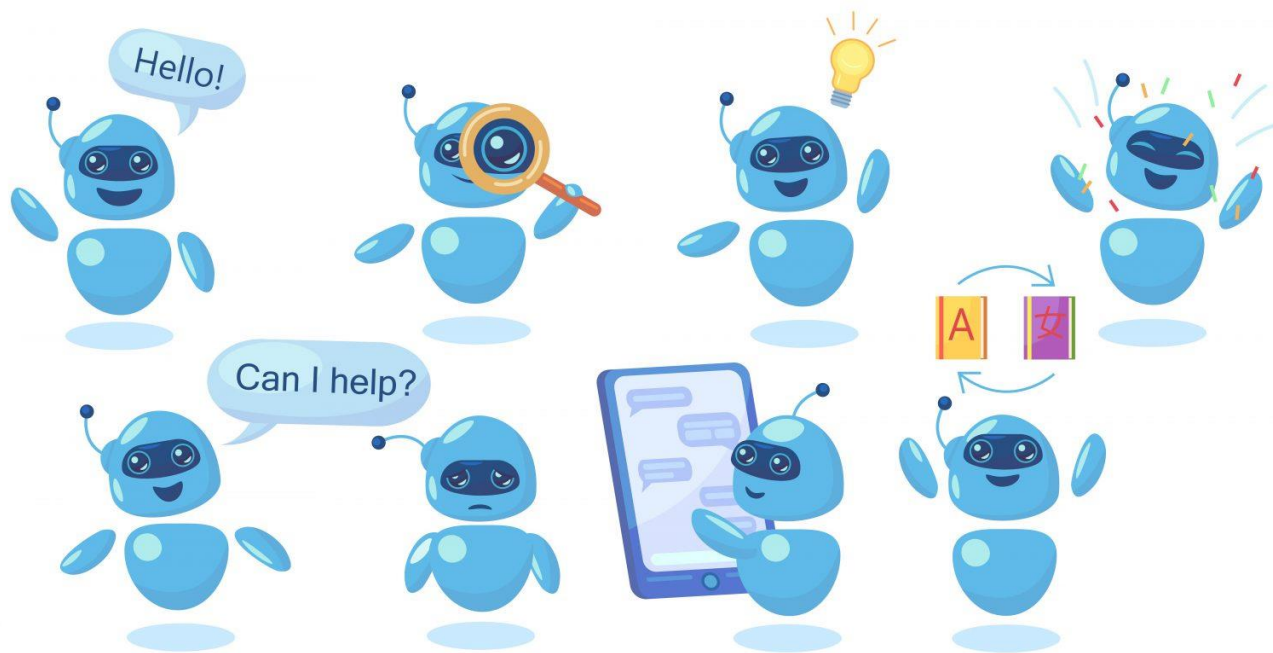
## ① 문장 생성

---

## ② 마르코프 체인을 활용한 문장 생성 실습

---

## 컴퓨터가 사람의 언어를 만들어내는 원리는?



단어의 조합으로 문장이 만들어진다는 사실에 착안하여  
단어의 순서를 고려하여, 새로운 문장을 생성

---

CHAPTER

01

# 문장 생성

## 1. 문장 생성이란?

# 컴퓨터가 인간이 작성한 문장과 유사한 형태로 새로운 문장을 생성하는 것

학습 데이터의 문장을 학습하여, 새로운 문장을 생성

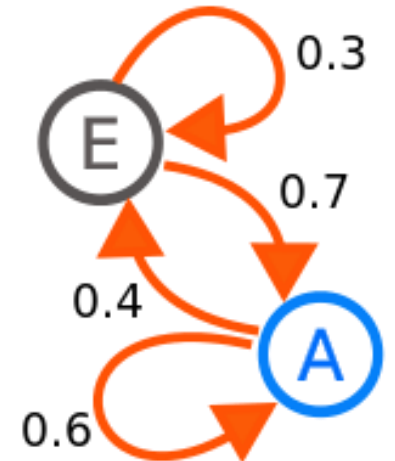
문장 생성 기술은 통계적 기법이나 인공지능 기법으로 구현 가능

통계적 기법은 주어진 문장 데이터를 바탕으로 확률 모델을 학습하여 새로운 문장을 생성함

인공지능 기법은 인공 신경망을 기반으로 주어진 문장을 학습하여 다음 단어를 예측하면서 문장을 생성함

## 2. 마르코프 체인의 개념

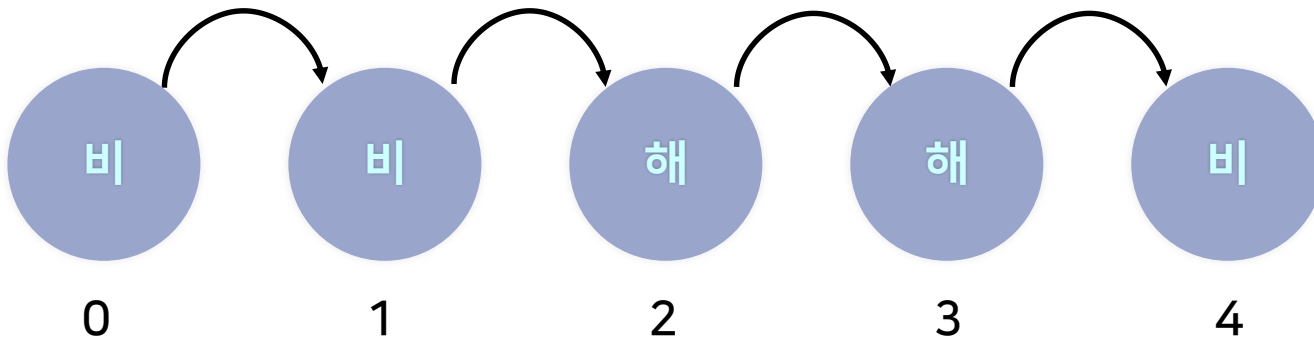
- » '특정 상태의 확률은 오직 최신의 과거 상태에 의존한다'는 마르코프 성질을 지닌 이산확률 과정
- » ' $n+1$ '회의 상태(state)는 ' $n$ '회의 상태나 그 이전 ' $n-1$ '의 상태에 의해 결정된다는 성질을 따름
- » 이전의 상태를 보고, 다음을 예측하는 확률 모형
- » 상태공간{State Space}, 전이확률(Transition Probability)로 마르코프 체인을 구성
- » 상태는 시스템이 취할 수 있는 상태를 나타내며, 전이확률은 한 상태에서 다른 상태로 이동할 확률을 나타냄





## 2. 마르코프 체인의 개념

### 마르코프 체인의 예



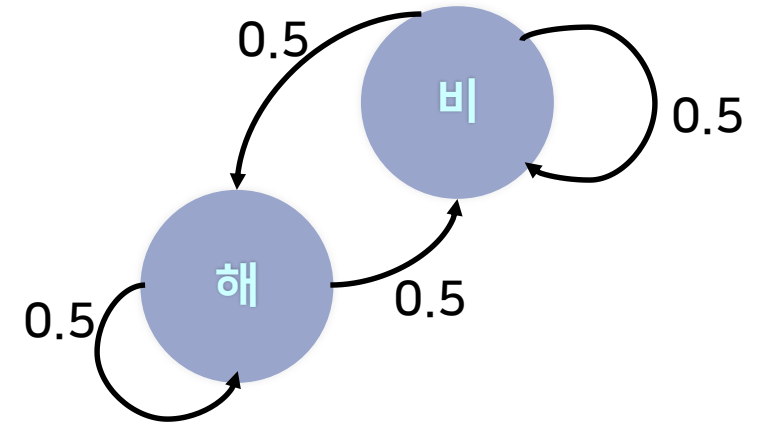
- 상태공간 : {해, 비}
- 상태전이 확률 :

	해	비
해	1	1
비	1	1

전이 빈도

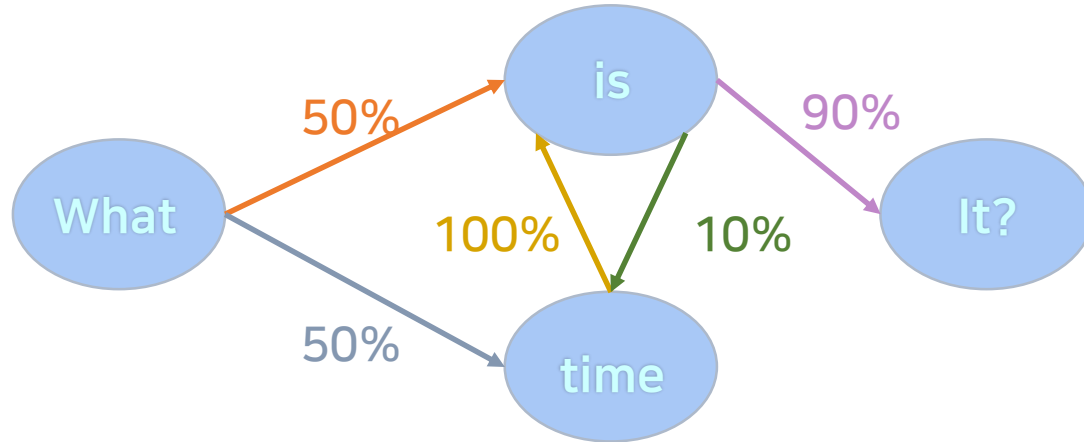
	해	비
해	1/2	1/2
비	1/2	1/2

전이 확률



### 3. 마르코프 체인을 이용한 문장 생성

- » 각 단어를 **상태**로 간주하고 **다음 단어로의 전이 확률**을 구해, 다음 단어를 예측하여 문장을 생성하는 방식을 사용
- » 전이 확률 계산을 위해 텍스트 데이터 수집, 전처리를 수행하여 상태 공간 및 전이 확률을 학습함
- » 단어의 실질적 의미 연관성을 반영하지 않고 문장을 조합하여, 완성도 있는 문장 생성에는 한계가 있음



### 3. 마르코프 체인을 이용한 문장 생성

#### ① 마르코프 체인을 이용한 문장 생성의 과정

1  
텍스트  
데이터 수집

문장 생성에 활용할 학습용 텍스트 셋 수집

2  
데이터  
전처리

수집한 텍스트 데이터를 전처리 하여 사용 가능한 형태로 변환  
이 단계에서는 문장 분리, 토큰화, 불용어 제거 등의 작업을 수행

3  
상태 공간  
생성

전처리 된 텍스트 데이터를 바탕으로 상태 공간을 생성.  
이 상태 공간은 마르코프 체인에서 상태(state)에 해당,  
각 단어는 고유한 상태로 취급

### 3. 마르코프 체인을 이용한 문장 생성

#### ① 마르코프 체인을 이용한 문장 생성의 과정

4

모델  
훈련

생성된 상태 공간을 바탕으로 각 상태 간 전이 확률을 계산.  
각 단어에서 다음 단어로의 확률을 나타냄

5

문장  
생성

훈련된 모델을 이용하여 새로운 문장을 생성.  
이 때, 모델이 예측한 다음 단어를 이용하여 새로운 문장을 구성.

예

"나는"이라는 단어 다음에는 "오늘" 또는 "내일" 등이 나올 확률이 높다면,  
모델은 "나는 오늘" 또는 "나는 내일"이라는 문장을 생성할 수 있음

## 3. 마르코프 체인을 이용한 문장 생성

### 📦 마르코프 체인을 이용한 문장 생성의 예

#### ① 텍스트 데이터 수집

ex

나는 밥을 먹었다.  
그는 밥을 좋아한다.  
나는 영화를 봤다.

#### ② 텍스트 데이터 전처리 : 구두점 제거

ex

나는 밥을 먹었다  
그는 밥을 좋아한다  
나는 영화를 봤다

## 3. 마르코프 체인을 이용한 문장 생성

② 마르코프 체인을 이용한 문장 생성의 예

### ③ 상태 공간 생성

ex

{나는, 밥을, 먹었다, 그는, 좋아한다, 영화를, 봤다}

### 3. 마르코프 체인을 이용한 문장 생성

⑥ 마르코프 체인을 이용한 문장 생성의 예

#### ④ 모델 훈련

	나는	밥을	먹었다	그는	좋아한다	영화를	봤다
나는	0	0.5	0	0	0	0.5	0
밥을	0	0	0.5	0	0.5	0	0
먹었다	0	0	0	0	0	0	0
그는	0	1	0	0	0	0	0
좋아한다	0	0	0	0	0	0	0
영화를	0	0	0	0	0	0	1
봤다	0	0	0	0	0	0	0

### 3. 마르코프 체인을 이용한 문장 생성

④ 마르코프 체인을 이용한 문장 생성의 예

- ⑤ 문장 생성 : 전이 확률대로 다음 상태를 결정할 수 있으며,  
다양한 문장 생성을 위해서 동일한 확률로 다음 단어를 선택할 수 있음

ex

나는 → 밥을 → 좋아한다.      나는 밥을 좋아한다.



### 3. 마르코프 체인을 이용한 문장 생성

#### N-gram을 활용한 마르코프 체인

- » 텍스트에서 n개의 연속적인 토큰(단어, 문자 등)을 사용하여 언어 모델을 구성하는 **N-gram을 마르코프 체인에 적용**
- » n개의 연속적인 단어를 마르코프 체인의 **상태**로 간주
- » N을 어떻게 설정하는 지에 따라 문장 생성의 성능이 달라질 수 있음

ex

나는 밥을 먹었다.  
그는 밥을 좋아한다.



ex

{(나는, 밥을), (밥을, 먹었다) (그는, 밥을) (밥을, 좋아한다)}

### 3. 마르코프 체인을 이용한 문장 생성

 N-gram을 활용한 마르코프 체인

	먹었다	좋아한다
{나는, 밥을}	1	0
{밥을, 먹었다}	0	0
{그는, 밥을}	0	1
{밥을, 좋아한 다}	0	0

### 3. 마르코프 체인을 이용한 문장 생성

#### 마르코프 체인의 구현

##### 마르코프 체인 구조

```
markov_chain = defaultdict(list) # 딕셔너리의 값을 리스트로 초기화
state_counts = defaultdict(int) # 단어 등장 빈도를 0으로 초기화

# markov chain 의 구조 : prefix- 현재 단어, suffix - prefix 뒤에 출현할 단어
# {(prefix):[suffix]}
# {('개봉', '되다니'):[민어지지, 기빠요, 실망임]}

# state_counts 의 구조
# {((('개봉', '되다니'),'민어지지')):1}
# {((('개봉', '되다니'),'기빠요')):2}
# {((('개봉', '되다니'),'실망임')):1}
```

### 3. 마르코프 체인을 이용한 문장 생성

#### 마르코프 체인의 구현

##### 단어 토큰 생성

```
prefix = tuple(words[i: i + order]) # 차수 만큼 문장을 슬라이싱하여, 튜플 자료형에 저장  
suffix = words[i + order] # prefix 다음에 나올 단어를 저장
```

```
# 마르코프 체인의 딕셔너리 사전의 key로 prefix, value로 suffix를 저장  
# {'(개봉', '되다니)': ['믿어지지', '기뻐요', '실망임']}
```

```
markov_chain[prefix].append(suffix)
```

```
# 단어 등장 빈도를 statecounts에 저장  
# {'(('개봉', '되다니'), '믿어지지'): 1}  
# {'(('개봉', '되다니'), '기뻐요'): 2}  
# {'(('개봉', '되다니'), '실망임'): 1}  
state_counts[(prefix, suffix)] += 1
```

### 3. 마르코프 체인을 이용한 문장 생성

#### 문장 생성

##### 문장 생성

```
current_prefix = random.choice(list(markov_chain.keys())) # 첫 단어 랜덤 선택  
sentence = list(current_prefix) # 새롭게 생성할 문장 초기화
```

```
# 빈도에 따른 확률 가중치 계산
```

```
weights = [state_counts[(current_prefix, suffix)] for suffix in next_words]
```

```
# 다음 단어 선택
```

```
next_word = random.choices(next_words, weights=weights)[0]
```



개 발 실 무  
**실습하기**

P r a c t i c a l P r o g r a m m i n g f o r A I

문장 생성 실습

## ① 문장 생성이란?

컴퓨터가 인간이 작성한 문장과 유사한 형태로 새로운 문장을 생성하는 것

## ② 마르코프 체인의 개념

이전의 상태를 보고, 다음을 예측하는 확률 모형

상태공간과 전이확률로 마르코프 체인을 구성

### ③ 마르코프 체인을 이용한 문장 생성

각 단어를 상태로 간주하고  
다음 단어로의 전이 확률을 구해,  
다음 단어를 예측하여 문장을 생성

텍스트 데이터 수집 → 데이터 전처리 →  
상태공간 생성 → 모델 훈련 → 문장 생성

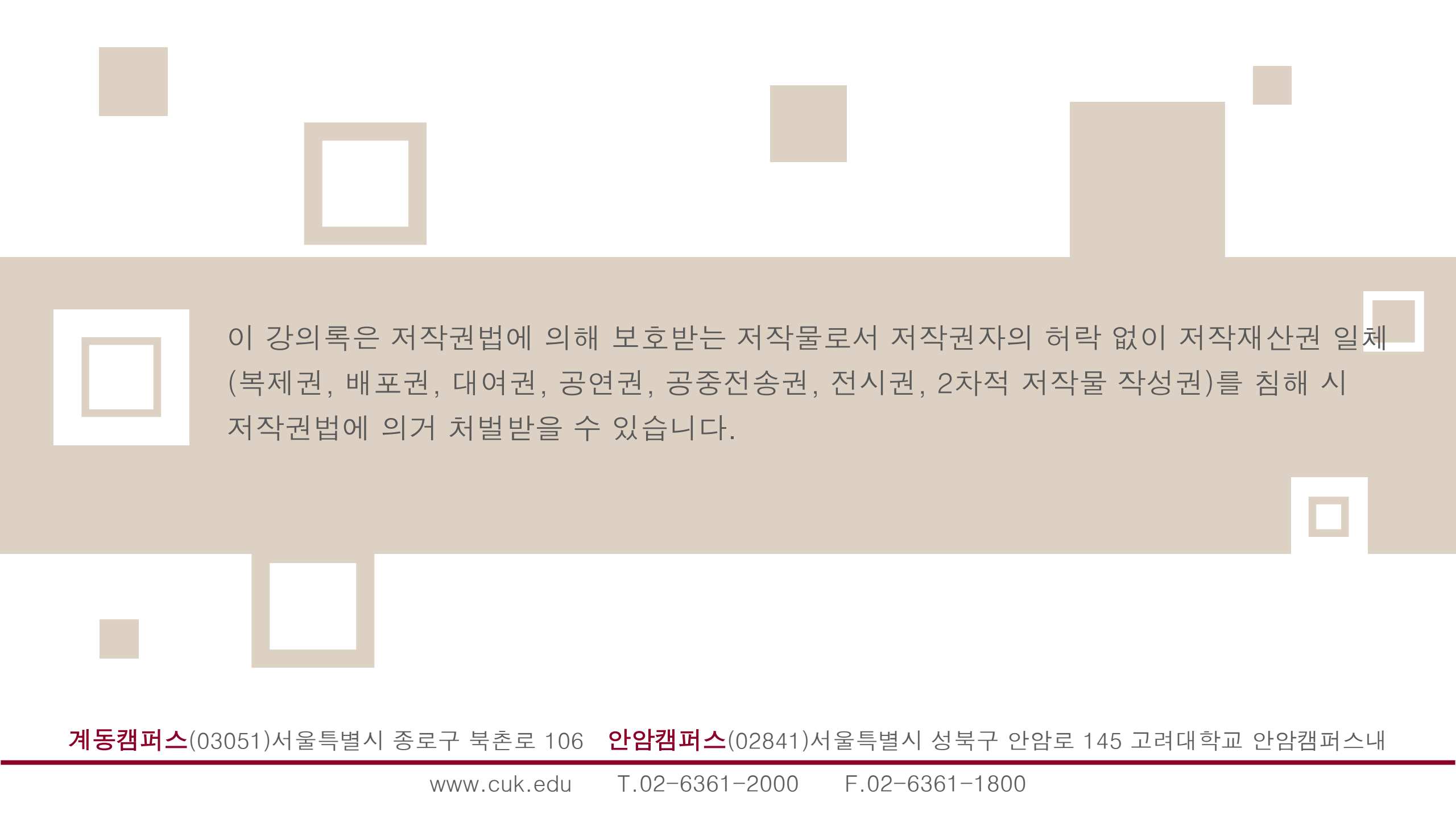
### ④ 문장 생성 실습

마르코프 체인을 이용한 문장 생성



## 참고문헌

📄 파이썬 머신러닝, 딥러닝 실전 개발 입문, 위키북스, 쿠지라 히코우즈쿠에



이 강의록은 저작권법에 의해 보호받는 저작물로서 저작권자의 허락 없이 저작재산권 일체 (복제권, 배포권, 대여권, 공연권, 공중전송권, 전시권, 2차적 저작물 작성권)를 침해 시 저작권법에 의거 처벌받을 수 있습니다.