산학협력 프로젝트 수행계획서

과제명	도메인 특화 Knowledge Graph 구축 및 활용 기술 개발		
협력기관명	(주)데이터스트림즈	과제멘토	이동욱 책임연구원
책임교수	백호기	소속	컴퓨터학부
참여인원	(총 07명) 기업체 01명, 참여교수 01명, 대학원과정 00명, 학부과정 05명		
수행기간	2022.03.01.~6.30.(4개월)	유형	중기

추진배경

○ 글로벌 빅테크 기업들의 주도하에, 초거대AI (Hyper-Scale AI)에 대한 경쟁이 이루어지고 있으나, 초거대AI 학습 및 운영은 전 지구적으로 손에 꼽힐만한 기관/기업에서만 가능

○ 여전히 많은 기업/기관은 자신들만의 특화된 영역에서의 AI기술 적용이 필요하나, Transformer 기반으로 학습된 Model은 의미있는 결과 및 정확도를 도출하기에는 어려움.

○ 본 프로젝트는 각 기업/기관들의 특화된 영역들의 데이터를 기반으로, Knowledge Graph를 구축하는 기술을 개발하고, 이를 기반으로 운영되는 서비스를 구축하는 것을 목표로 한다.

목표 및 내용

○ 프로젝트는 각 영역을 담당하는 마이크로서비스 형태로 구성할 것을 권장하며, 운영환경은 Kubernetes 환경을 추천하나, 일반 데스트탑, 서버, 클라우드 환경에서 진행해도 무방

○ 웹 프론트엔드
- Knowledge Graph 생성을 위한 학습데이터를 관리하고, 생성된 Knowledge Graph를 보여주며, Knowledge Graph 기반의 서비스의 동작을 확인할 수 있음.
- React, Vue, Svelte 등의 프레임워크 활용

○ 웹 백엔드
- 입력된 학습데이터를 처리하여 Knowledge Graph 엔진 및 서비스 엔진에 전달

○ KG(Knowledge Graph) Engine
- neo4j와 같은 Graph DB에 연동하여, Knowledge Graph를 생성하고, KG로의 query 등을 처리

○ Service Engine
- 생성된 Knowledge Graph를 기반으로, 도메인에 특화된 서비스를 제공

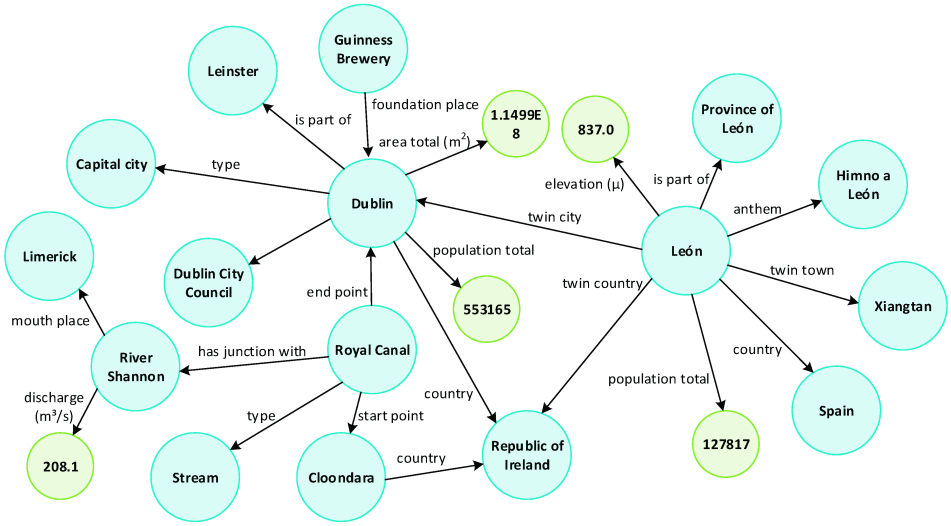
○ 협업 환경 및 CI/CD 프로세스 구축 (optional)
- 분산협력 개발을 위해 Jira, Github 등을 통해 프로젝트를 관리하고, Jenkins 등을 활용하여 DevOps를 위한 CI/CD 환경 구축

기대효과

○ Graph DB에 대한 이해 및 기술 습득

○ Knowledge Graph 구축 및 활용 서비스 개발 능력 배양

1. 과제 목적 및 필요성



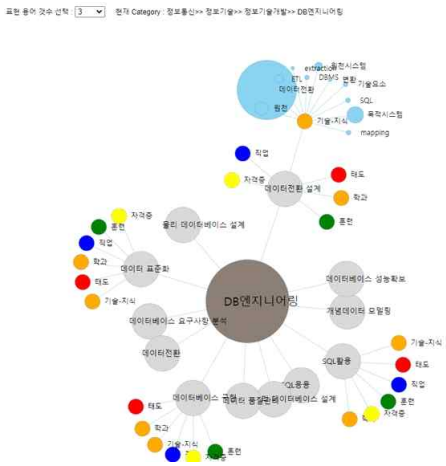
<그림 1. Knowledge Graph>

1. 프로젝트 목적

지식 그래프(Knowledge Graph)는 개별 객체의 데이터를 나타내는 정점과 객체 간의 연관성을 의미하는 간선으로 표현할 수 있는 그래프 형태의 저장된 지식을 의미한다.

지식 그래프는 컨텍스트 내의 링킹 및 의미론적인 메타데이터를 저장하며, 이러한 방식으로 데이터 통합, 통일, 분석, 공유(integration, unification, analytics, sharing)를 위한 프레임 워크를 제공한다.

지식 그래프로 데이터를 표현하는 가장 큰 목적은 그래프 구조가 연관성을 포함한 형태로 지식을 축적하고 전달하는 데 가장 유리한 자료구조이기 때문이다.



<그림 2. 지식 그래프 활용 예시>

그림 2는 '지식 그래프를 기반으로 한 직무 분석' 이다. 본 프로젝트에서는 그림2의 예시와 유사하게 도메인 특화 서비스를 제공한다. 예를 들어서, 지식 그래프가 웹 프로그래밍 공부 지식 그래프라고 한다면, 지식 그래프를 만들기 위해서 여러 가지 데이터들을 넣을 것이다. 사용자의 질문이 '웹 프론트엔드를 공부하려면 어떻게 해야 해?' 이라면 지식 그래프에서 데이터를 찾을 것이고, 간선으로 연결된 노드, 즉 정보들을 출력하여 질문에 응답할 수 있다. 이러한 Knowledge Graph를 구축하는 기술을 개발하고, 이를 기반으로 운영되는 웹 서비스를 구축하는 것을 목표로 한다.

2. 지식 그래프(Knowledge Graph)의 필요성

지식 그래프는 테이블 기반 데이터베이스에 비해 많은 장점을 갖고 있다.

첫째, 명시적으로 정의되지 않은 관계에 대해 추론(Reasoning)이 가능하다. 기존 테이블 기반의 데이터베이스는 테이블에 개체 간의 관계가 정의되지 않은 경우 관계를 유추하는 것이 불가능하지만, 지식 그래프는 기존에 정의된 지식 그래프의 관계를 토대로 새로운 관계를 탐색, 추론하는 것이 가능하다.

둘째, 여러 번의 참조를 통해 대답해야 하는 질문에 대해 지식 그래프는 성능 측면에서 장점을 갖고 있다. 예를 들면, "제안서가 인용한 문서가 인용한 문서를 보여주세요."라는 질문에 기존 데이터베이스는 문서 테이블과 인용 관계를 나타내는 테이블을 상호참조하면서 여러 번의 데이터 접근이 필요하지만, 지식 그래프는 문서들이 인용 관계로 연결되어 있어 몇 번의 접근으로 질문에 대한 답을 찾을 수 있다.

마지막으로, 기존 데이터와 새로운 관계를 갖는 데이터가 추가, 삭제, 변경될 경우 지식 그래프는 아주 간단하게 요청을 처리할 수 있다. 기존 데이터베이스에서는 새로운 타입의 데이터가 추가될 때마다 테이블을 늘리고, 기존 데이터와의 연결 관계를 모두 고려해야 한다. 이때 계산 비용이 많이 들 뿐 아니라, 데이터 정합성의 문제가 생길 가능성이 있다. 지식 그래프는 이미 구축된 지식 그래프에 정점을 추가하거나 삭제하고, 기존 정점 간의 관계를 고려해 간선을 새로 연결하거나 삭제하기만 하면 되기 때문에 관리가 매우 쉽다.

본 프로젝트는 지식 그래프가 대량의 데이터를 처리하는 데 용이하다는 점을 이용하여, 위의 원리를 활용하고 지식 그래프를 적용하여 도메인에 특화된 서비스를 제공하고자 한다.

3. 지식 그래프(Knowledge Graph)의 활용

구글은 지난 2012년 5월 16일 지식 그래프를 구글 영문 검색에 처음 적용하였고, 이를 통해 어떤 토픽에 대해 구조화된 정보와 다른 사이트로의 링크를 제공하여 다양한 소스로부터 축적한 시맨틱 검색 정보를 사용하여 검색 결과를 향상했다. 구글의 지식 그래프는 인터넷상의 각기 다른 출처에서 발견된 이런 사실들을 하나의 개념 지도에 연결된 조각들로 간주한다. 이러한 방식으로 표현된 데이터 네트워크는 훨씬 직관적이다. 뿐만 아니라 주어진 도메인에 있는 각 주체 간의 관계를 명료하게 만들어준다. 이를 통해 사용자는 모든 정보를 맥락에 맞게 접근하고 이해할 수 있다.

지식 그래프는 텍스트 분석에도 활용된다. 지식 그래프는 본문을 보다 정확하게 해석할 수 있도록 배경 지식, 인간과 유사한 개념 및 실체 인식을 제공한다. 분석 결과는 텍스트의 참조를 그래프의 특정 개념에 연결하는 의미 태그이다. 이러한 태그는 더 나은 검색과 추가 분석을 가능하게 하는 구조화된 메타데이터를 말한다. 텍스트에서 추출한 사실을 추가하여 지식 그래프를 풍부하게 할 수 있고, 이는 분석, 시각화 및 보고에 훨씬 더 가치를 부여한다.

도메인 특화 지식 그래프는 특정 용도와 응용 프로그램에 사용된다. 지능형 콘텐츠 및 패키지 재사용, 대응 및 상황 인식 콘텐츠 권장, 지식 그래프 구동 약물 검색, 의미 검색, 투자 시장 인텔리전스, 규제 문서의 정보 검색, 고급 약물 안전 분석 등과 같은 데이터 및 정보 집약적인 서비스 등이 있다.

4. 결론

지식그래프는 사람처럼 생각하고, 동음이의어를 구별할 줄 알며 방대한 양의 데이터를 빠르게 처리할 수 있다. 단순한 키워드 조합의 완전 일치/전방 일치의 방식과 다르게 질의 '의도'를 '이해'할 수 있어 위의 활용 예시뿐만이 아니라 인공지능(AI), 자연어 처리(NL), 자율주행, 항공 교통관제 등 4차 산업과 관련된 여러 분야에서도 빠질 수 없는 중요한 기술이다. 실시간 탐색 및 처리가 가능하기 때문에 인공지능, 자율주행, 항공 교통관제 등 즉각적인 판단을 요구하는 분야에서 앞 다투어 도입하는 추세이다. 따라서 도메인 특화 Knowledge Graph 최신 기술을 적용하여 사용자에게 데이터를 편리하게 제공해줄 수 있는 웹 서비스를 개발하는 것이 본 프로젝트의 목표이다.

2. 과제 내용 및 추진 방법

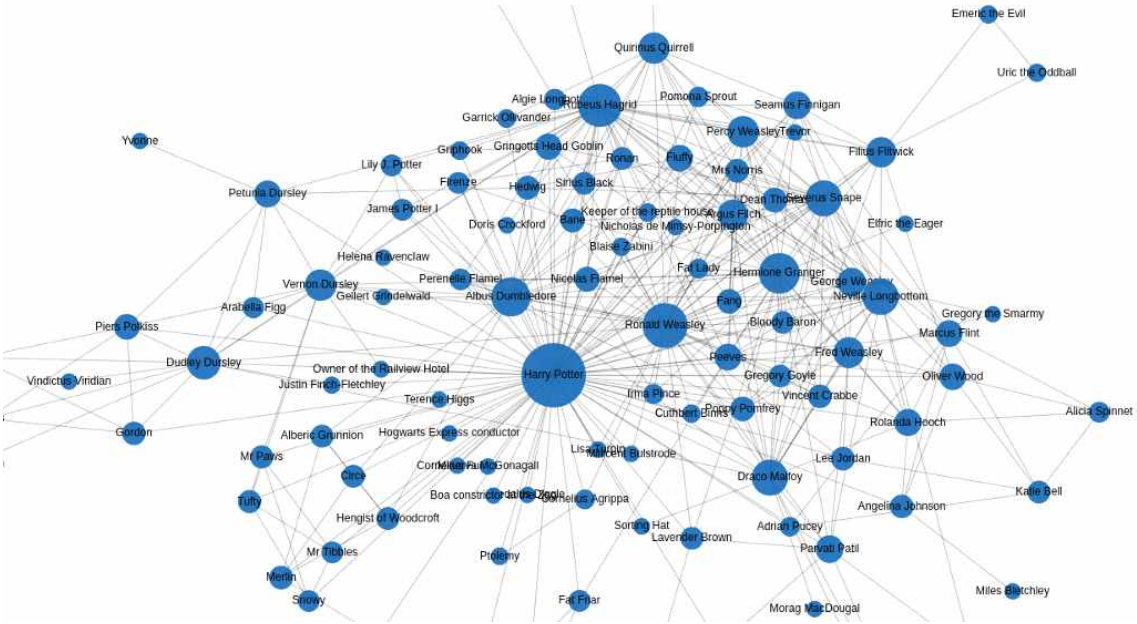
1. 프로젝트 내용

지식 그래프(Knowledge Graph)는 개별 객체의 데이터를 나타내는 정점과 객체 간의 연관성을 의미하는 간선으로 표현할 수 있는 그래프 형태의 저장된 지식을 의미한다. 정보를 지식 그래프의 형태로 저장하면 연관성 높은 정보들을 쉽게 확인할 수 있어 사용자에게 더욱 풍부한 정보를 제공할 수 있다.

지식은 축적된 분야에 따라서 개념과 용도가 달라, 지식 그래프를 구성하는 정점과 간선은 분야에 따라 다른 내용이 포함되어야 한다. 지식 그래프를 활용하여 최종적으로 제공하는 서비스에 따라 활용할 정점과 간선의 종류를 구체적으로 결정하는 과정을 지식 설계라고 한다.

지식 그래프는 여러 데이터 소스로부터 수집한 데이터를 분석하여 구성한다. 문서와 같은 텍스트 데이터로부터 자연어 분석 기술을 이용하여 텍스트에 포함된 내용을 정점과 간선의 형태로 변환하여 축적한다. 자연어 처리 기술의 한계를 극복하고 새롭게 등장하는 구성요소 등을 반영하기 위해 주기적으로 지식 전문가들이 지식 그래프를 점검하고 잘못 축적된 데이터를 수정해야 한다.

지식 그래프를 활용할 수 있는 분야 중 하나는 정보 검색이다. 기존에 정보 검색을 위해 이용한 역색인 방식은 연관된 지식을 같이 표시하기에 어려움이 있다. 검색을 고도화하고 사용자가 원하는 정보를 관련 있는 정보들과 함께 일목요연한 구조로 전달하기 위해 지식 그래프를 구축하는 것이 필수이다.



<그림 3. Knowledge Graph, neo4j, 2021.7.20.>

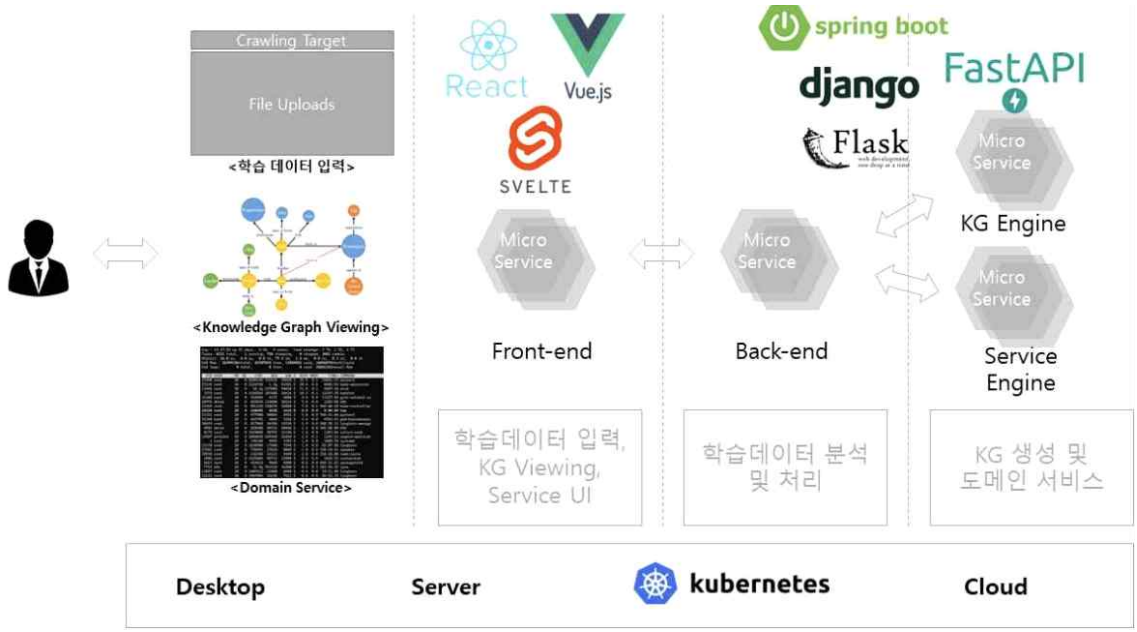
본 프로젝트는 특정 분야의 데이터를 입력하여 도메인 특화 Knowledge Graph를 구축하고, 지식 그래프를 구축하는 기술을 개발하여, 이를 기반으로 웹 서비스를 운영하는 것을 목표로 한다.

2. 지식 그래프(Knowledge Graph) 설계

지식 그래프를 구축하기 위한 기본 설계 절차는 다음과 같다. 우선 지식 그래프 데이터의 사용에 따라 지식 구조를 설계한다. 그 후, 입력된 학습데이터를 처리해야 하나, 대부분 데이터가 텍스트 형식이므로 AI 기반 자연어 처리와 태깅 작업을 거친다. 텍스트 데이터 이외의 기타 데이터는 지식으로 표현할 수 있도록 전처리를 거치고 데이터에 대한 구조화를 진행한다. 데이터의 구조를 바탕으로 모델링을 거친다. 모델링을 통해 데이터 간 연결 관계를 적재한다. 이렇게 형성된 지식 그래프를 조회하고 서비스하는 시각화 단계를 진행한다.

3. 프로젝트 설계 절차

본 프로젝트는 위의 설계 절차를 따라 진행하도록 한다. 우선 웹 서버는 입력된 텍스트 형식의 학습데이터를 분석 및 처리하여 지식 그래프(Knowledge Graph) 엔진 및 서비스 엔진에 전달한다. 지식 그래프 엔진에서는 neo4j와 같은 Graph DB와 연동하여 전달받은 데이터를 지식 그래프로 생성하고, 지식 그래프의 query 등을 처리한다. 여기까지의 과정을 1차 목표로 하고, 프로젝트의 중간 점검과 버그 수정을 진행한다. 지식 그래프의 구축이 완료된다면, 서비스 엔진에서는 생성된 지식 그래프를 기반으로, 도메인에 특화된 서비스를 제공한다. 이와 동시에 웹 프론트엔드에서는 지식 그래프 생성을 위한 학습데이터를 관리하고, 생성된 지식 그래프를 보여준다. 또한, 지식 그래프 기반 서비스의 동작을 확인할 수 있다. 마지막 과정까지 개발을 완료하며 프로젝트를 최종 마무리하도록 한다. 이 모든 프로젝트는 Kubernetes 환경에서 컨테이너화하며 운영을 자동화하여 관리한다.



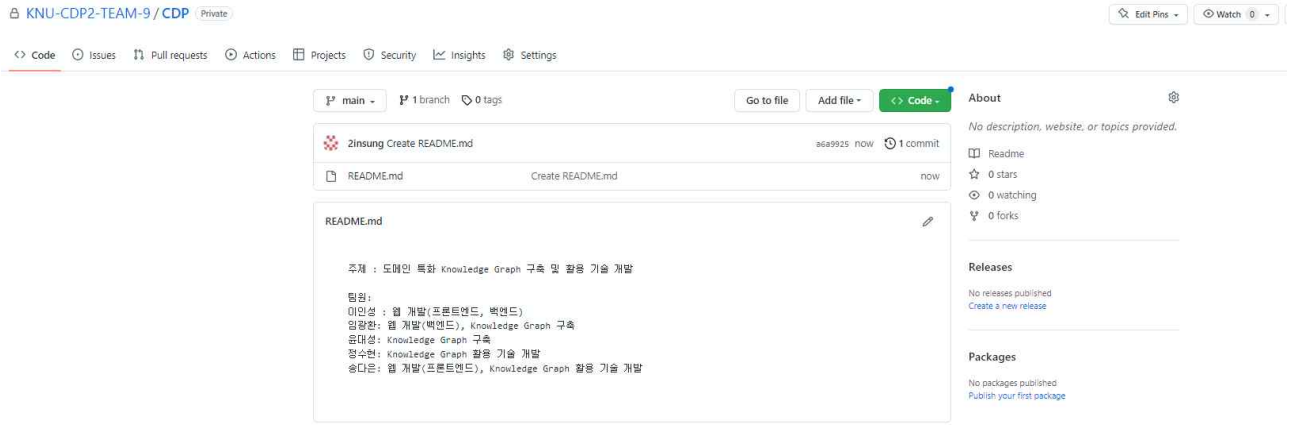
<그림 4. 프로젝트 설계 절차>

4. 개발 툴

프로젝트의 원활한 진행을 위해서 회의 애플리케이션 Notion을 통해 회의록 및 개발 현황을 기록한다.

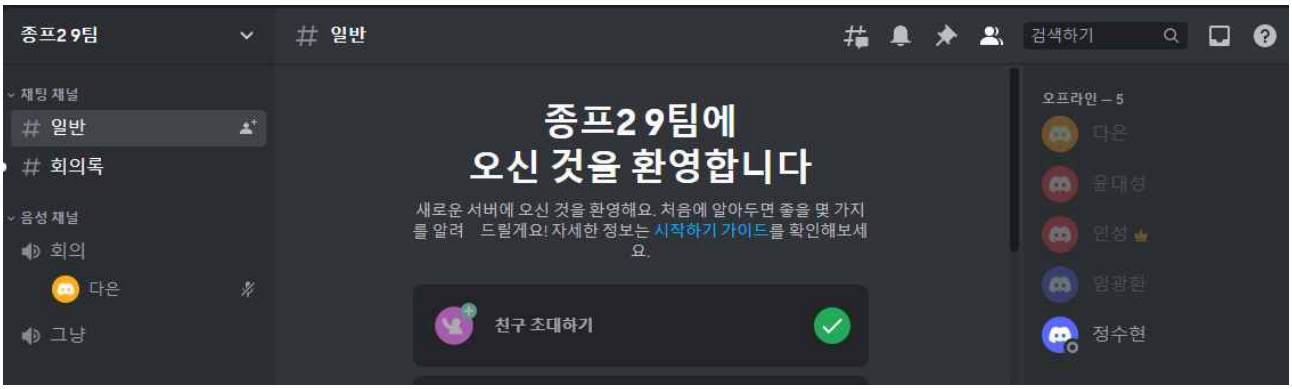
<그림 5. Notion 팀프로젝트 페이지>

팀원들 간의 분산협력 개발을 위해 Github를 통해 프로젝트를 관리하고, 프로젝트 관련 버그 및 개발 방향에 대해서는 Github issue 기능에서 논의한다.



<그림 6. 종합설계프로젝트2 9팀 Github>

온라인 회의는 디스코드를 통해 진행한다.

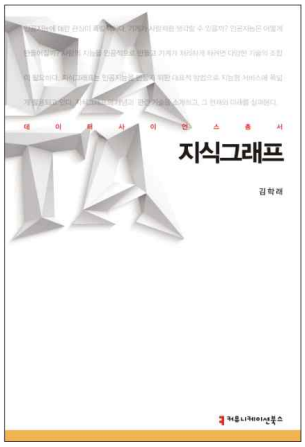


<그림 7. 종합설계프로젝트2 9팀 Discord>

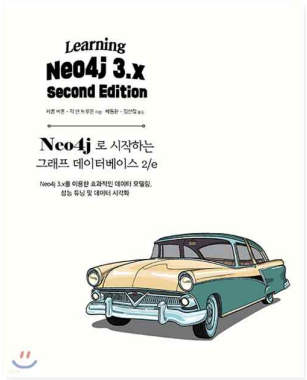
프로젝트의 1차 목표는 입력된 학습데이터를 처리하여 Knowledge Graph를 구축하는 것이다. 이를 위해 사용할 툴은 neo4j이며 사용 언어는 Java이다.

<그림 8. neo4j>

다음은 Knowledge Graph 구축을 위해 참고할 서적이다.



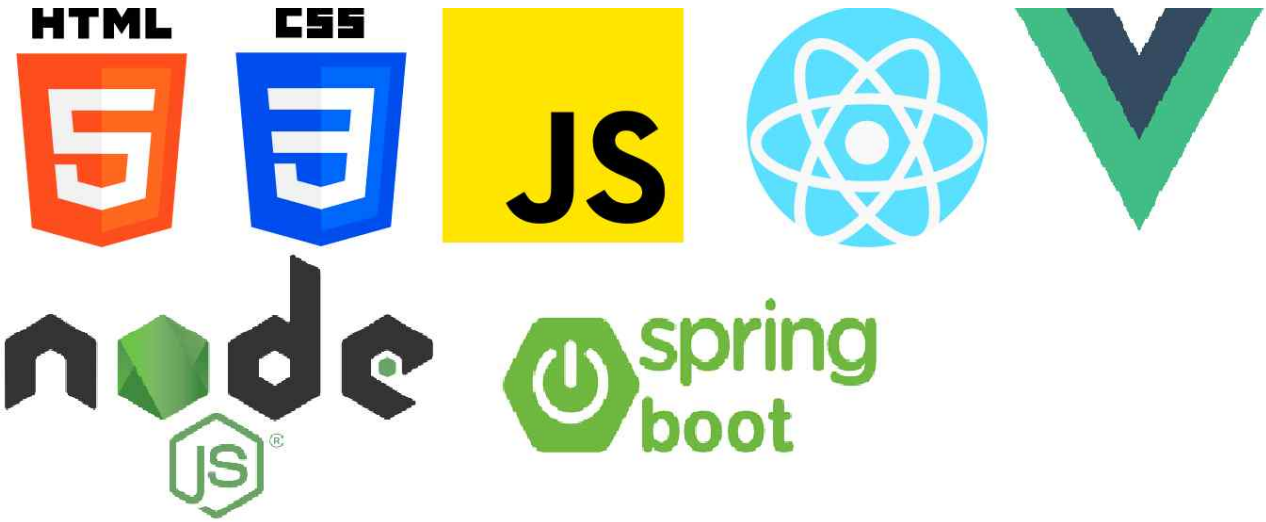
<그림 9. 김학래.『지식그래프』. 커뮤니케이션북스, 2017.>



<그림 10. 제롬 바톤, 릭 반 브루겐.『Neo4j로 시작하는 그래프 데이터베이스 2/e』. 에이콘, 2018.>

다음은 Knowledge Graph 구축을 위해 참고할 문서이다. https://neo4j.com/docs/

웹 프론트엔드 개발 도구는 html, css, javascript, react, Node.js, Vue.js 이다.
웹 백엔드 개발 도구는 spring boot, dJango, Flask이다.



<그림 11. 웹 개발 도구>

3. 과제 추진 일정 및 예산 활용 계획

1. 과제 추진 일정

3월	월	3																													
	일	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
분류	요일	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	담당자
회의&발표	프로젝트 회의																																
	프로젝트 계획 준비 & 발표																																
개발	프로젝트 이해 & 예상이슈 분석																																팀 전원
	웹 UI 설계 & 구현																																송다은, 이인성
	데이터의 구조(KG) 설계																																윤대성, 임광환
	데이터 처리																																윤대성, 임광환

<그림 12. 3월 과제 추진 일정>

4월	월	4																												
	일	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
분류	요일	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	담당자
회의&발표	프로젝트 회의																															
	프로젝트 중간 발표 준비																															
개발	데이터 처리																															윤대성, 임광환
	웹 기능 설계 & 구현																															이인성, 임광환
	데이터 구조화																															송다은, 이인성
	모델링																															윤대성, 임광환
논문	논문 작성																															팀 전원

<그림 13. 4월 과제 추진 일정>

5월	월	5																													
	일	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
분류	요일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	담당자
회의&발표	프로젝트 회의																																
	프로젝트 최종 발표 & 준비																																
개발	모델링																																윤대성, 임광환
	서비스 시각화																																송다은, 정수현
	KG 활용한 서비스 개발																																송다은, 정수현
	개발 최종 완성																																팀 전원
논문	논문 작성 & 제출																																팀 전원

<그림 14. 5월 과제 추진 일정>

6월	월	6																												
	일	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
분류	요일	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	담당자
논문	하계학술발표대회																															팀 전원

<그림 15. 6월 과제 추진 일정>

2. 예산 활용 계획

+)

2023년 한국정보 기술 학회 하계종합학술대회(논문 제출 마감일 5월 14일)

2023년 한국통신 학회 하계 종합 학술 발표회(일정 6/21~6/24)

2023년 한국방송 미디어 공학회 하계학술대회(논문 제출 마감일 5월 12일)(최종 논문 제출 6월 2일)

예산 활용 계획 : 논문 발표 교통비 및 학회등록비에 사용할 예정임.

4. 기대효과 및 활용방안

1.

지식 그래프(Knowledge Graph)는 기존의 테이블 기반 데이터베이스가 명시적으로 정의되지 않은 관계들에 대해 연관성을 찾기 어려운 것과는 달리, 입력된 노드들의 연결(Edge)을 활용해 직접 정의하지 않은 관계에 대해서도 추론 가능하며 이를 통해 새로운 지식의 지속적인 확장이 가능하며 더 자세하고 정확한 정보를 제공해 줄 수 있다.

또한, 지식 그래프를 이용하면 사용자의 명령에 대해서 AI가 '의도'를 파악할 수 있어 보다 정확하게 사용자가 원하는 결과물을 생성해낼 수 있을 것이다.

이러한 점들로 인해, 보다 적은 데이터로도 유의미한 결과와 정확도를 보여주는 AI 개발을 기대할 수 있으며 특화된 영역의 데이터로 구성된 지식 그래프는 각각의 기업/기관들의 자신들만의 특화된 영역에서의 AI 기술을 개발할 수 있을 것이다.

2.

지식 그래프(Knowledge Graph)는 특히 정보 검색의 분야에서 활용성이 높다.

기존에는 정보 검색을 위해 역색인(문서 집합 내에서 키워드의 내용과 위치를 연결)의 방식으로 데이터를 저장하였기에 질의내용에 내한 문서는 잘 나타낼 수 있지만, 연관성이 있는 추가적인 정보를 표시하기에는 어려웠다. 지식 그래프 방식으로 정보를 저장하였을 때 검색 문장에서 사용자의 '의도'를 파악하여 검색을 고도화하고 사용자가 원하는 정보와 관련된 정보를 함께 전달할 수 있기에 사용자가 양질의 정보를 더 쉽게 얻을 수 있을 것이다.

3.

기존의 관계형 데이터베이스가 아닌 영역 특화된 지식 그래프(Knowledge Graph)를 기반으로 사용하는 것만으로도 지식 그래프가 Data Access 부분과 데이터의 추가/삭제/변경 부분에 있어서도 보다 높은 성능을 보이기에 방대한 정보를 활용해야 하는 서비스를 제공하기에도 알맞을 것이다.

4.

영역 특화된 지식 그래프(Knowledge Graph)는 각각의 영역에서 서비스를 구축할 때, 보다 전문적이고 효과적인 서비스를 만드는 밑바탕이 될 것이다.

지식 그래프의 구현을 통해 앞서 설명한 지식 그래프의 장점을 살려 프로젝트에서 지식 그래프를 활용한 웹 서비스를 제공하면 사용자 입장에서 데이터를 편리하게 이용할 수 있게 될 것으로 기대된다.

5. 예상되는 주요 과제성과

1.

논문 발표는 2023년 한국정보 기술 학회 하계종합학술대회, 2023년 한국통신 학회 하계 종합 학술 발표회, 2023년 한국방송 미디어 공학회 하계학술대회 등에 제출할 예정이다.

2.

본 프로젝트를 통해서 SW 특허출원 및 서비스 상용화를 고려하고 있다.

6. 참여인력(세부)

지도교수	소속	컴퓨터학부		성명	백호기
참여인력 (산업체)	기업명	성명	직위	전화	Email
	(주)데이터스트림즈	이동욱	책임연구원	010-9866-7662	dwlee@datastreams.co.kr
과제참여학생	소속(학과)	학위과정 (성별)	학번	성명	담당업무
	컴퓨터학부	학사과정 (여)	2019115861	송다은	웹 개발(프론트엔드), Knowledge Graph 활용 기술 개발
	컴퓨터학부	학사과정 (남)	2020113738	이인성	웹 개발(프론트엔드, 백엔드)
	건설방재학과	학사과정 (남)	2017112918	임광환	웹 개발(백엔드), Knowledge Graph 구축
	컴퓨터학부	학사과정 (여)	2020115972	정수현	Knowledge Graph 활용 기술 개발
	컴퓨터학부	학사과정 (남)	2017117079	윤대성	Knowledge Graph 구축