PDF 논문 정보 처리

DeepShark Lap 학부연구생 이지언

개발 목적

1. 논문 본문 외 불필요한 노이즈(그림/표 캡션 등) 자동 제거.
2. 대주제-소주제-본문의 트리(Tree) 구조로 나타냄.

전체적인 흐름

1.fitz(PyMuPDF) 라이브러리를 사용하여 PDF 파일을 시스템 메모리에 로드. 이 코드는 문서의 모든 페이지를 전수 조사하며 포함된 텍스트 데이터를 추출하는데, 단순히 글자만 읽어오는 것이 아니라 줄바꿈 정보(\n)를 보존하여 이후 단계에서 문장의 맥락을 파악할 수 있는 기초 데이터를 형성.

2. 추출된 원문 텍스트에서 논문의 뼈대를 찾기 위해 정규표현식 기반의 파싱을 수행한다. 미리 정의된 main_pattern과 sub_pattern을 활용하여 대주제와 소주제를 실시간으로 식별한다.

3. current_main과 current_sub라는 상태 변수를 이정표로 활용한다. 제목이 탐지될 때마다 이 변수를 업데이트하고, 이후 나타나는 본문 텍스트들을 현재 활성화된 방에 자동으로 귀속시킴으로써 문서의 논리적 계층 구조를 완성한다.

4. garbage_keywords 분석을 통해 그림이나 표의 내용 등 본문과 무관한 요소를 제거한다.

핵심 코드

```python
# 대주제 및 소주제 인식 패턴
main_pattern = r'^([0-9]+\.|[IIIIIIVVVIVIIVIIIIXX]\.)\s+(.*)'
sub_pattern = r'^([0-9]+\.[0-9]+)\s+(.*)'

for line in lines:
    line = line.strip()
    if not line: continue

    # 참고문헌이 시작되면 추출 종료
    if re.match(r'^(References|참고문헌|REFERENCE)', line, re.IGNORECASE):
        break

    main_match = re.match(main_pattern, line)
    sub_match = re.match(sub_pattern, line)
```

```
# 대주제 탐지 및 제목 표준화 (서론/결론/관련연구 등)
if main_match:
    header_text = main_match.group(2).strip()
    if '서론' in header_text or 'Intro' in header_text: # 제목에 '서론'이나 'Intro'가 들어있다면
        current_main = "1. 서론"                        # 무조건 1. 서론으로 이름을 통일한다.
    elif '결론' in header_text or 'Conclusion' in header_text:
        current_main = "4. 결론"
    elif '관련' in header_text:
        current_main = "2. 관련연구"
    else:
        current_main = f"{main_match.group(1)} {header_text}"

    # 새로운 대주제를 저장할 공간 생성
    self.data["sections"][current_main] = {"content": "", "sub_sections": {}}
    current_sub = None # 새로운 대주제 시작에 따른 소주제 상태 초기화

# 소주제 탐지 및 구조적 저장
elif sub_match and current_main:
    current_sub = sub_match.group(1).strip()
    sub_title = sub_match.group(2).strip()
    self.data["sections"][current_main]["sub_sections"][current_sub] = {
        "title": sub_title, "content": ""
    }

# 일반 본문 텍스트를 현재 활성화된 섹션에 저장
elif current_main:
    # 소주제가 활성화된 상태라면 해당 소주제에 본문 추가
    if current_sub:
        self.data["sections"][current_main]["sub_sections"][current_sub]["content"] += line + "\n"
    # 소주제가 없다면 대주제 직속 본문에 추가
    else:
        self.data["sections"][current_main]["content"] += line + "\n"
```

```
garbage_keywords = [
    'Fig', 'Table', '그림', '표', 'DOI', 'ISSN', 'http',
    'Google', 'Facebook', 'Microsoft', 'Naver', 'Service', 'Company', # 표 1 관련
    'Problem', 'Analysis', 'Productivity', 'efficiency', 'False alarm', # 표 2 관련
    'Solution', 'Suggested', 'accuracy', 'detection rate', '90%', # 표 3 관련
    'zero-day', 'exploits', 'malware', 'Neural', 'Translation'
]

for line in lines:
    line_strip = line.strip()

    # 키워드 기반 필터링 (대소문자 구분 x)
    if any(kw.lower() in line_strip.lower() for kw in garbage_keywords):
        continue

    # 한글이 하나도 없는 영문/숫자 위주의 줄은 표 데이터일 확률이 높으므로 제거
    if len(line_strip) > 5 and not re.search('[가-힣]', line_strip):
        continue

    # 너무 짧은 텍스트 제거
    if len(line_strip) < 3:
        continue
```

실 행 결 과

```
================================================================================
 [ 논문 제목 ] 딥러닝 기술이 가지는 보안 문제점에 대한 분석
================================================================================

1. 서론
최근 4차 산업혁명 시대가 도래되면서 인공지능과 빅 데이터 처리에 대한 활용이 높아지고 있다. 인공지능인 식은 최근 전 세계적 관심을 모았던 이세돌 선수와 인공 지능 알파고의 바둑 대결을 기억하면 쉽게 이해할 수 있 다. 인공지능의 활용은 인공지능의 핵심기술의 하나인 딥 러닝이라 하는 인공신경망 기반의 기술로 어떠한 데이터 가 있을 때 이를 컴퓨터가 알아들을 수 있는 명령 중심으 로 학습을 통해 적용하도록 되어있으며 학습 후, 목적에 부합된 특수한 영역에 활용될 수 있는 알고리즘 형태의 기계학습이다. 본 논문에서는 인공지능 기술의 하나인 딥 러닝 기술이 인터넷과 연결된 다양한 비즈니스 분야에 새로운 형태의 친화적 서비스를 업무적으로 잘 활용할 것이라는 기대를 가지고 있다. 특히, 딥러닝과 같은 인공 지능의 핵심 기술은 사용자가 손쉽게 원하는 지식을 대 화하면서 원하는 학습 방향으로 유도하며 정보를 획득하 고 의사소통할 수 있도록 되어있기 때문에 그 가능성을 비즈니스 업무의 특수한 영역인 보안 업무에 활용할 수 있도록 타당성을 검토하여 문제점을 도출해 내고자 한다. 이러한 역할의 딥러닝이 경쟁력과 충분한 가능성을 갖추 고 있다면 이를 잘 훈련시키어 IT 업계 보안 시스템에 적용시킬 경우에는 기존 보안시스템에 문제점과 비용적 인 측면을 고려해 볼 때에 충분한 경쟁력이 확보될 것으 로 예상하고 있다[1]. 본 논문의 구성은 다음과 같다. 1장 서론에서는 인공 법의 기본적인 내용에 대해서 알아보고 2장 관련 연구에 서는 딥러닝 작동방식 및 특징에 대해서 알아보고 국내 외 전반적인 기술적 동향에 대해서 살펴본다. 3장에서는 딥러닝의 활용할 수 있는 기술적 특징에 대해서 알아보 고, 4장에서는 딥러닝이 안고 있는 문제점을 검토한 후 학습 방향으로 유도하여 비즈니스 업무에 활용할 수 있 도록 비인가자의 IP 및 세션정보에 대한 문제점을 분석하여, 5장에서 향후 기술적 기대 방향과 함께 결론으로 마무리한다.

2. 관련연구
딥러닝은 어떠한 문제 처리를 사람이 직접 지시하지 않아도 데이터를 통해 컴퓨터가 패턴 인식 문제 또는 특 징적 학습을 하여 그것을 스스로 처리하고 해결할 수 있 도록 하는 기계학습 기술이다. 이는 실제 인간의 뇌가 뉴 런들 간의 연결이 매우 깊은(deep) 구조를 가지고 있다 는 점에서 보다 진보된 학습과 추론에 대한 인공지능 기 술이라 정의할 수 있다[2].

2-1. 딥러닝 작동 방식
딥러닝의 학습 작동 방식 중 가장 중요한 부분은 분석 기술의 정보 추출 분야로 구조적 분류의 관점에서 어떻게 접근해야 하는지 이다. 이를 위해서는 해당 도메인에 서의 적정량의 학습데이터와, 해당 데이터로부터 최적의 분류 함수를 얻기 위한 다음과 같은 두 가지 기계 학습 (machine learning)이 필요하다. ① 첫 번째 기계 학습에 대한 부분은 경험적 역할로 기계 학습을 통해 스스로가 특정한 분야에 반복적 으로 적응하면서 지적 영역과 경험적 영역을 넓혀 나가는 방식이다. ② 두 번째 기계 학습에 대한 부분은 규칙과 패턴을 익히는 역할로 기계 학습을 통해 문제의 규칙을 익 히고 학습을 통해 필요한 과정을 습득하여, 같은 유형을 습득한 후, 비슷한 다른 문제의 영역까지도 해결 수 있는 강력한 응용력을 적응하는 방식이다.
```