AI for Future Workforce

Module 24: NLP를 위한 데이터 수집 및 처리

intel digital readiness

법률 고지사항

- Intel® 디지털 준비 프로그램 및 Intel® AI for Future Workfork 프로그램은 Intel Corporation에서 개발했습니다.
- © Intel Corporation. Intel, Intel 로고 및 기타 Intel 마크는 Intel Corporation 또는 자회사의 상표입니다. 다른 이름 및 브랜드는 다른 사람의 재산으로 주장될 수 있습니다. 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 한 가지는 무엇인가요?

또는 모듈 이후 구축한 한 가지는
무엇인가요?

데이터 수집 및 처리

AI for Future Workforce

intel digital readiness

학습 효과

이 워크샵이 끝나면 다음을 수행할 수 있습니다.

- 자연어 처리 및 기술의 현재 응용 프로그램 설명
- 자연어 처리의 이론과 응용에 대한 이해
- 인터넷에서 텍스트 데이터 다운로드
- 문장 분할, 토큰화, 중지 단어 제거 등으로 텍스트 데이터를 처리
- 다운로드 및 처리된 텍스트 데이터 탐색

텍스트 데이터의 출처는 무엇입니까?

데이터 출처(소스)

- 기존 공개 데이터 세트(오늘 할 것입니다!)
- 웹사이트에서 데이터 수집(오늘 할 것입니다!)
- 유료 데이터 세트 예. 뉴스 아카이브, 상업 데이터베이스
- 데이터의 수동 수집 예: 인터뷰, 설문 조사, 퀴즈

데이터 세트용 리소스

Link here

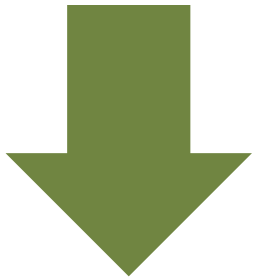
kaggle

Project Gutenburg

데이터 처리

배운 것을 기억 하나요?

Text



Numbers

어떻게 텍스트를 숫자로 변환하는가?

자기 주도 학습

Jupyter Notebook 사용 방법

- 위아래로 탐색하려면 키보드의 위아래 화살표 키를 사용할 수 있습니다.
- 이 통합 문서의 코드를 실행하려면 코드 블록을 선택하고 Shift + Enter를 누릅니다.
- 코드 블록을 편집하려면 Enter 키를 누릅니다.

시작하기 전에 원본 노트북을 복사해 두면 문제가 발생할 경우 항상 원본을 다시 참조할 수 있습니다

Link here

웹사이트 구조에 접근하기

- 웹사이트로 이동Ctrl-Shift-I를 눌러 Chrome devtools 콘솔을 엽니다(또는 다른 브라우저의 경우 F12).
- 웹사이트의 html 구조를 보려면 'Elements'를 클릭하십시오.
- Elements 선택기를 활성화하려면 Ctrl-Shift-C를 누르십시오.
- 웹 페이지에서 요소를 클릭하면 오른쪽에 있는 html Elements로 이동합니다.
- 태그 및 클래스 이름 참고

Link here

h2.entry-title.entry-title-portfolio | 450.41 × 28

#Startathon

Elements Memory Console Sources Audits Network Performance

```
<div id="x-iso-container" class="x-iso-container x-iso-container-portfolio cols-4 isotope" style="overflow: hidden; position: relative; height: 3657px;">
  <article id="post-3756" class="post-3756 x-portfolio type-x-portfolio status-publish has-post-thumbnail hentry portfolio-category-open-innovation x-portfolio-5f2a5be3d99ef20f40d8590b6fdae0a8 isotope-item" style="position: absolute; left: 0px; top: 0px; transform: translate3d(0px, 0px, 0px); opacity: 1;">
    <div class="entry-featured">
      <a href="http://www.sl2square.org/portfolio-item/startathon/" class="entry-thumb" title="Permalink to: "#Startathon""> == $0
        ::before
        <img width="1092" height="610" src="http://www.sl2square.org/wp-content/uploads/2015/04/startathon1-1170x654.jpg" class="attachment-entry-cropped-fullwidth size-entry-cropped-fullwidth wp-post-image" alt>
      </a>
    </div>
    <div class="entry-wrap cf">
      ::before
      <header class="entry-header">
        <h2 class="entry-title entry-title-portfolio">…</h2>
      </header>
      ::after
    </div>
  </article>
  <article id="post-3711" class="post-3711 x-portfolio type-x-portfolio status-publish has-post-thumbnail hentry portfolio-category-open-innovation portfolio-category-product-development-and-
```

a.entry-thumb img

Styles Computed

Filter :hov .cls +

```
element.style {
}

. integrity-light…ss?ver=6.3.8:1
entry-thumb {
  display: block;
  position: relative;
  background-color: #000;
}

a, h1 a:hover, h2 a:hover, h3 a:hover, h4 a:hover, h5 a:hover, h6 a:hover, .x-breadcrumb-wrap a:hover, .widget ul li a:hover, .widget ol li a:hover, .widget.widget_text ul li a, .widget.widget_text ol li a, .widget_nav_menu .current-menu-item > a, .x-accordion-heading .x-accordion-toggle:hover, .x-comment-author a:hover, .x-comment-time:hover, .x-recent-posts a:hover .h-recent-posts {  (index):45
  color: #02aed6;
}

a integrity-light…ss?ver=6.3.8:1
{
  text-decoration: none;
}

* integrity-light…ss?ver=6.3.8:1
, *:before, *:after {
  box-sizing: border-box;
}

a:-webkit- user agent stylesheet
```

Console Rendering Remote devices Sensors What's New ×

노트북에서 Section 1.1 완료

주요 내용

인터넷에서 데이터 수집

- 요청 패키지를 사용하면 파이썬 스크립트가 웹사이트와 통신하고 해당 사이트에서 정보를 '요청'할 수 있습니다.
- bs4라고도 하는 아름다운 수프 패키지는 웹사이트에서 원시 정보를 가져와 정보를 추출하는 데 유용한 기능을 제공합니다.

스크랩한 데이터 저장하기
섹션 1.2 완료

주요 내용

데이터 저장

- 데이터가 필요할 때마다 항상 웹을 스크랩할 필요가 없도록 데이터를 저장하세요.
- 웹사이트가 다운되는 경우에 대비하여 데이터를 보존하세요.

기타 데이터 소스
섹션 1.3 완료

Link here

주요 내용

기타 데이터 소스

- 공개적으로 컴파일되고 공유된 많은 데이터 세트가 있습니다.
  이를 사용하여 처리 시간과 노력을 줄일 수 있습니다.
  예) 트위터 데이터세트

NLTK(자연어 도구 키트)를 사용하여 NLP 데이터 작업 섹션 2.1 완료

주요 내용

텍스트 처리

1. 컴퓨터가 쉽게 처리할 수 있는 형태로 텍스트 나누기
2. 텍스트를 분석하고 시각화할 수 있습니다.
3. 사용 도구
   - Pandas
   - NLTK
4. 텍스트를 가장 작은 단위로 나누는 토큰화

NLP 데이터 처리
섹션 2.2 완료

NLP를 위한 데이터 전처리

Tokenization
(토큰화)

Stemming
(형태소 분석)

Normalization
(정규화)

Lemming
(레밍)

주요 내용

NLP를 위한 데이터 전처리

Gather Data

Web Scraping
(데이터수집
웹 스크래핑)

Storing Data
(데이터 저장)

Processing Text
(텍스트 처리)

- Requests Package
- Beautiful Soup Package

- List

- Pandas
- NLTK
- Matplotlib

프로젝트 (1/2)

4인 1팀을 구성하세요.

프로젝트

이 프로젝트는 각 팀이 얻은 텍스트 데이터를 수집, 처리 및 표시하는 것입니다.

각각 다른 레벨의 프로젝트가 있습니다.

- **Level 1:** 최소 50,000개의 토큰으로 데이터 수집(책/웹사이트에서)
- **Level 2:** 데이터를 사전 처리합니다. 데이터가 처리됨에 따라 토큰 수가 얼마나 감소하는지 계산합니다.
- **Level 3:** 토큰 분석: 상위 5개 토큰은 무엇입니까? 코퍼스를 설명하는 데 어떻게 도움이 됩니까?
- **Level 4:** 데이터세트를 사회적 영향 프로젝트에 사용할 수 있는 방법을 5가지 이상 제안합니다.

코딩을 시작하기 전 계획하고 전략을 세우세요.

- 어떤 주제에 관심이 있나요?
- 수집한 데이터가 어떻게 더 사용될 것이라고 생각하나요?
- 주어진 시간 내 작업이 완료될 수 있도록 팀의 작업을 어떻게 나눌 건가요?

하프타임!

각 팀은 진행 상황을 공유합니다

- 어느 수준에 도달했다고 생각하나요?
- 가장 높게 도달한 것에 대해 설명하세요.
- 프로젝트를 선보이기 전에 극복하고자 했던 가장 큰 애로사항을 설명하세요.
- 비슷한 애로사항이 있는 사람이 있나요?
- 도움을 주거나 조언을 하고 싶은 사람이 있나요?

프로젝트 (2/2)

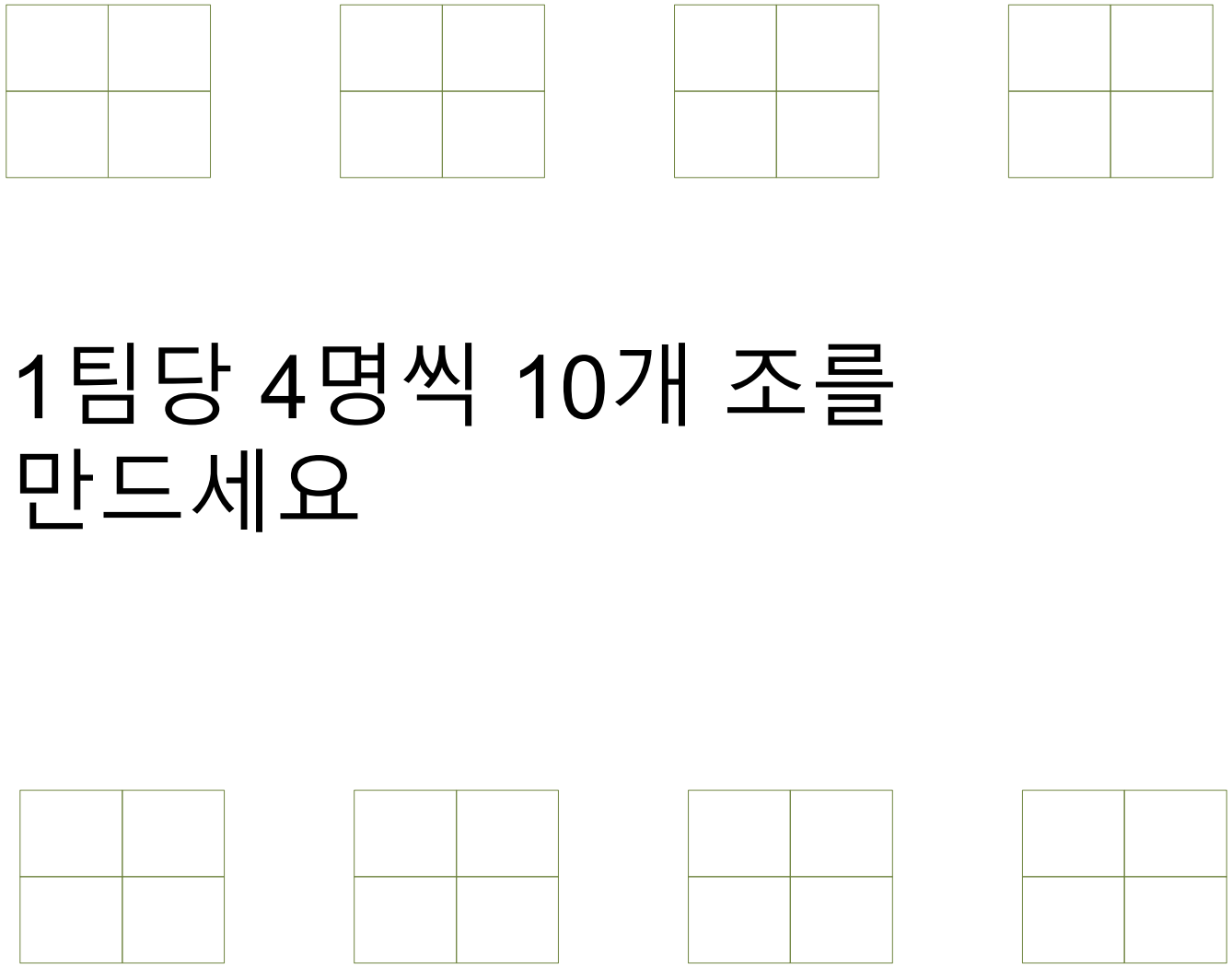
프로젝트

이 프로젝트는 각 팀이 텍스트 데이터를 수집, 처리 및 표시하는 것입니다.

각각 다른 레벨의 프로젝트가 있습니다.

- **Level 1:** 최소 50,000개의 토큰으로 데이터 수집(책/웹사이트에서)
- **Level 2:** 데이터를 사전 처리합니다. 데이터가 처리됨에 따라 토큰 수가 얼마나 감소하는지 계산합니다.
- **Level 3:** 토큰 분석: 상위 5개 토큰은 무엇입니까? 코퍼스를 설명하는 데 어떻게 도움이 됩니까?
- **Level 4:** 데이터세트를 사회적 영향 프로젝트에 사용할 수 있는 방법을 5가지 이상 제안합니다.

프로젝트 발표



1팀당 4명씩 10개 조를 만드세요

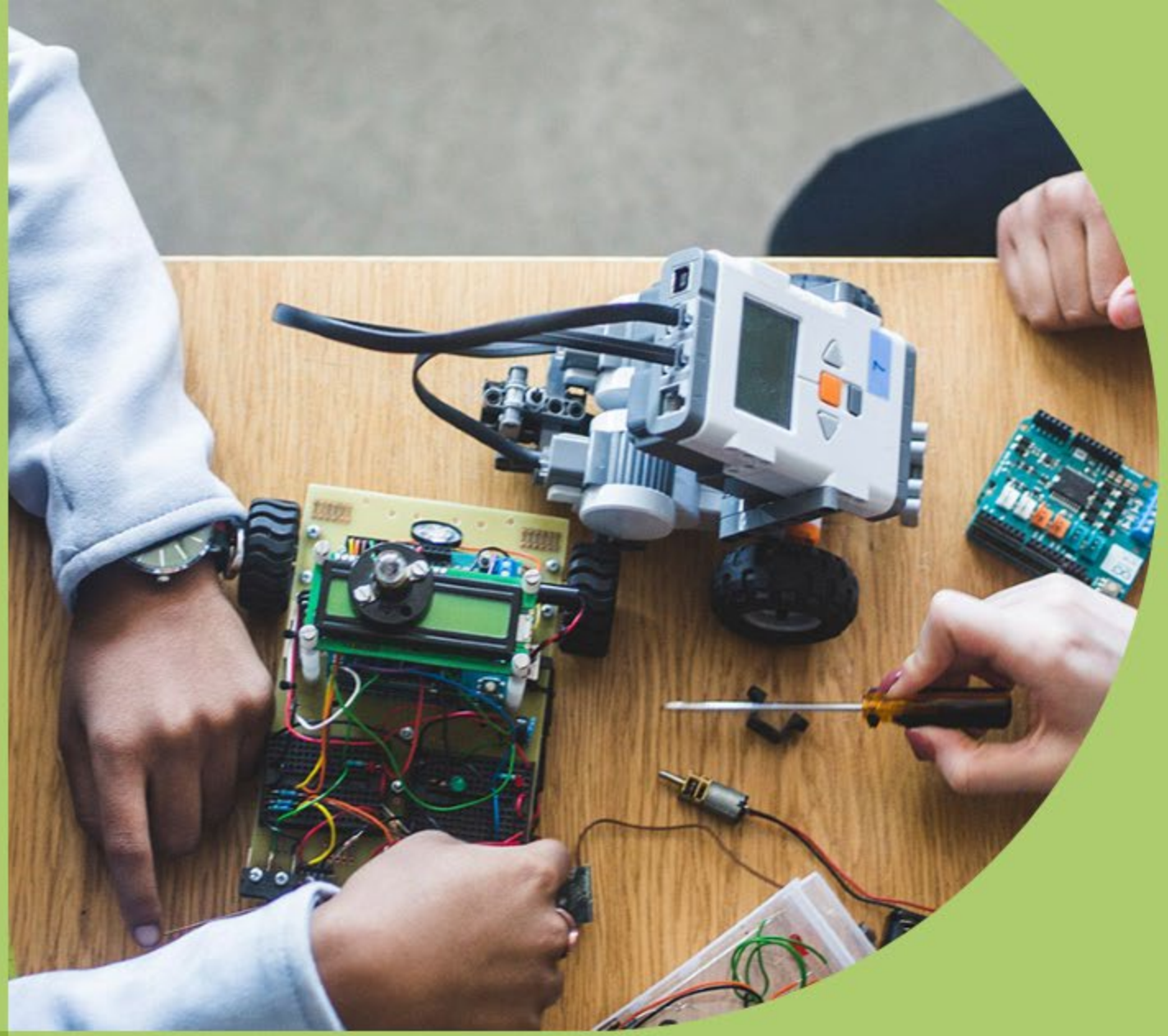
각 팀 발표 순서

1. 어떤 데이터 세트를 다운로드 했습니까?
2. 특정 데이터 세트를 다운로드하는 이유는 무엇입니까?
3. 데이터 세트를 어떤 애플리케이션에 사용할 수 있다고 생각합니까?
4. 업무 위임은 어떻게 하셨나요?
5. 데이터를 다운로드, 처리 및 분석하는 데 어떤 기술을 사용했습니까?
6. 데이터에 대해 무엇을 알게 되었습니까?
   - 가장 일반적인 단어
   - 희귀 단어
   - 토큰 수

모두 수고하셨습니다.

논의해 봅십다.

- 당신의 접근 방식은 무엇이었습니까?
- 어떤 어려움을 겪었습니까?
- 어떻게 극복했나요?
- 어떻게 개선할 것인가요?



요약

intel digital readiness

오늘 배운 것 중 개인적으로 유용하다고 생각하는 한 가지를 얘기해보세요.

오늘 사용한 새로운 기술 하나를 공유하세요!

오늘 배운 내용으로 함께 해보고 싶은
한가지를 공유해볼까요?
아니면 배운 것을 어떻게 적용할 것인가에
대해 얘기해볼까요?

학습 효과

이 워크샵이 끝나면 다음을 수행할 수 있습니다.

- 자연어 처리 및 기술의 현재 응용 프로그램 설명
- 자연어 처리의 이론과 응용에 대한 이해
- 인터넷에서 텍스트 데이터 다운로드
- 문장 분할, 토큰화, 중지 단어 제거 등으로 텍스트 데이터를 처리
- 다운로드 및 처리된 텍스트 데이터 탐색

퀴즈

Link here

적용

- 오늘 배운 것을 어떻게 이 수업의 맥락을 넘어 어떻게 적용하고 싶습니까?
- 오늘 배운 것을 어떻게 보는지, 현재의 세계에서 도움이 되나요?
- AI 애플리케이션을 구축할 때 주의해야 할 사항은 무엇입니까? 개인 정보 보호 및 안전 고려 사항이 있습니까?


intel®
digital
readiness