소프트웨어학부 2020203090 한용옥

주제	주제에 대한 간단한 설명
연합학습 (Federated Learning, FL)	여러 개별 클라이언트가 데이터를 직접 공유하지 않고, 각각의 데이터를 로컬에서 학습한 후 모델의 가중치만 서버에 공유, 서버에서 가중치 집계, 연산 후 클라이언트로 배포하여 전체 모델을 훈련하는 머신러닝 방법

데이터 수집 및 처리

수집 과정	결과
Selenium활용, WOS 다운로드 자동화	총 13485개의 논문 정보를 엑셀로 다운로드

고급 검색에 활용한 쿼리는 아래와 같다

TS=(Federated Learning) AND WC=(Communication OR Computer Science, Artificial Intelligence OR Computer Science, Cybernetics OR Computer Science, Hardware & Architecture OR Computer Science, Information Systems OR Computer Science, Interdisciplinary Applications OR Computer Science, Software Engineering OR Computer Science, Theory & Methods OR Telecommunications)

데이터 전처리 과정 및 결과

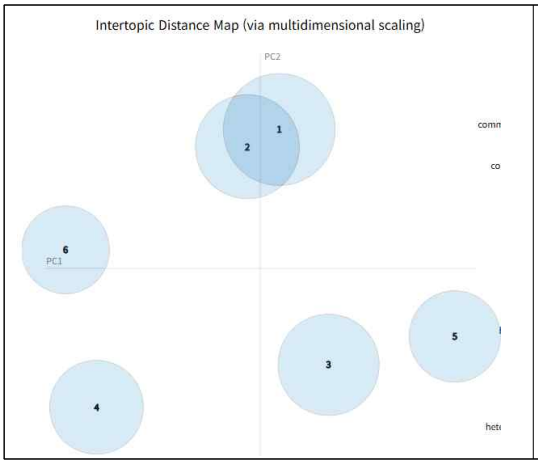
LDA를 사용하기 때문에 BoW 제작을 위해 문장에 대해 단어 단위 전처리를 수행함

데이터 프레임 통합	다운로드한 엑셀의 안 쓰는 열을 제거, 특문 제거 후 데이터프레임으로 병합 Crossref API를 활용하여 DOI 기반으로 논문 발행 날짜 정보 추가 중복 데이터 제거, 유효하지 않은 날짜, 결측치 제거, 특수문자 제거
소문자화, 부호 제거	대소문자는 단어 의미에 영향이 없고 문장부호로 연결된 단어를 토큰으로 분리해도 된다고 판단
토큰화	NLTK TreebankWordTokenizer를 이용해 문장을 단어 단위로 토큰화
표제어 추출	의미가 동일한 단어를 하나로 통합하기 위해 NLTK WordNetLemmatizer를 이용함
불용어 처리	모든 문서에 공통적으로 쓰이는 단어는 토픽 추론에 악영향을 준다고 판단 불용어(NLTK stopwords) 및 논문 초록, 연합학습 분야에서 거의 공통으로 쓰이는 단어('data', 'model', 'federate', 'federated', 'framework', 'learn', 'train', 'propose', 'problem', 'challenge', 'methods', 'server', 'clients', 'client', 'privacy', 'local', 'global', 'however') 삭제
짧은 단어 제거	길이가 2 이하인 단어는 의미가 없다고 판단, 제거
결과	11240개의 논문 정보(제목, 초록, 날짜) 수집 완료

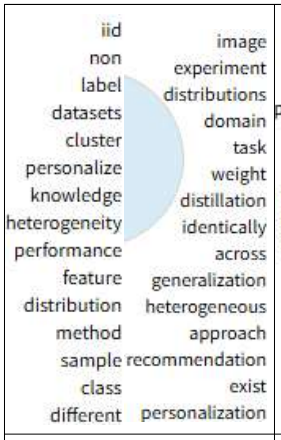
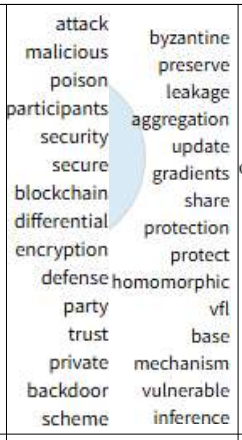
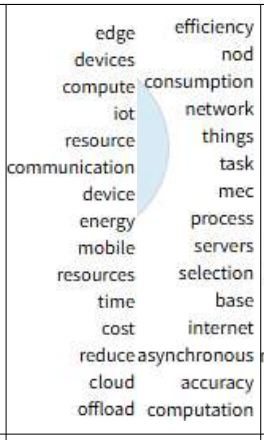
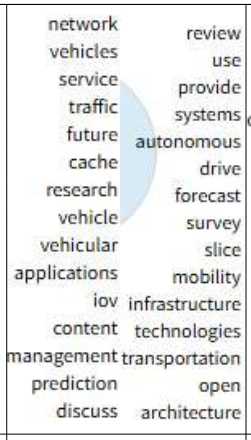
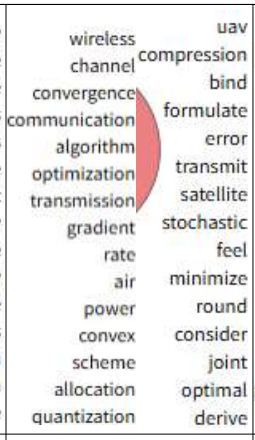
토픽 모델링

사용기법	LDA	구현체	GenSim (gensim.models. LdaMulticore)
토픽 개수	6개	시각화 구현체	pyLDAvis

각 문서의 토픽 할당 및 결과



각 토픽을 저차원에 투영한 그래프이다
원의 크기는 토픽이 할당된 문서의 개수를 의미한다
토픽별 원의 크기가 비교적 유사하므로 이는 모델이 특정 토픽에 과도하게 치우치지 않고 다양한 연구 주제를 고르게 반영하였음을 의미한다

					
1	2	3	4	5	6

pyLDAvis 에서 제공하는 토픽별 단어 관련도 relevance(term w|topic t)=λ*p(w|t)+(1-λ)*p(w|t)/p(w)
λ=0.5의 상위 30개를 뽑아 각 토픽이 자주 사용하는 일반적인 단어와 다른 토픽과 구별되는 특이 단어를 동시에 고려하였다

해당 토픽으로 분류된 논문의 제목 (토픽 당 2개)

1	Personalized Location-Preference Learning for Federated Task Assignment in Spatial Crowdsourcing Mitigating bias in heterogeneous federated learning via stratified client selection
2	Vertical Federated Learning Based Privacy-Preserving Cooperative Sensing in Cognitive Radio Networks FedCut: A Spectral Analysis Framework for Reliable Detection of Byzantine Colluders
3	A Blockchain-based Data Sharing Marketplace with a Federated Learning Use Case Using Knowledge Graphs for Machine Learning in Smart Home Forecasters
4	DAG-based swarm learning: A secure asynchronous learning framework for Internet of Vehicles Towards Efficient Learning Using Double-Layered Federation Based on Traffic Density for Internet of Vehicles
5	Adaptive Control of Local Updating and Model Compression for Efficient Federated Learning Accelerating federated learning based on grouping aggregation in heterogeneous edge computing
6	Synergy conformal prediction applied to large-scale bioactivity datasets and in federated learning DYNAMIC SCHEDULING FOR FEDERATED EDGE LEARNING WITH STREAMING DATA

토픽 연관 단어와 분류된 논문의 제목으로 추론한 토픽 결과

1	non-iid 상황에서도 모델을 효과적으로 학습시키기 위한 주제
2	모델의 보안 위협과 프라이버시 보호 기술에 관한 연구
3	엣지 및 모바일 디바이스에서의 효율적인 자원 활용 및 계산 분산 전략에 대한 연구
4	자율주행, 차량, 이동수단의 네트워크 기반 연합학습
5	연합학습 과정의 통신 비용 절감 및 학습 최적화를 위한 알고리즘 설계
6	의료 분야에서 연합학습을 활용한 질병 탐지와 민감한 데이터 분석

전처리된 데이터에서 5개를 빼서 학습에서 제외하였다 아래 표는 그들의 모델 추론 결과이다

2	A Privacy-Preserving and Verifiable Federated Learning Scheme
5	On-demand Quantization for Green Federated Generative Diffusion in Mobile Edge Networks
2	Privacy-Preserved Credit Data Sharing Integrating Blockchain and Federated Learning for Industrial 4.0
4	dsMTL: a computational framework for privacy-preserving, distributed multi-task machine learning
2	DeSMP: Differential Privacy-exploited Stealthy Model Poisoning Attacks in Federated Learning

dsMTL 논문이 4로 잘못 매핑되었다는 점을 제외하면
나머지 4개의 논문들은 적절한 토픽에 매칭되었다고 볼 수 있다

생성형 AI 적용 및 비교 결과

사용한 생성형 AI	사용한 프롬프트
GPT-4o, 검색 기능 이용	연합학습 연구 분야에 대해 토픽모델링 실시 토픽 개수는 자유

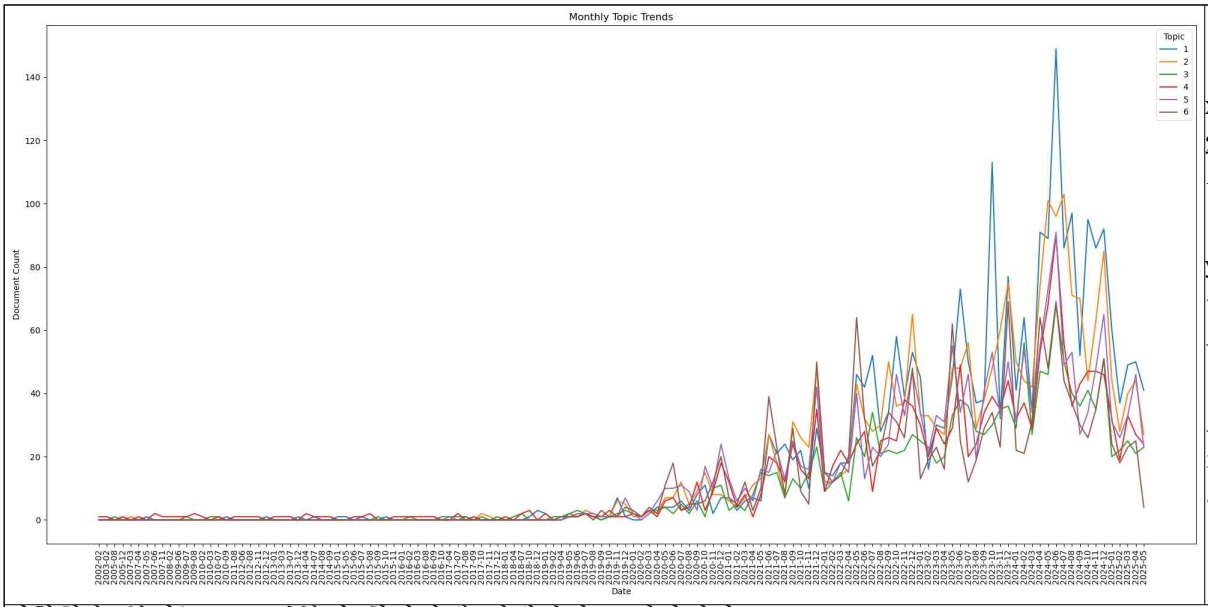
GPT 토픽 모델링 결과

프라이버시 보호 및 보안 강화	차등 프라이버시(Differential Privacy), 동형 암호(Homomorphic Encryption), 안전한 집계(Secure Aggregation) 등의 기술을 활용하여 데이터 프라이버시를 강화합니다.
비동기 및 통신 효율화	비동기 통신, 통신 비용 최적화, 모델 압축 및 프루닝 등을 통해 통신 효율을 향상시킵니다.
이질적인 데이터 및 시스템 처리	데이터 및 시스템의 이질성을 처리하기 위한 개인화 모델, 클라이언트 선택 전략, 메타러닝 기법 등을 연구합니다.
연합 토픽 모델링	분산된 텍스트 데이터를 활용하여 주제를 추출하는 토픽 모델링 기법을 FL 환경에 적용합니다.
응용 분야별 FL 연구	의료: 환자 데이터의 프라이버시를 보호하면서 질병 예측 및 진단 모델을 학습합니다. 금융: 개인 금융 데이터를 보호하면서 사기 탐지 및 신용 평가 모델을 개발합니다. 스마트 시티: 도시 인프라 데이터를 활용하여 교통 흐름 예측, 에너지 관리 등을 수행합니다. 자율 주행: 차량 간 데이터 공유를 통해 자율 주행 모델의 정확도를 향상시킵니다.

비교 분석

공통점	연합학습의 핵심 응용 분야 및 기술적 이슈를 중심 주제로 도출, 아래는 대응관계 토픽1 - 이질적인 데이터 및 시스템 처리 / 토픽2 - 프라이버시 보호 및 보안 강화 토픽3 - 비동기 통신 효율화
차이점	지피티에서는 기존에 볼 수 없었던 분야인 연합 토픽 모델링이 추가됨
	LDA는 자율주행, 의료 같은 응용분야를 각 분야마다 하나의 토픽으로 설정했지만 GPT는 응용분야 전체를 하나의 토픽으로 삼음
	LDA는 토픽 개수가 하이퍼파라미터로 고정이지만, GPT는 토픽 개수를 모델이 판단하여 정함

기타 개인 의견 - 해당 분야 연구 추세



x축
2025-05까지의
월 단위 눈금

y축
해당 월에 발표된
논문 개수

그래프
각 토픽에 따른
연구 추세

연합학습 분야는 2020년부터 활발하게 진행됨이 드러나있다
모든 토픽의 연구추세가 비슷하다 2024년까지는 활발하게 진행되다가 올해 소폭 감소하였다
연합학습 분야에서는 토픽1에 해당하는 non-iid 상황에서의 학습 연구가 인기있음을 확인할 수 있다