

AI 개발 실무

12. 문장 유사도 분석

김 윤 기 교수



12_{week}

A I 개 발 실 무 | 김 윤 기

문장 유사도 분석



- » 문장 유사도 분석이 무엇인지 알 수 있다.
- » 레벤슈타인 거리를 이용해 문장 유사도를 측정할 수 있다.
- » N-gram을 이용해 문장 유사도를 측정할 수 있다.

① 문장 유사도 분석

② 문장 유사도 분석 실습

형태가 다르지만 같은 의미의 문장

나는 { 서울특별시
서울시 } 에 살아요.
서울

텍스트의 형태는 다르지만 유사한 의미의 문장들을
찾아 문서 분류, 도용 검사 등에 활용

CHAPTER

01

문장 유사도 분석

1. 문장 유사도 분석이란?

자연어와 자연어 사이의 의미적 유사성을 점수화 하여 유사도를 측정하는 것

비교하고자 하는 자연어 문장 간 의미론적으로 얼마나 유사한 지를 컴퓨터가 분석할 수 있어야 함

어순과 단어가 서로 다르더라도 유사한 의미를 나타내는 수 많은 형태의 문장을 컴퓨터가 분석할 수 있도록 하는 것이 목적

기계 번역, 텍스트 요약, 질문 응답 시스템, 검색 엔진 등의 분야에서도 유사도 분석을 활용하여 자연어 처리의 품질을 향상 시킴

2. 문장 유사도 분석의 활용 분야

검색의 정확도 향상

정보 검색 시 입력 텍스트의 형태에만 의존하지 않고,
검색하고자 하는 의도를 파악하여 정확한 검색에 도움을 줌

도용 판별

논문, 도서 등의 지적 재산권의 무단 사용을 분석

문서 분류

비슷한 의미를 가지는 문서들을 카테고리화

추천 시스템

도서, 영화를 줄거리를 기반으로 선호도 분석을 하고,
비슷한 내용의 다른 콘텐츠를 추천

3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

- » 두 문자열 간의 편집 거리를 측정하여 문자열 간의 유사성을 측정
- » 편집 거리란, 한 문자열을 다른 문자열로 변환하는 데 필요한 최소 편집 연산의 횟수를 의미함
- » 편집 연산의 종류

삽입
(Insertion)

한 문자열에
새로운 문자를 삽입하는 것

삭제
(Deletion)

한 문자열에서 문자를
제거하는 것

교체
(Replacement)

한 문자열의 문자를
다른 문자로 교체하는 것

3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

📦 레벤슈타인 거리 알고리즘의 예



3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

📦 레벤슈타인 거리 알고리즘의 예

1. 행과 열로 각각의 비교군을 배치하고 숫자를 순서대로 입력

		유 사 도 나 분 석 할 까 요											초 기 화
열	0	1	2	3	4	5	6	7	8	9	10	11	
	1												
마 나	2												
	3												
	4												
	5												
분 석 이	6												
	7												
	8												
	9												
될 까 요	10												
	11												

3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

📦 레벤슈타인 거리 알고리즘의 예

2. 다음으로 한 글자씩 비교하며 삽입, 변경, 삭제 여부를 결정

- ❶ 글자가 서로 동일하면 대각선의 숫자를 가져옴
- ❷ 변경이 필요하면 대각선 수의 +1
- ❸ 삽입이 필요한 경우 왼쪽 수에서 +1
- ❹ 삭제가 필요한 경우 위쪽 수에서 +1
- ❺ 1 ~ 4까지의 숫자 중 가장 낮은 수를 입력

3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

② 레벤슈타인 거리 알고리즘의 예

2. 다음으로 한글자씩 비교하며 삽입, 변경, 삭제 여부를 결정

삽입

변경

삭제

		유 사 도 나 분 석 할 까 요													초 기 화
		0	1	2	3	4	5	6	7	8	9	10	11		
열 마 나 분 석 이 될 까 요	1		1	변경은 대각선의 + 1											
	2			2	삽입은 왼쪽수의 + 1										
	3					3									
	4						3								
	5							3							
	6								3						
	7									4	삭제는 위의 수의 + 1				
	8										4				
	9											5	변경은 대각선의 + 1		
	10												5		
	11													5	

삽입

변경

삭제

3. 유사도 분석 알고리즘

① 레벤슈타인 거리(Levenshtein Distance)

📦 레벤슈타인 거리 알고리즘의 예

3. 모든 경우의 수를 반복하면 비교하면, 최종적으로 마지막 수가 비용값이 됨

삽입

변경

삭제

		유 사 도 나					분 석		할 까 요			
열 마 나 분 석 이 될 까 요	0	1	2	3	4	5	6	7	8	9	10	11
	1	1	2	3	4	5	6	7	8	9	10	11
	2	2	2	3	4	5	6	7	8	9	10	11
	3	3	3	3	4	5	6	7	8	9	10	
	4	4	4	4	4	5	6	7	8	9		
	5	5	5	5	5	4	5	6	7	8		
	6	6	6	6	6	5	4	5	6	7		
	7	7	7	7	7	6	5	4	5	6	7	
	8	8	8	8	8	7	6	5	4	5	6	7
	9	9	9	9	9	8	7	6	5	6	7	
	10	10	10	10	10	9	8	7	6	6	5	6
	11	11	11	11	11	10	9	8	7	7	6	5

삽입

변경

삭제

3. 유사도 분석 알고리즘

② N-gram

» 연속적인 단어의 나열을 N개씩 끊어서 이를 하나의 토큰으로 간주하여 비교

- $N=1 \rightarrow$ 유니그램
- $N=2 \rightarrow$ 바이그램
- $N=3 \rightarrow$ 트라이그램
- $N=4$ 이상
→ 앞에 그대로 숫자를 붙여서 부름

» N 값을 조정함으로써, 얻을 수 있는 정보의 양과 질을 조절할 수 있음

» N 값이 너무 작으면 정보의 양이 적어 질 수 있고, 너무 크면 정보가 희소해져 적절한 n 값을 찾는 것이 중요

3. 유사도 분석 알고리즘

② N-gram

N gram 예시

My dream is having cute baby	
unigram	My, dream, is, having, cute, baby
bigram	My dream, dream is, is having, having cute, cute baby
trigram	My dream is, dream is having, is having cute, having cute baby
4-gram	My dream is having, dream is having cute, is having cute baby

3. 유사도 분석 알고리즘

② N-gram

» N-gram을 이용한 유사도 측정 방법

ex "I love cats" vs "I love dogs"

① 두개의 문장을 n-gram으로 분석

- "I love cats"의 2-gram 집합: {"I ", " l", "lo", "ov", "ve", "e ", " c", "ca", "at", "ts"}
- "I love dogs"의 2-gram 집합: {"I ", " l", "lo", "ov", "ve", "e ", " d", "do", "og", "gs"}

3. 유사도 분석 알고리즘

② N-gram

» N-gram을 이용한 유사도 측정 방법

ex "I love cats" vs "I love dogs"

② 두 개의 집합 간에 공통으로 나타나는 n-gram의 수를 측정

- {"I ", " I", "lo", "ov", "ve", "e "}가 공통으로 나타나므로 공통 n-gram 수치는 6



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

문장 유사도 분석 실습

① 문장 유사도 분석이란?

자연어와 자연어 사이의 의미적 유사성을
점수화하여 유사도를 측정하는 것

② 문장 유사도 분석의 활용 분야

검색의 정확도 향상

도용 판별

문서 분류

추천 시스템

③ 문서 유사도 분석 알고리즘

레벤슈타인 거리

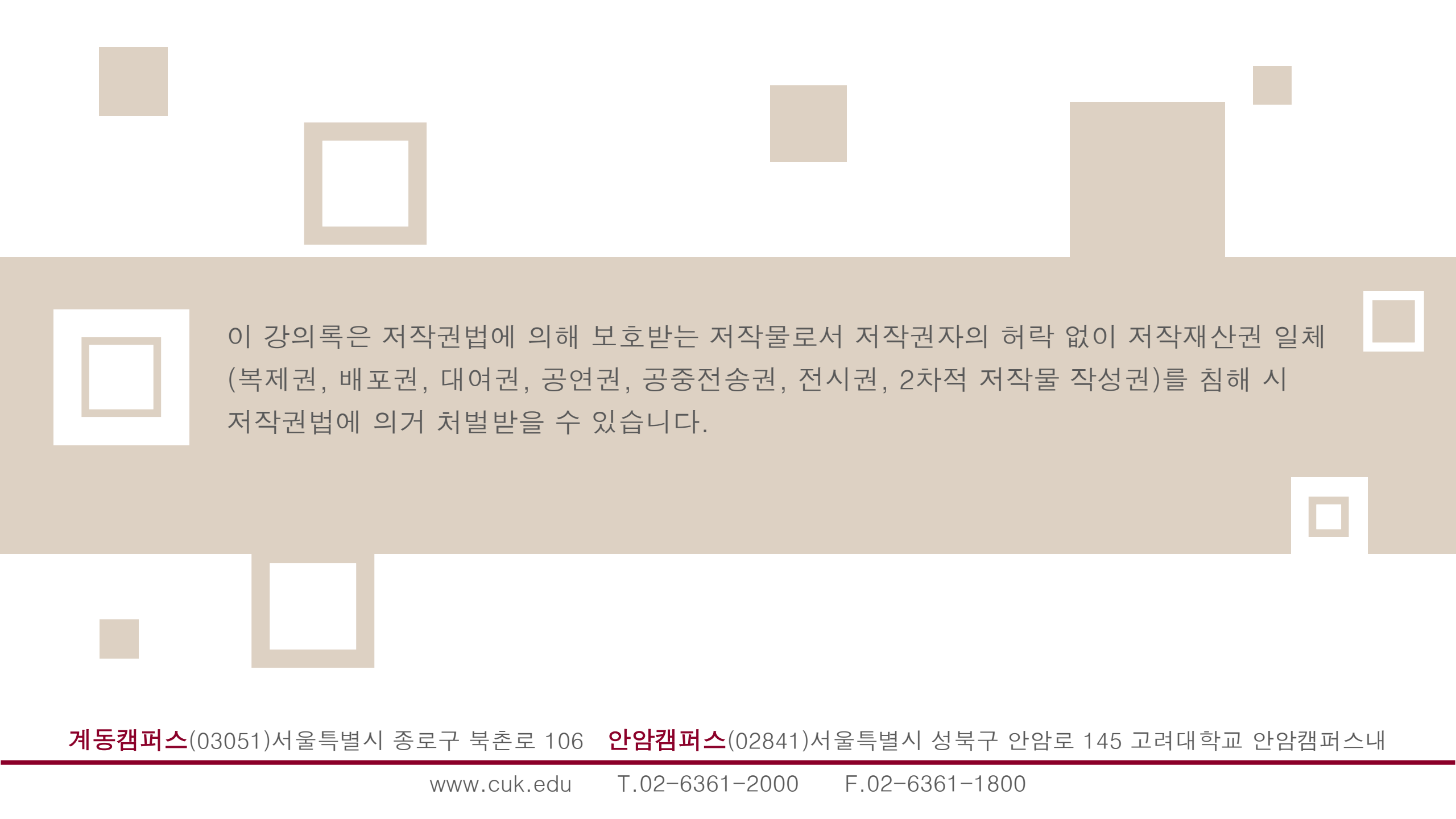
N-gram

④ 문서 유사도 분석 실습

레벤슈타인 거리 및 N-gram

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.