

AI개발 실무

14. 챗봇

김 윤 기 교수



14 week

A I 개 발 실 무 | 김 윤 기

챗봇

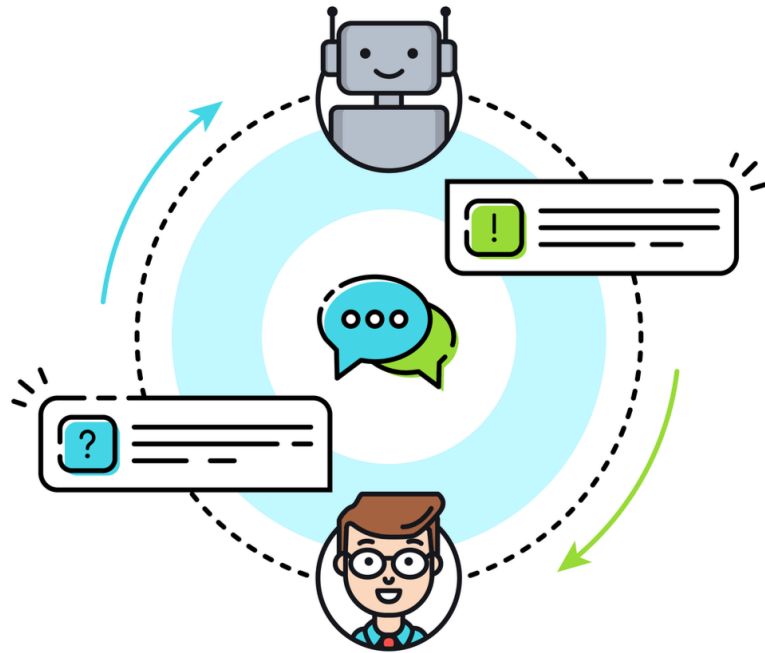


- » TF-IDF의 개념을 이해할 수 있다.
- » 코사인 유사도의 개념을 설명할 수 있다.
- » 챗봇을 구현할 수 있다.

① 챗봇의 원리

② 챗봇 구현 실습

챗봇은 어떤 원리로 구현될까?



사용자의 질문과 유사한 질의에 대한 답변을 찾아
가장 적절한 답변으로 응답

CHAPTER

01

챗봇의 원리

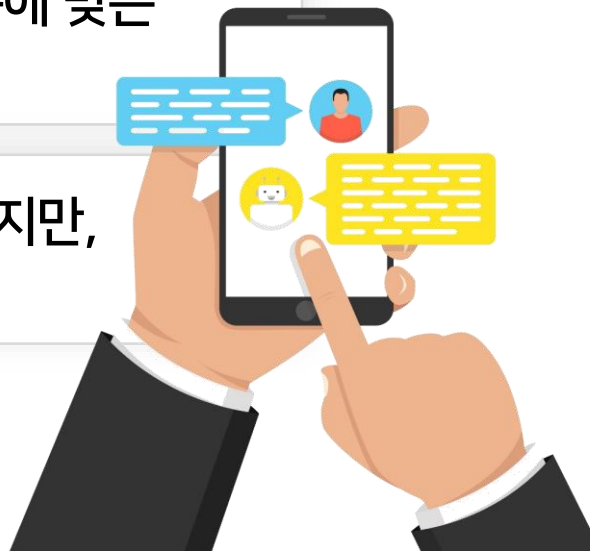
1. 챗봇이란?

인공지능 기술을 이용하여 자연어로 된 질문에 대해 답변을 제공하는 프로그램

채팅 형식으로 사용자와 대화를 나누며, 자연어 처리 기술과 기계 학습 알고리즘을 이용하여 사용자의 의도를 파악하고 적절한 답변을 제공

데이터 수집을 통해 질문, 답변을 학습하고 질문에 맞는 최적의 답을 구함

대화의 흐름에 따라 상황에 맞는 대화를 제공하지만, 사용자의 감정이나 의도를 파악하기는 어려움



2. 챗봇의 처리 단계

사용자 입력

사용자가 질문한 내용을 입력 단계
사용자의 입력은 자연어 형태로 입력

입력된 내용 분석

분석하여 그에 맞는 처리를 수행
자연어 처리 기술과 기계 학습 알고리즘이 사용

적절한 응답 생성

챗봇은 사용자의 의도에 맞는 적절한 응답을 생성
단순한 텍스트 응답 뿐만 아니라 이미지, 음성 등
다양한 형태의 응답이 가능

응답 출력

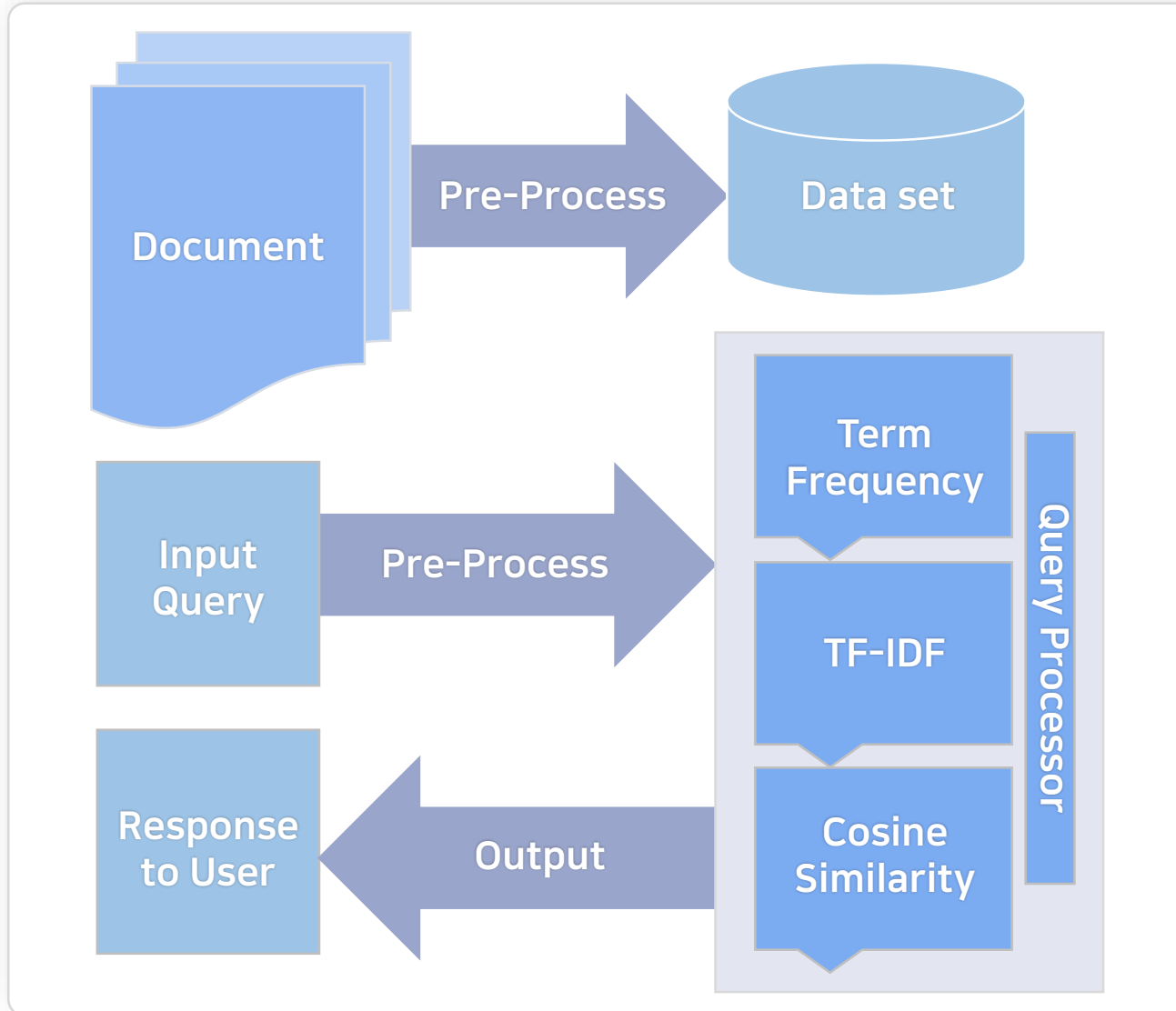
생성된 응답을 출력하여 사용자에게 제공

3. 챗봇의 구조



**질문과 유사한 질의에 대한 답변을 찾아
가장 적절한 답변으로 응답**

4. 챗봇 실습 개요



5. 단어 수치화

① 문서 단어 행렬(Document Term Matrix, DTM)

**다수의 문서에서 등장하는 각 단어들의 빈도를
행렬로 표현하여 문서를 수치화 한 것**

행과 열이 바뀌면 **단어 문서 행렬(Term Document Matrix, TDM)**
이라고 하기도 함

문서들을 **서로 비교**할 수 있도록 **수치화** 할 수 있다는 의의가 있음

빈도수가 높은 **불필요한 단어들이**
중요한 단어로 인식될 수 있는 **문제점**이 존재

5. 단어 수치화

① 문서 단어 행렬(Document Term Matrix, DTM)

문서 단어 행렬 예시

		고프다	너무	다리가	배가	비가	아프다	와서	저기	지나간다
	A : 배가 너무 너무 아프다	A	0	2	0	1	0	1	0	0
	B : 배가 너무 고프다	B	1	1	0	1	0	0	0	0
	C : 저기 배가 지나간다	C	0	0	0	1	0	0	1	1
	D : 비가 와서 다리가 아프다	D	0	0	1	0	1	1	0	0

5. 단어 수치화

② TF-IDF(Term Frequency-Inverse Document Frequency)

**단어의 빈도(TF)와 역 문서 빈도(IDF)를 사용하여
DTM내의 각 단어들마다 중요한 정도를 가중치로 주는 방법**

TF와 IDF를 곱한 값으로 점수가 높은 단어일수록 다른 문서에는 많지 않고 해당 문서에서 자주 등장하는 단어를 의미함

벡터 형태이므로 후에 인공지능망의 입력으로도 사용할 수 있음

5. 단어 수치화

② TF-IDF(Term Frequency-Inverse Document Frequency)

$tf(d,t)$ (Term Frequency)

단어의 빈도. DTM에서 각 단어들이 가진 값

$df(t)$ (Document Frequency)

- 특정 단어 t 가 등장한 문서의 수
- 특정 단어가 문서에서 몇 번 등장 했는지는 고려하지 않음

5. 단어 수치화

② TF-IDF(Term Frequency-Inverse Document Frequency)

idf(d,t) (Inverse Term Frequency)

$$\text{Idf}(d,t) = \log \left(\frac{n}{1+\text{df}(t)} \right)$$

- 기본적인 개념은 df의 역수
- n : 문서의 갯수
- **Log 사용 이유** : 문서의 갯수 n이 커질 수록 IDF 값이 급격하게 증가
- **1을 더한 이유** : 분모가 0이 되지 않도록 하기 위해

5. 단어 수치화

② TF-IDF(Term Frequency-Inverse Document Frequency)

예제의 IDF

	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

$$\text{Idf}(d,t) = \log \left(\frac{n}{1+\text{df}(t)} \right)$$

문서 1 : 먹고 싶은 사과

문서 2 : 먹고 싶은 바나나

문서 3 : 길고 노란 바나나 바나나

문서 4 : 저는 과일이 좋아요.

단어	IDF(역 문서 빈도)
과일이	$\ln(4 / (1 + 1)) = 0.693147$
길고	$\ln(4 / (1 + 1)) = 0.693147$
노란	$\ln(4 / (1 + 1)) = 0.693147$
먹고	$\ln(4 / (2 + 1)) = 0.287682$
바나나	$\ln(4 / (2 + 1)) = 0.287682$
사과	$\ln(4 / (1 + 1)) = 0.693147$
싶은	$\ln(4 / (2 + 1)) = 0.287682$
저는	$\ln(4 / (1 + 1)) = 0.693147$
좋아요	$\ln(4 / (1 + 1)) = 0.693147$

5. 단어 수치화

② TF-IDF(Term Frequency-Inverse Document Frequency)

예제의 TF-IDF

TF										IDF	
	과일이	길고	노란	먹고	바나나	사과	싫은	저는	좋아요	단어	IDF(역 문서 빈도)
문서1	0	0	0	1	0	1	1	0	0	과일이	$\ln(4 / (1 + 1)) = 0.693147$
문서2	0	0	0	1	1	0	1	0	0	길고	$\ln(4 / (1 + 1)) = 0.693147$
문서3	0	1	1	0	2	0	0	0	0	노란	$\ln(4 / (1 + 1)) = 0.693147$
문서4	1	0	0	0	0	0	0	1	1	먹고	$\ln(4 / (2 + 1)) = 0.287682$
										바나나	$\ln(4 / (2 + 1)) = 0.287682$
										사과	$\ln(4 / (1 + 1)) = 0.693147$
										싫은	$\ln(4 / (2 + 1)) = 0.287682$
										저는	$\ln(4 / (1 + 1)) = 0.693147$
										좋아요	$\ln(4 / (1 + 1)) = 0.693147$
	과일이	길고	노란	먹고	바나나	사과	싫은	저는	좋아요		
문서 1	0	0	0	0.287682	0	0.693147	0.287682	0	0		
문서 2	0	0	0	0.287682	0.287682	0	0.287682	0	0		
문서 3	0	0.693147	0.693147	0	0.575364	0	0	0	0		
문서 4	0.693147	0	0	0	0	0	0	0.693147	0.693147		

6. 유사 문장 검색

① 문서의 유사도 검색

➔ 코사인 유사도(Cosine Similarity)란?

**두 벡터 간의 코사인 각도를 이용하여
구할 수 있는 두 벡터의 유사도**

두 벡터의 방향이 완전히 동일한 경우에는 1의 값을, 수직을 이루면 0, 180도로 반대의 방향을 가지면 -1의 값을 갖게 됨

-1부터 1 사이의 값을 가지며 값이 1에 가까울 수록 유사도가 높다고 판단함

6. 유사 문장 검색

① 문서의 유사도 검색

➔ 코사인 유사도(Cosine Similarity)



➔ 코사인 유사도 수식

$$\text{cosine similarity} = (A \cdot B) / (\|A\| \|B\|)$$

6. 유사 문장 검색

① 문서의 유사도 검색

➔ 문서 간 코사인 유사도 계산해보기

📦 문서 단어 행렬 예시

문장 1의 벡터 $A = [0, 0.3, 0, 0, 0]$

문장 2의 벡터 $B = [0, 0, 0, 0, 0.3]$

$$(A \cdot B) = 0 * 0 + 0.3 * 0 + 0 * 0 + 0 * 0 + 0 * 0.3 = 0$$

$$\|A\| \|B\| = \sqrt{0^2 + 0.3^2 + 0^2 + 0^2 + 0^2} * \sqrt{0^2 + 0^2 + 0^2 + 0^2 + 0.3^2}$$

$$(A \cdot B) / (\|A\| \|B\|) = \frac{0}{0.09} = 0$$



개발 실무
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

챗봇의 원리 실습

① 챗봇이란?

챗봇은 인공지능 기술을 이용하여 자연어로 된 질문에 대해 답변을 제공하는 프로그램

② 챗봇의 처리 단계

사용자 입력 → 입력된 내용 분석 →
적절한 응답 생성 → 응답 출력

③ 챗봇의 구조

질문과 유사한 질의에 대한 답변을 찾아
가장 적절한 답변으로 응답

④ 챗봇 실습 개요

텍스트를 TF-IDF로 수치화한 뒤,
Cosine-Similarity를 통해 유사 답변을 구함

⑤ 단어 수치화

TF-IDF를 통해 문장의 출현 빈도를 기반으로
단어 수치화

⑥ 유사 문장 검색

Cosine-Similarity를 통해
수치화 된 문장간 유사도를 계산

⑤ 단어 수치화

TF-IDF를 통해 문장의 출현 빈도를 기반으로
단어 수치화

⑥ 유사 문장 검색

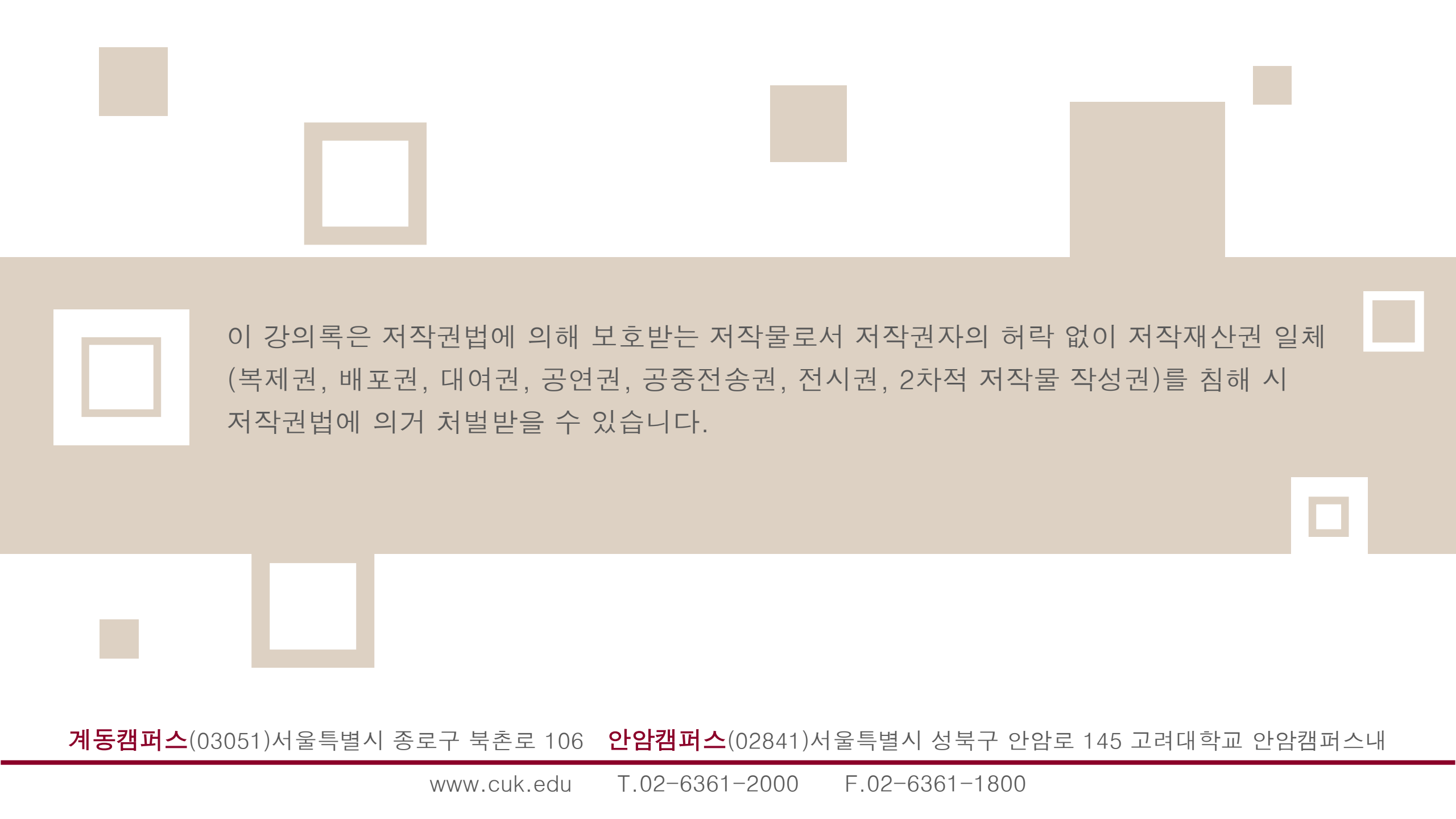
Cosine-Similarity를 통해
수치화 된 문장간 유사도를 계산

⑦ 챗봇 구현 실습

TF-IDF 및 Consine 유사도를 이용한
챗봇 구현

참고문헌

 딥러닝을 이용한 자연어 처리 입문(<https://wikidocs.net/book/2155>)



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