2024-1학기 DU-도전학기 계획서

<table>
<tr><td>과제명</td><td colspan="4">효율 및 정확도 향상을 위한 표절 검사 프로그램 개발</td></tr>
<tr><td>신청 유형</td><td colspan="4">☐ 개인 ■ 팀(팀명: CopyCat)</td></tr>
<tr><td rowspan="2">도전 영역</td><td colspan="4">☐ 학생설계 ■ 대학제안</td></tr>
<tr><td colspan="4">■ 전공(주전공 또는 복수전공) ☐ 일반선택</td></tr>
<tr><td>신청 학점</td><td colspan="4">3학점</td></tr>
<tr><td rowspan="5">참여자</td><td>성명</td><td>소속</td><td>학번</td><td>비고</td></tr>
<tr><td>박OO</td><td>컴퓨터공학전공</td><td></td><td>팀장</td></tr>
<tr><td>안OO</td><td>컴퓨터소프트웨어전공</td><td></td><td>팀원</td></tr>
<tr><td>강OO</td><td>컴퓨터공학전공</td><td></td><td>팀원</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>지도교수 의견</td><td colspan="4">상기 학생들이 제안한 해시 암호 및 딥러닝 기술 기반 표절 검사 프로그램 기술은 효율성이 높은 표절 검사 프로그램 기술 개발을 위한 고도화된 방법이라고 생각됩니다. 본 도전학기 프로그램을 통해 정보보호, 인공지능, 지능시스템, 파이썬 교과목을 심화학습함으로써 학생들의 전공역량이 강화될 것이라고 기대하고, 최종 목표를 달성할 수 있도록 지도교수로서 성심껏 지도하겠습니다.

(소속) 컴퓨터공학전공 (성명) 김지연 (서명 또는 날인)</td></tr>
</table>

1. 도전 배경

표절검사는 주로 논문, 과제 등 다른 출처에서 복사했거나, 인용되었는지를 감지하는 프로세스를 의미한다. 특히 챗gpt와 같은 Open AI의 사용률이 증가하면서 이를 출처없이 사용하는 사례가 늘어나고 있다. 2022년 [1]대학내 과제물 중 절반에 가까운 46.03%가 '표절 위험' 인 것으로 나타났다. 31%이상 표절한 경우 '표절 위험'으로 분류했다고 한다. 이러한 문제를 해결하기 위해, 본 팀은 제안받은 표절검사 프로그램을 효율적이고 정확한 표절 검사로 개선해보려고한다. word2vec, 중심성 계수 및 해시값 비교 알고리즘을 사용할 계획이다. 또한 형태소 분석기를 통하여 문장의 조사를 제외한 명사 시퀀스를 추출하고 단어의 순서와 사용 빈도 등의 여러 시나리오를 구축하여 명사의 유사도와 해시 값을 분석하고, 그래프로 표시되는 중심성 계수의 유사도를 계산 및 단어 비율의 결과를 밴다이어그램으로 표시할 것이다. 그리고 각 개발한 알고리즘과 직접 개발한 크롤러를 이용하여 구축한 데이터셋을 기반으로 기존 표절검사 프로그램보다 효율 및 정확성 있는 표절검사 프로그램 개발을 도전해보려한다.

1) 2023.04.20.,"AI기반 표절검사서비스 '카피킬러'도입 대학 통계" https://www.fieldnews.kr/news/articleView.html?idxno=860

2. 도전 과제의 목표

가. 팀 목표

본 팀의 목표는 도전학기 프로그램에서 제안받은 “표절 검사 프로그램 개발”을 기존 상용 프로그램에서 효율성과 정확도를 향상시키기 위한 기술을 개발하고자 한다. 먼저 직접 개발한 크롤러를 이용해 데이터셋을 구축하고, 중심성 계수, word2vec, 해시 비교 및 단어의 비율을 계산하여 밴다이어그램으로 표시한다. 또한 결과 그래프의 유사도를 계산해 차별화된 표절 검사 프로그램을 개발하려고 한다.

나. 개인 목표

1) 단어 비율 계산 및 그래프 유사도 계산과 밴다이어그램 제작 (박OO)

- 형태소 분석기로 명사 시퀀스 추출 및 불용어 처리하여 밴다이어그램 표시
- 중심성계수 분석으로 나온 그래프 결과의 유사도를 계산하며 전공 역량 강화

2) 해시 값 비교, 데이터 셋 구축을 위한 크롤러 제작 및 인터페이스 설계 (안OO)

- 키워드 입력 시 학술DB에서 데이터 크롤링하여 자체 데이터셋 구축
- 해시 알고리즘을 선택하여 각 문장의 해시 값 추출
- 프로젝트 통합 및 인터페이스 설계를 통한 전공 역량 강화

3) 중심성 계수 분석 및 word2vec 사용 (강OO)

- 데이터의 중심성 계수를 분석하여 단어간 거리 그래프 제작
- word2vec 사용으로 그래프 유사성 계산을 통한 전공 역량 강화

3. 도전 과제 내용

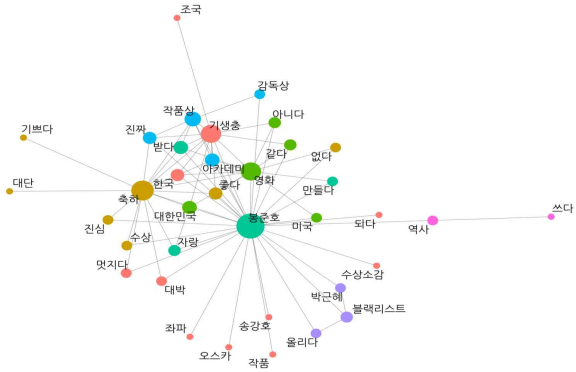
가. 기존 표절 검사 프로그램 개선점 조사

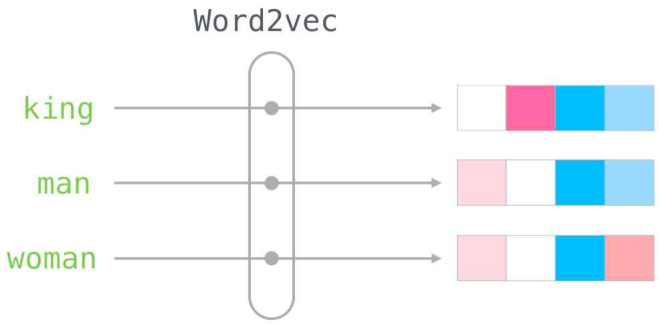
표절 검사 프로그램은 학문적 부정행위를 방지하기 위해 주로 논문, 보고서 등의 내용에서 구축된 DB를 통해 문서 비교를 진행하여 표절 여부를 판단해주는 도구이다. 기존의 표절 검사 프로그램 중 대표적인 프로그램인 ‘카피킬러’를 살펴보면 일반적으로 A4 15~20매의 1000문장 이하의 문서 경우 1-2분내로 검사가 완료 된다고 한다. 하지만 실상 과제물을 제출하기 위해 문서를 업로드하면 5분 이상 지연되는 경우가 발생한다. 또한 카피킬러의 표절 인식 범위는 6어절/1문장 이상의 일치도가 나타난 경우 표절로 판단한다. 이는 유사성 확인을 위한 간단한 규칙 중 하나이며 모든 문장이 6어절 이내로 표현되지 않고 구절이 길어질 수도 있어 정확성이 제한된다고 볼 수 있다. 카피킬러에선 pdf 형식의 문서를 변환한 이미지 파일은 비교가 가능하지만 실제 파일 내의 들어있는 그림 자체의 비교는 제공하지 않아 이미지 표절률을 측정하기 어려운 실태이다. 또한 카피킬러를 통해 특정 표절률 기준만 통과하면 연구 윤리를 지킨것이라는 사회적 분위기가 생성되면서 카피킬러 맞춤형 논문이 생산되는 점이 문제가 될 수 있다.

나. 프로그램 개발 및 데이터 운용

표절 검사 프로그램 제작에 있어서 데이터 운용은 개발 알고리즘 검증 및 학습에 대한 직접적인 영향을 주는 핵심 부분 중 하나로, 정형화된 데이터 셋이 필요하며, 구조화된 방식으로 이를 운용하는 것이 중요하다. 본 팀의 목표인 효율적이고 정확한 표절 검사를 위한 프로그램을 개발하고자 텍스트 비교 및 매칭 알고리즘을 개발하여 중복 단어 및 구조를 효과적으로 식별하고, 정확한 표절 여부를 판별할 수 있도록 할 것이며 추가적으로 중심성 계수와 word2vec을 활용해 문장을 예측하는 방식으로 표절률을 측정할 것이다. 데이터는 자체 개발 크롤러를 활용하여 학습 데이터셋을 구축할 것이며 데이터셋은 비교 문서와 표절 문서 각각에 대한 데이터를 정제하여 학습에 활용할 수

있는 형태로 구축한다. 또한 앞선 알고리즘의 산출물로 나타나는 그래프를 2차 분석하여 심층적인 프로그램을 개발할 것이다.





다. 업무 분장 내용

팀원 성명	소속	담당 업무
박OO	컴퓨터공학전공	- 형태소 분석기로 명사 시퀀스 추출 및 불용어 처리하여 밴다이어그램 표시 - 중심성계수 분석으로 나온 그래프 결과의 유사도를 계산하며 전공 역량 강화
안OO	컴퓨터소프트웨어전공	- 키워드 입력 시 학술DB에서 데이터 크롤링하여 자체 데이터셋 구축 - 해시 알고리즘을 선택하여 각 문장의 해시 값 추출 - API 통합 및 인터페이스 설계를 통한 전공 역량 강화
강OO	컴퓨터공학전공	- 데이터의 중심성 계수를 분석하여 단어간 거리 그래프 제작 - word2vec 사용으로 그래프 유사성 계산을 통한 전공 역량 강화

4. 도전 과제 추진일정

주차	활동 목표	활동 내용	투입 시간
1주차	구축 환경 조사 및 시나리오 제작	박 팀장): 사용할 형태소 분석기 조사 및 선정	6시간
		안 팀원): 해시 알고리즘 조사 및 선정	6시간
		강 팀원): 중심성 계수 조사 및 사전 공부	6시간
2주차	구축 환경 조사 및 시나리오 제작	박 팀장): 클러스터링 조사 및 밴다이어그램 공부	6시간
		안 팀원): 사용할 라이브러리 조사 및 공부	6시간
		강 팀원): word2vec 조사 및 사전 공부	6시간
3주차	팀원별 환경 구축	박 팀장): 형태소 분석기 환경 제작	8시간
		안 팀원): 크롤러 프로세스 개발	8시간
		강 팀원): 시나리오 설계 및 환경 구축	8시간
4주차	환경 구축 및 데이터 수집	박 팀장): 명사 시퀀스 추출 및 데이터 수집	6시간
		안 팀원): 크롤러 개발 및 인터페이스 구축	6시간
		강 팀원): 그래프 구축 프로세스 개발	6시간
5주차	팀원별 알고리즘 개발	박 팀장): 불용어 처리 및 밴다이어그램 제작	8시간
		안 팀원): 크롤러 개발 및 인터페이스 구축	8시간
		강 팀원): 중심성 계수를 통한 그래프 제작	8시간
6주차	알고리즘 개발 및 데이터 처리	박 팀장): 밴다이어그램 제작 및 그래프 유사도 분석	8시간
		안 팀원): 문서, 문장별 해시값 조사	8시간
		강 팀원): 중심성 계수를 통한 그래프 제작	8시간
7주차	알고리즘 개발 및 최적화 진행	박 팀장): 그래프 유사도 분석	5시간
		안 팀원): 해시 값 비교 알고리즘 최적화	5시간
		강 팀원): word2vec을 이용하여 벡터화	5시간
8주차	중간 보고서 작성	박 팀장): 중간 보고서 작성	3시간

		안	(팀원): 중간 보고서 작성	3시간
		강	(팀원): 중간 보고서 작성	3시간
9주차	알고리즘 개발 및 최적화 진행	박	(팀장): 그래프 유사도 분석	3시간
		안	(팀원): 웹 페이지 제작 및 코드 최적화	3시간
		강	(팀원): 벡터 유사도 계산	3시간
10주차	시스템 통합을 위한 최적화 진행	박	(팀장): 코드 유지보수 및 최적화	5시간
		안	(팀원): 코드 유지보수 및 최적화	5시간
		강	(팀원): 코드 유지보수 및 최적화	5시간
11주차	팀원별 알고리즘 및 코드 피드백	박	(팀장): 오류 검증 및 코드 최적화	8시간
		안	(팀원): 오류 검증 및 코드 최적화	8시간
		강	(팀원): 오류 검증 및 코드 최적화	8시간
12주차	코드 성능 재검증 및 프로젝트 통합	박	(팀장): 데이터셋 활용하여 성능 재검증	5시간
		안	(팀원): 데이터셋 활용하여 성능 재검증	5시간
		강	(팀원): 데이터셋 활용하여 성능 재검증	5시간
13주차	프로젝트 통합 및 인터페이스 개발	박	(팀장): 알고리즘 개선 및 프로젝트 통합	8시간
		안	(팀원): 알고리즘 개선 및 프로젝트 통합	8시간
		강	(팀원): 알고리즘 개선 및 프로젝트 통합	8시간
14주차	프로그램 테스트케이스 실행 및 문제 해결	박	(팀장): 프로그램 테스트 및 문제해결	8시간
		안	(팀원): 프로그램 테스트 및 문제해결	8시간
		강	(팀원): 프로그램 테스트 및 문제해결	8시간
15주차	최종 보고서 작성	박	(팀장): 최종 보고서 작성	3시간
		안	(팀원): 최종 보고서 작성	3시간
		강	(팀원): 최종 보고서 작성	3시간

5. 활동 지원비 상세 내역

활동 지원비 신청내역		
항 목	산출근거	금액(원)
자료구입비	- Do it! 쉽게 배우는 R 텍스트 마이닝 - 24,000원	24,000원
회의비	- 회의비= 348,000원	348,000원
등록비	- 대한임베디드공학회 등록비 - 220,000원 * 3명 = 660,000원	660,000원
교통비	- 하양 - 동대구 왕복 기차비 - 6,000원(왕복) * 3명 * 1회 = 18,000	18,000원
항공료	- 대구 – 제주 왕복 항공료 - 150,000원(왕복) * 3명 * 1회 = 450,000원	450,000원
합계(원)		**1,500,000**

6. 과제 수행 후 제출할 수 있는 결과물

도전 학기 활동은 수행하면서 제출할 수 있는 결과물로는 팀 공통 결과물이 있다. 팀 공통 결과물은 기존에 있던 카피킬러와는 차별화된 프로그램을 개발하고 이를 보여줄 수 있도록 웹 페이지를 제작할 것이며, 중심성 계수와 word2vec을 이용해 단어 벡터간의 거리를 구하여 다음 문장을 예측하고 표절률을 검사하는 프로그램이 될 것이다. 그리고 제작한 표절 검사 프로그램을 중심으로 학술대회 논문을 작성하여 대한임베디드공학회 학술대회에 참여하고자 한다.

가. 팀 공통 결과물 : 효율 및 정확도 향상을 목적으로 한 차별화된 표절 검사 프로그램과 웹 페이지, 학술대회 논문