《 기하 수행평가 》

하이에듀

주제	코사인 유사도와 벡터

요약

데이터를 벡터로 표현하고, 그 벡터 사이의 각도를 이용(코사인 값)하여 데이터 간의 유사도를 판별하는 방법으로 '코사인 유사도'가 있습니다. 코사인 유사도는 자연어 문장들 사이의 유사도를 판단하고, DNA 염기들 사이의 유사도를 관찰해서 유전자를 분석하는 등 다양한 작용을 하고 있습니다.

가이드 자료를 준비하였습니다. 더 많은 자료가 필요하실 경우 언제든지 다시 연락주시길 부탁드리겠습니다. 감사합니다. :)

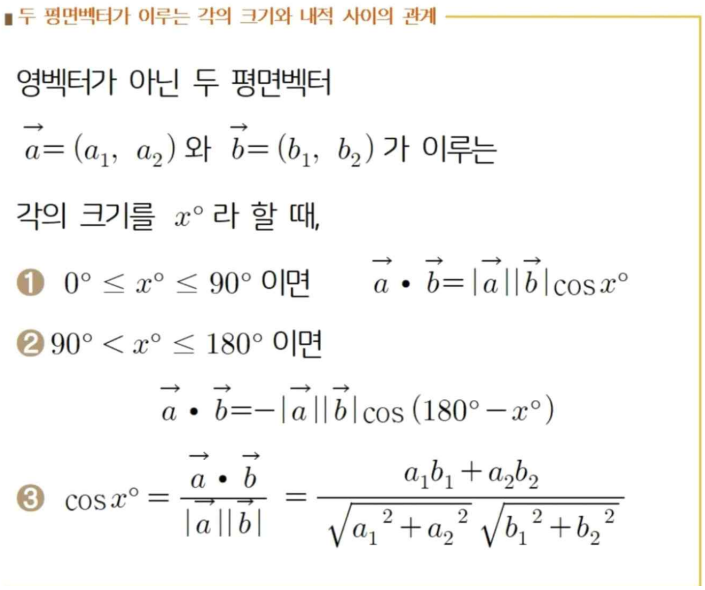
가이드

1. 탐구한 주제

'컴퓨터공학에서의 코사인 유사도와 벡터 내적의 이용' 정도로 적으면 됩니다. 탐구 제목에는 '벡터 내적'과 '코사인 유사도'라는 단어가 들어가도록 설정 부탁드립니다.

2. 탐구 주제에 적용한 기하 개념 또는 공식

평면벡터가 이루는 각의 크기와 내적 사이의 관계를 중심으로, 두 벡터가 이루는 각의 cos 값과 내적값 등을 기술해주시면 됩니다. 구체적으로는 아래 사진과 같이 적으면 좋을 것이라 판단됩니다.



두 평면벡터가 이루는 각의 크기와 내적 사이의 관계

영벡터가 아닌 두 평면벡터

$\vec{a}=(a_1,\ a_2)$ 와 $\vec{b}=(b_1,\ b_2)$ 가 이루는

각의 크기를 x° 라 할 때,

❶ $0^\circ \le x^\circ \le 90^\circ$ 이면 $\quad \vec{a}\cdot\vec{b}=|\vec{a}||\vec{b}|\cos x^\circ$

❷ $90^\circ < x^\circ \le 180^\circ$ 이면

$$\vec{a}\cdot\vec{b}=-|\vec{a}||\vec{b}|\cos(180^\circ-x^\circ)$$

❸ $$\cos x^\circ=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}=\frac{a_1b_1+a_2b_2}{\sqrt{{a_1}^2+{a_2}^2}\sqrt{{b_1}^2+{b_2}^2}}$$

3. 탐구 동기 (100자 이상)

컴퓨터공학에서 벡터가 사용되는 사례에 대해 찾아보았고, 그 결과 유사도를 판단하는 방법으로 벡터가 쓰인다는 식으로 기술해주시면 됩니다.

예시) 벡터가 컴퓨터공학에서 어떻게 쓰이는지 알고 싶어 관련 자료를 찾아보았다. 그 결과 데이터를 벡터로 표현하는 유사도 판단 방법이 있다는 것을 알게 되었다. 두 벡터 사이의 각도를 코사인을 이용해 수치화하는 코사인 유사도에 대해 더 알아보고 싶어서 탐구 주제를 선택하였다.

4. 탐구 내용 (200자 이상)

코사인 유사도란 벡터와 벡터 사이의 유사도를 비교할 때 두 벡터 간의 사잇각을 구해 얼마나 유사한지 수치로 나타낸 것입니다.

벡터 방향이 비슷할수록 두 벡터는 서로 유사하며, 두 벡터가 90도일 때는 관련성이 없고, 방향이 반대가 될수록 두 벡터는 반대 관계가 됩니다.

두 벡터의 크기가 모두 1이라고 가정하고, 같은 방향에 있으면 cos0=1이므로 코사인 유사도는 1입니다. 90도를 이루고 있으면 cos90=0이므로 코사인 유사도는 0입니다. 서로 반대 방향을 향하고 있으면 cos180=-1이므로 코사인 유사도는 -1입니다.

위에서 분석한 것과 같이, 백터의 내적은 각 벡터의 절댓값에 코사인을 곱한 것과 같습니다.

$$A \cdot B = |A| * |B| * \cos\Theta$$

그렇다면 식을 다음과 같이 만들 수 있습니다. 즉 평면좌표에서뿐만이 아니라, 다양한 차원에서의 벡터를 이렇게 다룰 수 있습니다.

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

문서 간 유사도를 측정하는 방법 중 유클리드 거리 기반의 좌표도 있지만, 단어의 빈도수에만 기반해서 유사도를 구하는 방법은 정확도가 떨어집니다. 따라서 문서 사이의 유사도를 측정할 때에는 코사인 유사도가 제일 많이 쓰입니다.

코사인 유사도는 다음과 같은 경우에서 주로 이용됩니다.
- 문장 혹은 문서간의 유사도를 구할 때
- 벡터의 크기가 중요하지 않고, 벡터 사이의 각도가 중요할 때

5. 느낀 점 (100자 이상)
다음과 같은 내용이 들어가면 보다 풍부한 느낀 점을 작성할 수 있습니다.
- 평면에서뿐만이 아니라 다양한 차원에서의 벡터를 다룰 수 있다는 것이 신기했다.
- 컴퓨터공학에서도 기하와 벡터가 폭넓게 이용되고 있다는 것이 신기했다.
- AI 기술을 다루기 위해서 기하와 벡터를 열심히 공부해야겠다는 생각이 들었다.

예시) 컴퓨터공학과 AI 기술에서도 벡터가 이용되고 있는지 몰랐는데, 이번 기회로 벡터의 폭넓은 쓰임을 알게 되었다. 또한, 평면좌표에서의 벡터뿐만이 아니라 컴퓨터에서 더 다양한 차원에서의 벡터를 다룰 수 있다는 것이 신기했다.

학생의 이해를 돕기 위해 정리가 잘 되어 있는 블로그 자료를 첨부하였습니다:
https://bkshin.tistory.com/entry/NLP-8-%EB%AC%B8%EC%84%9C-%EC%9C%A0%EC%82%AC%EB%8F%84-%EC%B8%A1%EC%A0%95-%EC%BD%94%EC%82%AC%EC%9D%B8-%EC%9C%A0%EC%82%AC%EB%8F%84