


중간 발표

악성 웹사이트 모니터링 시스템 제작



목차 소개

Table of Contents

01

프로젝트 소개

- 프로젝트
- 프로젝트 순서

02

지금까지 해온 것

- 진행 사항
- 흐름도

03

제작 현황

- URL 크롤러
- 데이터 정제 및 분류기

04

해야 할 것

- 사이트 중요도 분석
- 단어 중요도 분석

05

계획

- 최종 결과물
- 앞으로의 계획

06

마무리

- 인사말
- QnA



01. 프로젝트 소개

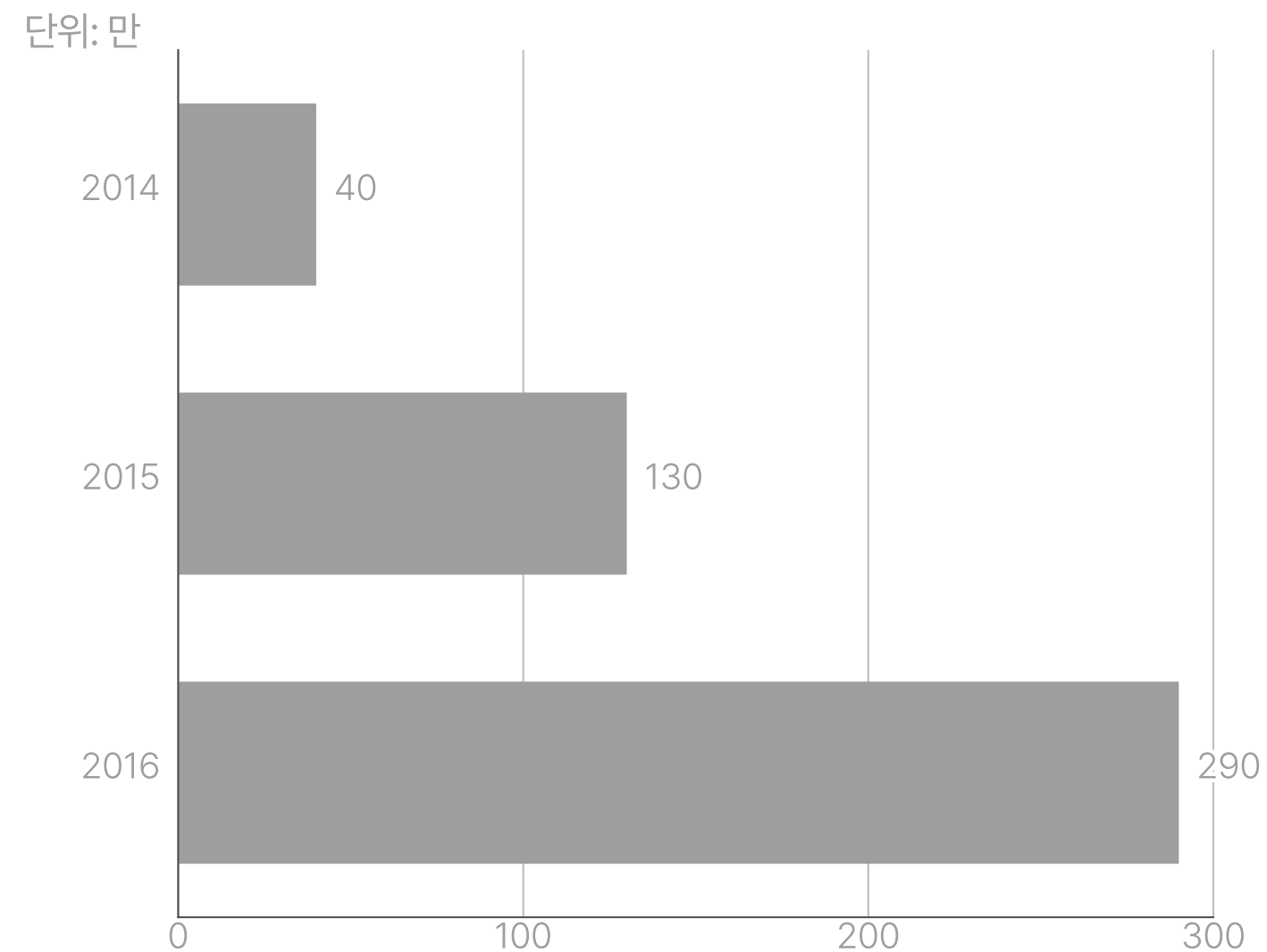
프로젝트 정의

“악성 웹사이트 탐지 및 모니터링 시스템 구축”



01. 프로젝트 소개

프로젝트 필요성



악성 사이트 증가 그래프

» 사이버 범죄 예방

- 악성 웹사이트는 저작권과 정보를 침해하고, 실제 재산 손실을 초래
- 악성 웹사이트에 대한 선제적 차단으로 위와 같은 사회 문제를 예방하고 해결할 수 있음

» 악성 웹사이트의 탐지와 빠른 대처

- 악성 웹사이트의 범위는 갈수록 늘어나는 추세
- 악성 웹사이트 탐지 범위를 확장하고 차단을 우회, 새롭게 등장한 사이트를 추적할 수 있음



01. 프로젝트 소개

프로젝트 목표

» 프로젝트 목표

- 악성 웹사이트를 자동으로 탐지하고, 모니터링한다.
- 악성 웹사이트를 종류별로 구분한다. (도박/불법웹툰/음란물/...)
- 악성 웹사이트를 한 번에 볼 수 있는 웹페이지를 제작하고,
악성 웹사이트를 분류해주는 프로그램을 제작한다.



01. 프로젝트 소개

프로젝트 순서

01 URL 크롤러 제작

02 키워드 기반 악성 웹사이트 분류 + LLM을 이용한 악성 웹사이트 분류

03 악성 사이트 게시 웹페이지 제작 + 악성 웹사이트 분류 프로그램 제작



02. 지금까지 해온 것

진행사항

» 웹사이트 링크 추출

- 링크를 추출해 큐에 삽입
- 연속적으로 링크를 타고 들어가 수집할 수 있음

» 키워드 기반 사이트 분류

- HTML 내용 추출
- 악성 웹사이트인지 아닌지 구분

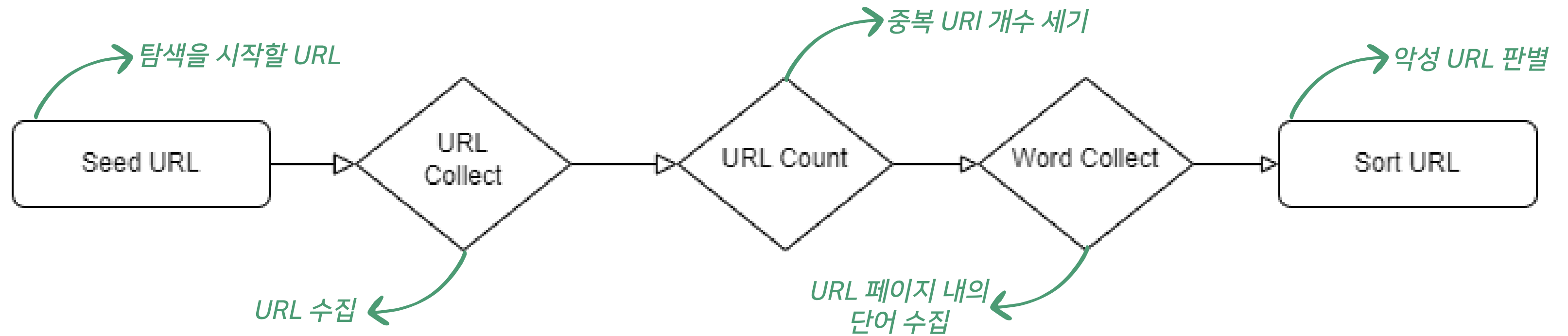
» 웹사이트 중요도 산출

- 사이트 참조 빈도를 파악해 중요도 산출
- PageRank 알고리즘 사용



02. 지금까지 해온 것

흐름도





03. 제작 현황

URL 크롤러

» a tag의 href 속성 활용

```
# 모든 링크를 찾아서 리스트에 추가
for link in soup.find_all('a'):
    href = link.get('href')
    if href and (href.startswith('http://') or href.startswith('https://')):
        # URL 인코딩
        href = quote(href, safe='/:?=&%')
        scraped_urls.append(href)
```

구현 코드 예시

» 큐를 이용한 URL 저장



발견한 URL 큐에 저장
→ 큐에 저장한 URL에 들어가 URL 수집
→ 반복

» URL 정제

- HOST URL
- 접속 가능한 URL
- http → https

» 화이트 리스트 생성



구글, 네이버, 페이스북, ...
자주 등장하는 악성이 아닌 웹사이트 제외



03. 제작 현황

데이터 정제 및 분류기

» 단어 기반 분류

키워드 기반 악성,정상으로 분류

```
malicious_keywords_webtoon = ['무료웹툰', '보증토토', '토토보증업체', '레진', '레진코믹스', '네이버웹툰', '다음웹툰', '카카오웹툰']
```

```
malicious_keywords_gambling = ['배팅', '배팅', '배팅하기', '카지노', '슬롯', '입금', '리그', '중계중', '스포츠', '스포츠중계', '토너먼트', '고스톱', '포커', '셋다', '맞고', '룰렛']
```

단어 분류를 위한 리스트

```
casino = ["도박", "카지노", "또또", "로또", "환불", "포인트"]
```

```
adult = ["19", "유혹", "오피", "정화", "한국", "야동", "BJ"]
```

선정한 키워드 예시



04. 현재 하고 있는 것

크롤러 & 데이터 정제 및 분류기 발전

사이트 중요도 분석

단어 중요도 분석

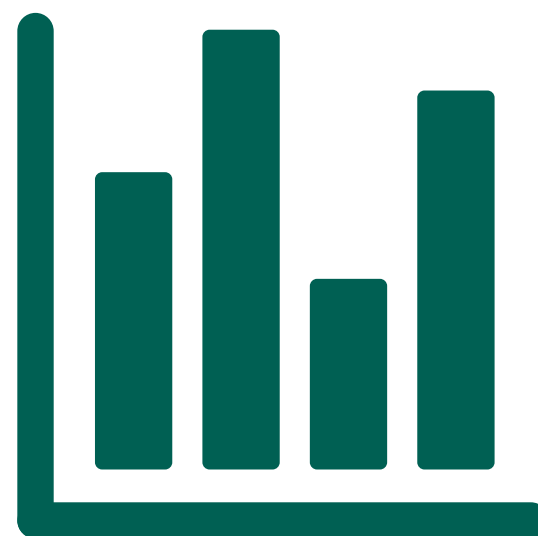


04. 해야 할 것

사이트 중요도 분석

» PageRank Algorithm

우선적인 차단



사이트 중요도 측정



더 나은 분류체계 수립



04. 해야 할 것

단어 중요도 분석

다크웹 은어 탐지

단어 중요도 측정



분류 작업 속도 향상

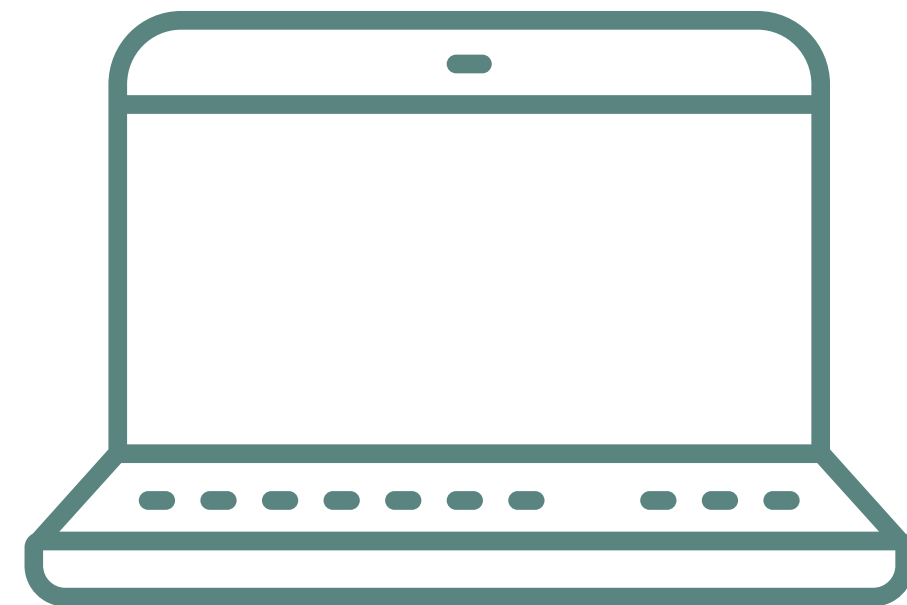


05. 계획

최종 결과물



악성 웹사이트를 자동으로
탐지하는 프로그램 제작



악성 웹사이트를 게시한
웹페이지 제작



05. 계획

앞으로의 계획

» 발전

- 크롤러의 정확도와 속도를 높이기
- 더 나은 방식의 크롤러를 구현할 수 있는지 생각해보기

» 분류

- 키워드 별 악성 웹사이트 분류 및 LLM을 이용한 악성 웹사이트 분류 진행
- 정확도가 높고 속도가 빠른 분류 방법 찾기
- 악성 웹사이트 종류 별로 분류하기

» 제작

- 악성 웹사이트를 구분하여 웹사이트에 게시
- 검색 기능 및 추적 기능 고안
- URL 입력 시 해당 사이트가 악성 웹사이트인지 구분하는 기능 고안



06. 마무리

인사말

감사합니다



06. 마무리

QnA

QnA