AI for Future Workforce

Module 25: NLP 분류

intel digital readiness

법률 고지사항

- Intel® 디지털 준비 프로그램 및 Intel® AI for Future Workforce 프로그램은 Intel Corporation에서 개발했습니다.
- 프로그램 날짜와 수업 계획은 변경될 수 있습니다.
- Intel 기술에는 활성화된 하드웨어, 소프트웨어 또는 서비스 활성화가 필요할 수 있습니다.
- 모든 제품과 구성 요소는 안전을 보장 할 수 없습니다.
- 결과물은 추정되거나 시뮬레이션 되었습니다.
- Intel은 타사 데이터를 제어하거나 감사하지 않습니다. 정확성을 평가하려면 다른 출처를 참조해야 합니다.
- 당신이 투자한 비용과 그에 대한 결과물은 다를 수 있습니다.

지난 모듈에서 배운 것 1가지는
무엇입니까?

or

모듈에서 만들어 본 것 1가지는
무엇입니까?

학습 목표

이 모듈을 통해 다음과 같은 역량을 습득할 수 있습니다:

- 데이터를 단어 형식의 가방으로 표현
- 학생들은 TFIDF 형식을 적용하는 법을 배움
- 학생들은 모델 교육을 위해 별도의 데이터 세트를 구성하는 법을 배움
- 학생들은 분류 모델을 구성하고 훈련시킴
- 데이터 파이프라인 생성

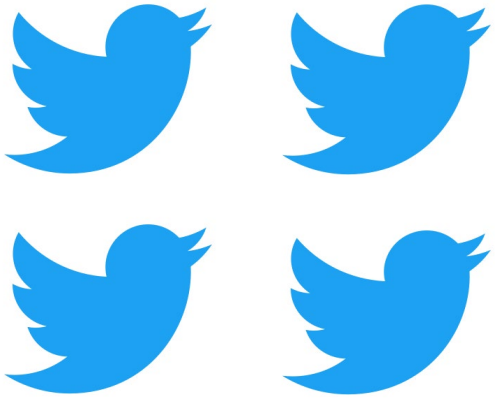
단어 가방
(The Bag of Words)

AI for Future Workforce

intel digital readiness

문서는 단일 텍스트 데이터(예: 트윗)이며, 말뭉치는 문서 모음입니다!

문서



말뭉치

단어 가방

- 단어 가방은 말뭉치에 있는 문서를 숫자 형태로 표현하는 방법입니다.
- 각 행은 하나의 문서를 나타내는 배열입니다. 그 열들은 어휘에 있는 단어들을 나타내는데, 이것은 말뭉치에 나타나는 모든 단어들에 대해 하나의 열이 있다는 것을 의미합니다.
- 각 행의 숫자는 특정 단어가 문서에 나타나는 횟수를 나타냅니다.
- 예를 살펴보겠습니다.

여러분의 말뭉치

- I love cats
- I do not like dogs
- I like fish
- Cats eat fish

여러분의 단어 가방

어휘 → (Vocabulary)

I	love	cats	do	not	like	dogs	fish	eat
1	1	1	0	0	0	0	0	0
1	0	0	1	1	1	1	0	0
1	0	0	0	0	1	0	1	0
0	0	1	0	0	0	0	1	1

자기 주도 학습

주피터 노트북 사용 방법은?

- 탐색하려면 키보드의 **위쪽** 및 **아래쪽** 화살표 키를 사용할 수 있습니다.
- 이 통합 문서에서 코드를 실행하려면 코드 블록을 선택하고 **Shift + Enter** 키를 누릅니다.
- 코드 블록을 편집하려면 **Enter** 키를 누릅니다.

시작하기 전에 주피터 노트북을 복사해 보는 것이 좋습니다.
문제가 있는 경우 항상 원본을 참조할 수 있습니다.

단어 가방 알고리즘

노트북 Section 1을 완료하십시오.

주요 학습 포인트

단어 가방

- 텍스트 데이터를 숫자로 변환합니다.
- 단어 가방은 각 트윗에 대해 각 단어가 나타나는 횟수를 카운트하여 입력 데이터로 사용합니다.

중요하고 관련성 있는 단어 찾기

노트북 Section 2를 완료하십시오.

주요 학습 포인트

전체 빈도, 역 문서 빈도

- 가장 일반적인 단어는 트윗에 대한 정보를 제공하는 데 유용하지 않을 수 있습니다.
- 중요한 특성은 일부 문서에는 자주 나타나지만 모든 문서에는 나타나지 않는 단어에서 찾을 수 있습니다.

모델 훈련

노트북 Section 3을 완료하십시오.

주요 학습 포인트

기계 학습으로 모델을 훈련시킵니다

- 다양한 기계 학습 모델을 제공하는 sklearn 라이브러리 사용
- 데이터를 훈련 세트 및 테스트 세트로 분할

파이프라인 구축

노트북 Section 4를 완료하십시오.

주요 학습 포인트

단어 가방 만들기

- 정규화
- 단어 벡터

중요 특성 찾기

- 가장 일반적인 단어

ML 모델 훈련

- sci-kit learn

기술의 전체적인 흐름을 간략하게 소개합니다.

본 모듈을 계속 학습하십시오!

분류 도전 과제

노트북 Section 5를 완료하십시오.

주요 학습 포인트

기계 학습으로 모델을 훈련시킵니다

- 감정 분석은 특정 주제에 대한 작가의 태도를 확인하기 위해 텍스트 조각에서 의견을 분류하는 과정입니다.
- CountVectorizer 함수를 사용하여 단어 가방을 만들 수 있습니다.
- TfidfTransformer 함수는 단어 가방을 TFIDF로 변환하는 데 사용할 수 있습니다.
- 영화 리뷰에 대한 감정 분석을 성공적으로 수행하셨습니까?

프로젝트 (Part 1)

이제 4인 1조로 프로젝트를 진행할 시간입니다!

프로젝트

이 프로젝트는 각 팀이 선택한 주제의 NLP 기반 분류 프로젝트를 만드는 것입니다.

다양한 레벨은 무엇입니까?

- **Level 1:** 데이터를 전처리합니다. 데이터가 처리 될 때 감소한 토큰 수를 카운트 하십시오.
- **Level 2:** 감정 분석 모델을 훈련시키고 성능을 분석하십시오.
- **Level 3:** NLP의 분류가 사회적 영향 프로젝트에 어떻게 사용될 수 있는지 적어도 5 가지 방법을 제안하고 수집해야하는 데이터 세트를 제안하십시오.

코딩을 시작하기 전에 계획하고 전략을 세우십시오.

- 어떤 주제에 관심이 있습니까?
- 어떤 종류의 분류 작업에 데이터를 사용 하시겠습니까?
- 그 분류가 어떻게 유용할 것이라고 생각하십니까?
- 주어진 시간 내에 작업을 완료할 수 있도록 팀의 작업을 어떻게 나누시겠습니까?

Half-Time!

각 팀은 진행 상황을 공유합니다.

- 어느 레벨을 달성했다고 생각하십니까?
- 지금까지 가장 큰 성공을 설명하십시오.
- 프로젝트를 시작하기 전에 극복하고 싶은 가장 큰 장애물에 대해 설명하십시오.
- 비슷한 도전 과제를 진행하는 팀이 있나요?
- 프로젝트와 관련하여 조언을 해 주실 분이 있습니까?

프로젝트 (Part 2)

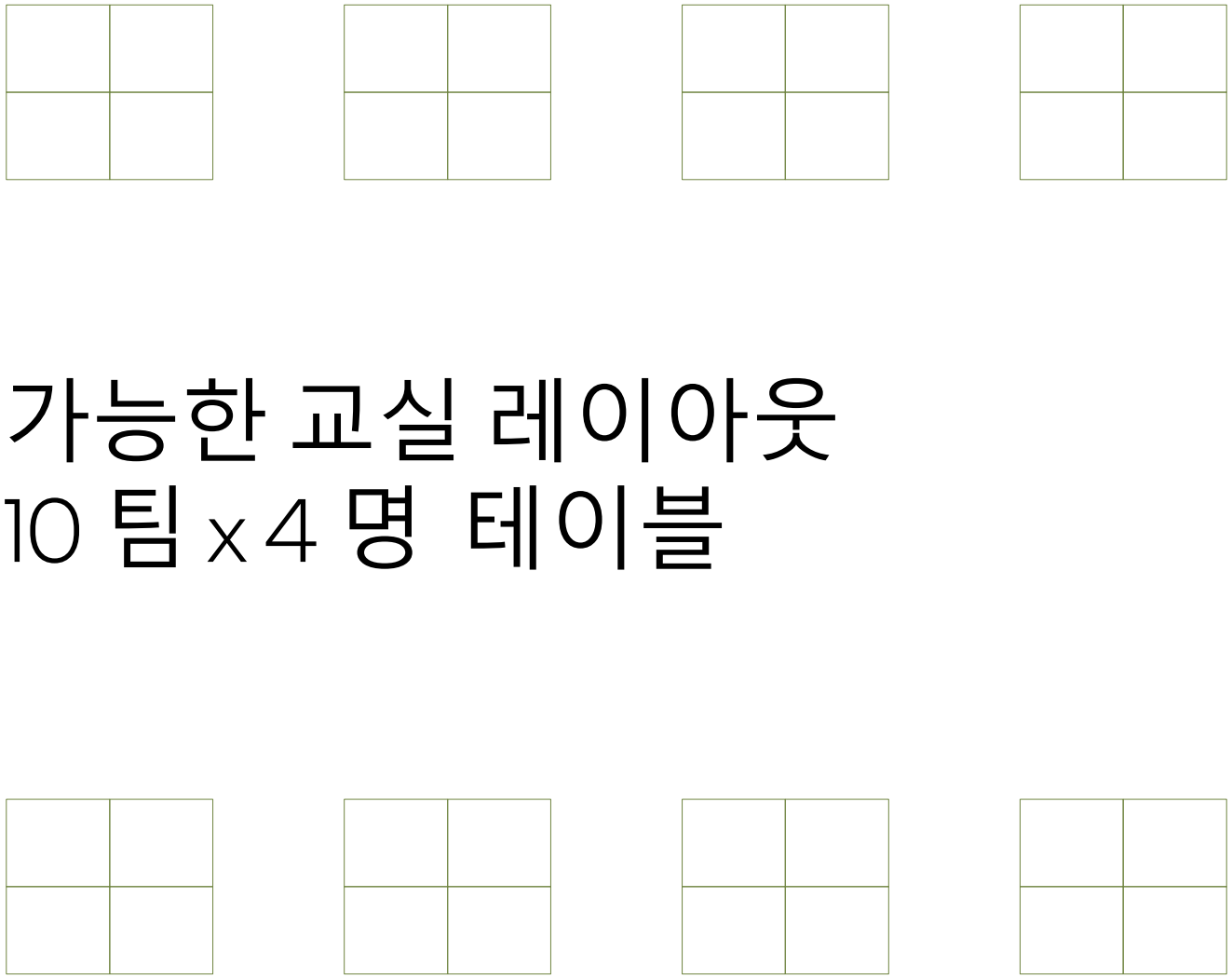
프로젝트

이 프로젝트는 각 팀이 선택한 주제의 NLP 기반 분류 프로젝트를 만드는 것입니다.

다양한 레벨은 무엇입니까?

- **Level 1:** 데이터를 전처리합니다. 데이터가 처리 될 때 감소한 토큰 수를 카운트 하십시오.
- **Level 2:** 감정 분석 모델을 훈련시키고 성능을 분석하십시오.
- **Level 3:** NLP의 분류가 사회적 영향 프로젝트에 어떻게 사용될 수 있는지 적어도 5 가지 방법을 제안하고 수집해야하는 데이터 세트를 제안하십시오.

프로젝트 발표



가능한 교실 레이아웃
10 팀 x 4 명 테이블

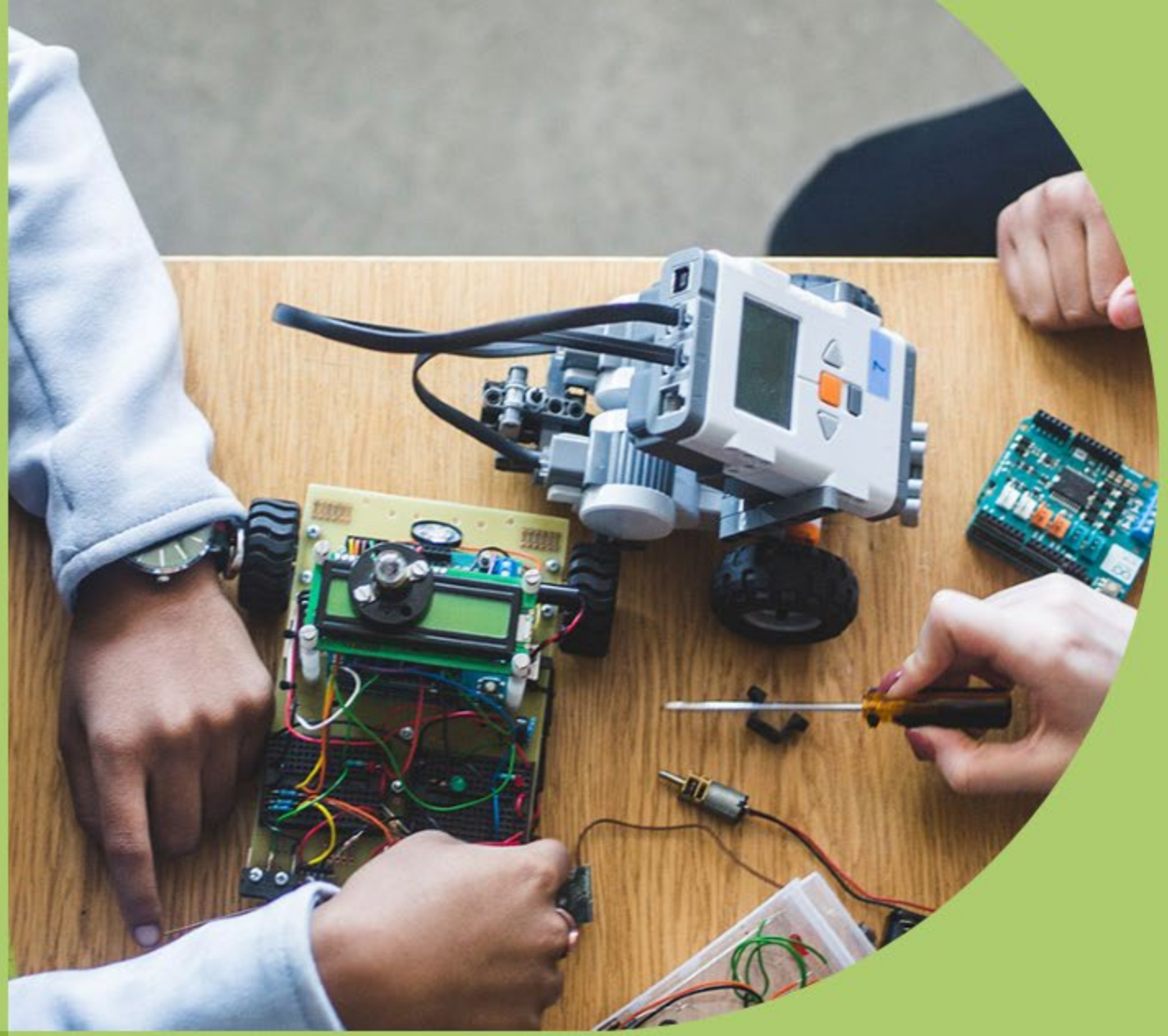
각 팀의 발표 순서

1. 어떤 데이터 세트를 다운로드했습니까?
2. 이 데이터 세트를 다운로드한 이유는 무엇입니까?
3. 데이터 세트를 어떤 애플리케이션에 사용할 수 있다고 생각하십니까?
4. 어떻게 업무를 분담하였습니까?
5. 데이터를 다운로드, 처리 및 분석하기 위해 어떤 기술을 사용했습니까?
6. 어떻게 모델을 훈련 시켰습니까?
7. 모델에 대해 무엇을 알았습니까?
8. 분류 모델이 어떻게 유용 할 수 있습니까?

모든 팀이 열심히 참여해 주셔서 감사합니다!

프로젝트에 대해 논의해 봅시다!

- 어떻게 접근하셨습니까?
- 어떤 도전에 직면하셨나요?
- 어떻게 극복하셨나요?
- 프로세스를 어떻게 개선할 수 있습니까?



요약

intel digital readiness

오늘 배운 것 중 개인적으로
유용하다고 생각되는 것은?

오늘 사용한 새로운 기술 중 하나를 공유해 보세요!

오늘 배운 내용으로 하고 싶은 일 한가지는 무엇입니까? 아니면 배운 것을 어떻게 적용하시겠습니까?

학습 목표

이 모듈을 통해 다음과 같은 역량을 습득할 수 있습니다:

- 데이터를 단어 형식의 가방으로 표현
- 학생들은 TFIDF 형식을 적용하는 법을 배움
- 학생들은 모델 교육을 위해 별도의 데이터 세트를 구성하는 법을 배움
- 학생들은 분류 모델을 구성하고 훈련시킴
- 데이터 파이프라인 생성

퀴즈

링크

적용

- 오늘 배운 것을 어떻게 적용하고 싶습니까?
- 일상 생활에 분류 모델을 적용 할 수 있습니까? 어떤 사업이 이로 인해 가장 큰 이익을 볼 수 있을까요?


intel®
digital
readiness