《 기하 발표 주제 》

하이에듀

주제	**텍스트 유사도 비교**
요약	텍스트 유사도에 대한 기본적인 개념을 확인하고, 다양한 방법으로 텍스트 유사도를 구하는 방법들을 살펴봅니다. 그 뒤, 텍스트 유사도의 비교가 산업공학과 통계학에서 어떻게 쓰이는지 확인해보면 좋을 것 같습니다.

자료 1. 텍스트 유사도

텍스트 유사도(Text Similarity)는 두 개 이상의 텍스트 콘텐츠가 얼마나 서로 유사한지를 정량적으로 측정하는 방법입니다. 이는 자연어 처리(NLP)의 중요한 개념 중 하나로, 다양한 언어 관련 애플리케이션에서 광범위하게 사용됩니다.

측정 방법: 텍스트 유사도를 측정하는 방법에는 여러 가지가 있으며, 각 방법은 텍스트를 다루는 방식에 따라 다릅니다. 여기에는 주로 두 가지 접근 방식이 포함됩니다:

벡터 공간 모델: 텍스트를 수치적 벡터로 변환하고, 이 벡터들 사이의 유사도를 측정합니다. 대표적인 방법으로는 코사인 유사도, 유클리드 거리, 자카드 유사도 등이 있습니다.

시맨틱 분석: 텍스트의 의미를 분석하여 유사도를 측정합니다. 이는 종종 머신 러닝, 특히 딥러닝 모델을 사용하여 구현됩니다.

응용 분야: 텍스트 유사도는 다양한 분야에서 활용됩니다. 예를 들어, 검색 엔진은 사용자의 쿼리와 관련된 문서를 찾기 위해 유사도를 측정합니다. 또한, FAQ 시스템, 플래그리아리즘 검출, 문서 분류, 감성 분석 등에서도 중요한 역할을 합니다.

도전 과제: 텍스트 유사도 측정은 복잡할 수 있습니다. 다른 단어가 같은 의미를 가질 수 있으며(예: 동의어), 같은 단어가 다른 맥락에서 다른 의미를 가질 수도 있습니다(예: 다의어). 또한, 문맥, 비유, 은유 등 언어의 복잡한 특성을 적절히 처리하는 것이 중요합니다.

기술 발전: 최근에는 딥러닝 기반의 NLP 모델(예: BERT, GPT)이 등장하면서, 텍스트의 의미를 더 깊이 이해하고 복잡한 언어적 특성을 처리할 수 있게 되었습니다. 이러한 모델들은 문맥적 의미를 파악하는 데 탁월하여, 텍스트 유사도 측정의 정확도를 향상시키고 있습니다.

❖ 유사도

• 두 텍스트 사이의 유사한 정도를 수치화한 것
• '유사한 정도'는 주관적 기준이므로 이를 정량화하여 분석해야 함
• 유사도를 수치화(계산)하는 방법
▪ 같은 단어의 개수로 판단하는 방법
▪ 형태소 또는 자소 단위로 나누어 비교하는 방법
- 형태소 : 뜻을 가진 가장 작은 말의 단위
- 자소 : 한 언어의 문자 체계에서 음소를 표시하는 최소의 변별적 단위로서의 문자 혹은 문자 결합
▪ 벡터로 나타내어 두 벡터 사이의 거리나 각 등을 이용하는 방법

• 유사도의 종류
▪ 자카드 유사도
▪ 유클리드 유사도
▪ 코사인 유사도
▪ 맨해튼 유사도

자료 2. 유사도를 구하는 방법

❖ 유사도를 구하는 방법• 수학 내용 요소와 유사도의 관계
▪ 집합 → 자카드 유사도
▪ 두 점 사이의 거리 → 유클리드 유사도
▪ 벡터 → 코사인 유사도

❖ 자카드 유사도
• 두 텍스트 P, Q를 각각 단어 집합으로 표현한 뒤집합을 통해 유사도를 측정하는 방법(집합 기반 유사도)
• 두 텍스트 P와 Q를 구성하는 단어의 집합이 각각 P, Q일 때,P와 Q의 자카드 유사도를 기호로 J(P,Q)와 같이 나타내고,P와 Q의 교집합의 원소의 개수를 P와 Q의 합집합의 개수를 나눈 값으로 구함

❖ 자카드 유사도
• 항상 0 이상 1 이하의 값을 가짐
• 교집합의 원소의 개수가 많을수록 유사함→ 자카드 유사도의 값이 1에 가까울수록 두 텍스트가 유사하다고 판단
• 예) 텍스트 P의 단어 집합 : 오늘, 날씨, 비, 우산, 조심
텍스트 Q의 단어 집합 : 오늘, 날씨, 기온, 낮음, 조심
→ P와 Q 사이의 자카드 유사도는 3/7

❖ 유클리드 유사도

• 두 텍스트 P, Q를 각각 벡터로 표현한 뒤 두 점 사이의 거리를 이용하여 유사도를 측정하는 방법(유클리디안 유사도, 거리기반 유사도)
• 두 텍스트 P와 Q를 나타내는 벡터가 각각 P, Q일 때, P와 Q의 유클리드 유사도를 기호로 L(P, Q)와 같이 나타내고 P와 Q의 각 성분의 차의 제곱의 합의 양의 제곱근으로 구함
• 텍스트 P와 Q의 벡터를 좌표평면에 나타내면 P와 Q의 유사도는 좌표평면 위의두 점 사이의 거리와 같음→ 유클리드 유사도의 값이 0에 가까울수록 두 텍스트가 유사하다고 판단
• 예) 두 텍스트 P와 Q를 나타내는 벡터가 P=(4, 1), Q=(1, 5) 일 때, 두 텍스트의 유클리드 유사도는 5=두 점 (1, 5)와 (4, 1) 사이의 거리

❖ 코사인 유사도
• 두 텍스트 P, Q를 각각 벡터로 표현한 뒤 두 벡터 사이의 각도를 이용하여 유사도를 측정하는 방법
• 두 텍스트 P와 Q를 나타내는 벡터가 P, Q일 때, P와 Q의 코사인 유사도를 기호로 C(P, Q)와 같이 나타내고, P와 Q의 내적을 P와 Q의 크기로 나누어서 구함
→ 내적의 개념을 설명하지 않고, 두 벡터의 같은 위치의 성분의 곱들의 합으로 지도
• 코사인 유사도의 값은 항상 -1에서 1 이하
• 코사인 유사도의 값이 1에 가까울수록 두 텍스트가 유사하다고 판단
• 예) 두 텍스트 P와 Q를 나타내는 벡터가 P=(2,1)과 Q=(1,3)일 때,P와 Q의 코사인 유사도는 1/root(2). 두 벡터가 이루는 각 45도

https://wonhwa.tistory.com/26
https://velog.io/@jaeyun95/NLP%EC%9D%B4%EB%A1%A02.%EC%9E%90%EC%97%B0%EC%96%B4-%EC%B2%98%EB%A6%AC-%EA%B0%9C%EC%9A%94-%EC%9C%A0%EC%82%AC%EB%8F%84-%EB%B0%8F-%EB%AC%B8%EC%A0%9C%EB%93%A4
https://brunch.co.kr/@kakao-it/189

텍스트의 유사한 정도 분석.pdf 파일 참고 부탁드립니다.

자료 3. 산업공학, 통계학에서의 활용

산업공학은 효율성, 최적화, 공정 개선을 목표로 하는 학문 분야입니다. 최근에는 데이터 분석과 기계학습의 발전이 산업공학의 다양한 영역에서 새로운 기회를 창출하고 있습니다. 이 중에서도 텍스트 유사도 분석은 산업공학의 여러 분야에서 유용하게 활용될 수 있는 중요한 도구입니다.

1. **고객 피드백 분석**: 산업공학에서 고객 만족도는 중요한 지표 중 하나입니다. 텍스트 유사도 분석을 통해 고객 피드백, 리뷰, 설문 응답 등에서 유사한 의견이나 문제점을 식별할 수 있습니다. 이를 통해 고객의 요구사항을 더 잘 이해하고, 제

품이나 서비스 개선에 필요한 통찰력을 얻을 수 있습니다.

2. **제품 리뷰와 시장 분석**: 산업공학자들은 텍스트 유사도 분석을 활용하여 경쟁 제품의 리뷰를 분석할 수 있습니다. 이를 통해 시장의 동향, 소비자의 선호, 경쟁사의 강점 및 약점을 파악하는 데 도움이 됩니다.

3. **품질 관리**: 제조 과정에서 발생하는 기술적 문서나 보고서에서 반복되는 문제점을 찾는 데 텍스트 유사도 분석이 사용될 수 있습니다. 이를 통해 품질 문제를 조기에 식별하고, 공정 개선을 위한 조치를 취할 수 있습니다.

4. **지식 관리**: 산업공학에서는 대량의 기술 문서와 데이터를 관리하는 것이 중요합니다. 텍스트 유사도 분석을 통해 관련 문서를 빠르게 검색하고, 중복되거나 관련된 내용을 식별할 수 있습니다.

통계학은 데이터 분석과 해석에 중점을 두는 학문 분야로, 다양한 데이터 소스에서 유의미한 정보를 추출하고 이를 기반으로 결론을 도출하는 데 사용됩니다. 텍스트 데이터의 급증과 함께 텍스트 유사도 분석은 통계학적 접근과 결합되어 데이터 분석의 새로운 영역을 탐구하고 있습니다.

1. **데이터 마이닝과 텍스트 분석**: 통계학은 대규모 텍스트 데이터 세트에서 패턴을 식별하고 분류하는 데 사용됩니다. 텍스트 유사도 분석은 이러한 텍스트 데이터에서 유사한 항목을 그룹화하거나 주요 트렌드를 식별하는 데 중요한 역할을 합니다.

2. **감성 분석**: 소셜 미디어, 제품 리뷰, 고객 피드백 등에서 수집한 텍스트 데이터를 분석하여 사람들의 감정과 의견을 파악합니다. 텍스트 유사도 분석은 이러한 감성 분석에서 중요한 역할을 하며, 통계적 방법과 결합하여 더 정확한 감성 분류를 가능하게 합니다.

3. **시장 조사와 고객 행동 분석**: 텍스트 유사도 분석을 활용하여 시장 조사 보고서, 고객 행동에 대한 연구, 경쟁사 분석 등에서 수집된 데이터를 분석할 수 있습니다. 이를 통해 시장의 동향, 고객 선호도 및 행동 패턴을 통계적으로 분석하고 예측합니다.

4. **품질 관리와 피드백 분석**: 제품이나 서비스에 대한 고객 피드백과 리뷰를 분석하여 품질 관리 및 개선에 필요한 통찰력을 얻습니다. 텍스트 유사도 분석은 이러한 피드백에서 반복되는 주제나 문제점을 식별하는 데 유용합니다.