텍스트 마이닝을 통한 AI 기술 창업 시장 분석

「미래산업혁명 및 기술창업」과 「프로그래밍 기초 및 실습」강의페어링

전자공학과, 안지연 교수님 지도

연구 목적

독일의 인더스트리 4.0에서 처음 언급된 4차 산업혁명의 파급력으로 인해 기술의 변화가 이루어졌으며 이에 따라 창업시장 또한 급변하였다. 성공적인 창업을 위해 4차 산업혁명의 핵심기술 중 하나인 **AI(인공지능)**와 관련된 연구 결과를 자연어 처리 기반으로 수집하고 체계적으로 지도화(시각화)하여 세계 AI관련 창업의 현황을 알아보고 향후 방향을 전망한다.

-4차 산업 혁명 기반 기술

‘미래산업혁명 및 기술창업’에서의 강의페어링

4차 산업혁명은 디지털 혁명에 기반하여 물리적, 디지털적 및 생물학적 공간의 경계가 희석되는 기술 융합의 시대로 초연결성(Hyper-Connected), 초지능화(Hyper-Intelligent)를 통해 모든 것이 상호 연결되고 보다 지능화된 사회로의 변화이며 대표적인 기술 AI, IOT, big data, Drone이 있다. 4차 산업혁명이 일어난 이래로 성공적인 창업을 위해서 현재 4차 산업혁명으로 인한 기술의 변천과 연구개발 동향 분석이 필요하며 대표적인 기술인 **AI**를 바탕으로 연구를 전개한다.

‘프로그래밍 기초 및 실습’에서의 강의페어링

자연어는 사람이 의사소통에 사용하는 언어로, 컴퓨터에서 사용하는 프로그래밍 언어와 같이 사람이 의도적으로 만든 인공어(constructed language)에 대비되는 개념이다. 자연어 처리에는 자연어 분석, 자연어 이해, 자연어 생성 등의 기술이 사용된다. 자연어 분석은 그 정도에 따라 형태소 분석(morphological analysis), 통사 분석(syntactic analysis), 의미 분석(semantic analysis) 및 화용 분석(pragmatic analysis)의 4 가지로 나눌 수 있다. 현재의 프로젝트 진행에선 형태소 분석을 하며 형태소 분석을 위해 NTLK 자연어 처리 라이브러리를 사용하였다.

연구 방법

체계적 문헌 고찰방법은 반복적인 분석을 통해 유의미한 2차 데이터를 추출하고 분석하는 고찰방법이다. 체계적 문헌 고찰방법의 단계는 연구 목적의 정의 -> 적절한 데이터 수집-> 데이터 추출-> 데이터의 적격성 평가 -> 데이터의 분석 및 결합-> 결론 도출로 이루어진다. 또한 유의미한 결론을 도출하기 위해서는 데이터를 수집할 때의 정확한 기준의 도입을 통해 자료를 분별하는 과정이 필요하다. 본 프로젝트에서 체계적 문헌 고찰방법을 적용해서 논문의 database, 검색 조건, 출판연도 등을 기준으로 데이터를 선별하였다. 또한 문자(비정형) 데이터를 기반으로 텍스트를 분석하여 통계적으로 유의미한 결과를 도출해내는 과정을 **텍스트마이닝**이라고 하며 체계적 문헌 고찰방법을 통해 수집된 데이터를 텍스트 마이닝을 통해서 세계의 AI 기술 경향을 파악하고 창업과 관련해서 유의미한 결과를 도출한다.

연구 과정

1 단계: 데이터 수집

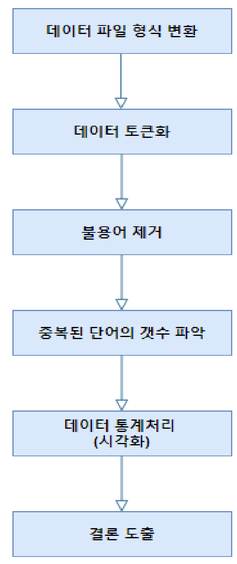
체계적 문헌고찰을 하기 위해 적절한 데이터의 수집을 하기 위해 scopus에 등재된 논문을 대상으로 1차 수집을 한 후에 필터링 기준을 설정하여 필터링 후 다시 자료를 처리하였다. scopus 대상의 논문으로 한 이유는 해외 논문들을 대상으로 하여 보다 포괄적인 연구결과를 얻으려고 했기 때문이다. 논문을 검색한 기준은 다음과 같다. 첫째, Title(제목), Abstract(초록), Keyword(핵심어)에서 AI의 유사어가 포함된 논문이며 둘째, review article 이다. review article을 기준으로 한 이유는 여러 연구논문을 대상으로 분석을 한 논문이 review article이므로 1개의 review article에 다른 여러 연구논문이 집약적으로 포함되어 있다고 판단했기 때문이다. 위와 같은 기준으로 검색했을 때 총 6,417개의 논문이 검색되었다.

2단계: 데이터 필터링

1차 수집이 끝난 자료 중 논문이 출판된 년도를 기준으로 2013~2021년도에서 벗어나는 논문들을 모두 제외하였다. 본 연구의 목적은
4차 산업혁명의 영향을 알아보는 것이므로 독일의 인더스트리 4.0이 본격화된 2013년 이후의 자료를 수집하는 것이 연구목적에 적합하다.
필터링 이후 논문은 총 4,634개가 검색됐으며 논문의 제목과 출판연도를 통해 텍스트 마이닝 처리를 하기 위해
검색된 모든 논문의 정보를 제목과 출판연도만 포함하여 csv 파일 형식으로 추출했다.

3단계: 텍스트 마이닝 구현

자연어 처리 기반의 텍스트마이닝을 구현하기 위해 python 언어의 기반으로 인터넷으로호환이 가능한 google colab환경에서 설계하였다.
자연어 전처리 과정을 위해 NLTK 라이브러리,자료 통계처리를 위해 pandas 라이브러리, wordcloud를 그리기 위해 wordcloud 라이브러리를 사용했으며
다음과 같은 과정으로 이루어졌다.
1. 추출된 Data를 포함하는 csv 파일 불러오기=> 2. csv 파일 list에 data 저장 => 3. list에 저장된 data에서 제목 정보추출 후 단어 토큰화 작업 수행
=> 4. 토큰화 된 단어 중 불용어 처리 및 표제어 추출 => 5. 단어의 집합 중 중복되는 단어의 빈도 파악 => 6. 중복빈도 높은 단어 순으로 단어 배열 후 그래프화
=> 7. 출판년도 순으로 논문 개수 그래프화 => 8. 단어의 집합을 통해 빈도 높은 단어 기반으로 wordcloud 이미지 생성



-텍스트 마이닝 흐름도

data

	Title	Year
0	Artificial intelligence (AI) applications in a...	2021
1	A survey on time-sensitive resource allocation...	2021
2	Applications of artificial intelligence in COV...	2021
3	Adjunctive cytoprotective therapies in acute i...	2021
4	Clinical applications of artificial intelligen...	2021
...	...	...
4629	Exploring the Role of Artificial Intelligence ...	2021
4630	Comparison of Artificial Intelligence Techniqu...	2021
4631	A review on selective L-fucose/D-arabinose iso...	2021
4632	Artificial Intelligence for the Future Radiolo...	2021
4633	Research progress of artificial intelligence i...	2021

4634 rows × 3 columns

-선별된 데이터 입력값

연구 결과 및 의의

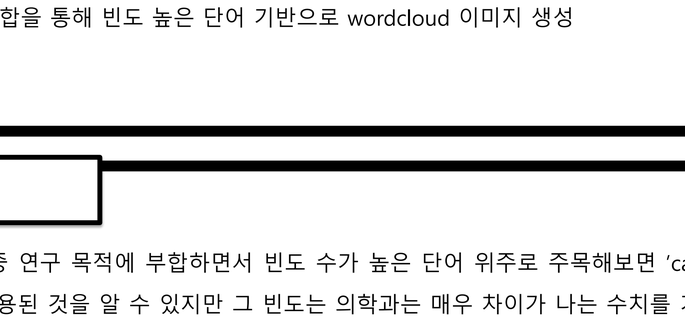
논문의 제목에서 불용어를 제외한 핵심어를 추출한 결과 중 연구 목적에 부합하면서 빈도 수가 높은 단어 위주로 주목해보면 ‘cancer’, ‘stroke’, ‘patient’ ‘covid’ ‘scloliosis’ 등등 다양한 의학 용어들의 집합이 빈도 수가 압도적이었으며 ‘social’, ‘society’,’iot’ 등 의학 이외에도 다른 분야의 기술에 AI이 사용된 것을 알 수 있지만 그 빈도는 의학과는 매우 차이가 나는 수치를 기록했다. 의학 용어 중에 ‘cancer’, ‘scoliosis’, ‘stroke’ 과 같은 병의 진단명을 나타내는 단어가 빈도 수가 높았는데 이 단어들과 더불어 ‘clinical’, ‘detection’, ‘prediction’, ‘potential’ , ‘diagnosis’와 같이 잠재적이고, 질병을 판단하고 예측한다는 의미를 가진 의학용어 또한 더불어 많은 빈도를 기록했다. 그 이외에 ‘iot’, ‘image sensing’. ‘enery harvest’ 같은 다른 신기술의 빈도는 50회를 넘지 못하였다.
의학 분야에서의 AI의 적용되는 사례를 알기 위해 세부 범주표를 만들어서 의학 용어와 함께 더불어 쓰이는 단어의 경향을 파악한 결과 병의 진단명과 ‘처치’,’ 약’ , ‘탐지’ 이 세 단어의 중 한 단어의 조합으로 이루어진 경향이 높았다.
그 이외에 수술을 뜻하는 ‘surgery’가 58회를 기록했으며 구체적으로 논문을 찾아본 결과 AI 기술을 통해 만들어진 로봇으로 수술하는 논문들로 확인되었다. 또한 의학 연구계통인 병리학, 영상의학을 뜻하는 ‘pathology’, ‘radiology’와 같은 단어들도 45회 정도로 공학 계열 범주의 단어들보다도 높거나 같은 빈도를 보였다. 이를 통해서 AI 기술을 이용해 원인 불명이거나 특정 잠재적인 질병을 진단하는 기술, 응급 수술에 대한 처치, 로봇을 이용한 수술, 또한 병리학과 영상의학 대한 연구가 많았다고 판단했으며 실제로 AI 기술을 통해 stroke(뇌졸중)을 판단할 수 있다는 것도 잘 알려져 있다. 또한 2020년에 처음 발생한 신종 바이러스 호흡기 감염질환 코로나의 영향으로 ‘covid’, ‘pandemic’ 단어 또한 찾아볼 수 있었다. 이처럼 AI 기술의 연구는 다른 용도보다 현재 의학적인 용도로 연구개발이 활발히 되고 있으며 AI 를 통해 **특정 질병을 판단하는 즉 영상의학을 위한 툴을, 질병에 대한 처치 혹은 약 투여, 응급 수술을 위한 수술 로봇 개발을 하는 것이 현재 세계 ai 관련 창업의 현황이라고 생각한다.** 또한 논문의 수를 정렬한 결과 2013년을 기점으로 기하급수적으로 논문의 숫자가 늘어나는 모습을 볼 수 있는데 이를 통해 4차 산업혁명의 기술의 연구빈도가 계속해서 높아지며 앞으로도 4차 산업기술이 세계의 시장에 영향을 줄 것이라고 전망할 수 있다.

비고	의학	횟수
암	cancer	384
뇌졸중	stroke	273
급성의, 질병 단어와 함께 사용	acute	207
치치, 질병 단어와 함께 사용	treatment	204
가슴	breast	198
질병	disease	195
임상의	clinical	194
영상, 질병 단어와 함께 사용	imaging	
코로나	covid	168
진단, 질병 단어와 함께 사용	diagnosis	161
치료	therapy	157
약	medicine	144
척추측만증	scoliosis	139
건강	health	136
원인 불명의, 질병 단어와 함께 사용	idiopathic	125
탐지	detection	121
약	drug	108
부신	adrenal	68
잠재적인, 질병 단어와 함께 사용	potential	64
폐	lung	58
수술	surgery	58
유행병	pandemic	55
예견	prediction	80
수용기	receptor	55
내분비	endocrine	53
내시경검사	endoscopy	50
혈관 내의	endovascular	47
심혈관의	cardiovascular	46
병리학	pathology	46
영상의학	radiology	45

-고빈도 단어 세부 범주

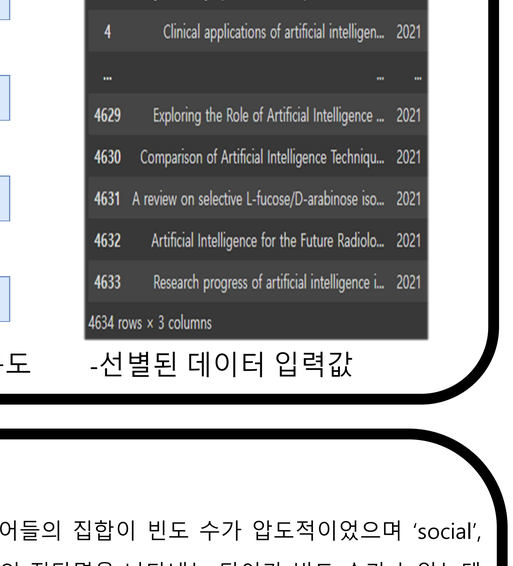
	공학	횟수
	science	47
	computational	47
machine learing의 인공신경망에서 nerual	neural	48
	sensing	46
	energy	45
	internet	41
imaging과 함께 사용되는 경향	classification	39
	robot	33
internet of things	iot	29
	robotics	23

-고빈도 단어 세부 범주



-논문 출판연도 경향(2013~2021)

-고빈도 단어 경향



-고빈도 단어로 생성된 wordcloud

참고 자료

이철주, 최종인 , 우리나라 대학의 기술사업화 영향요인 연구 : 국내 논문에 대한 체계적 문헌 고찰,기술혁신학회, 2019년 2월, p 50~84

김대호, 4차 산업혁명, 커뮤니케이션북스, 2016년

김성근, 조혁준 , 강주영, 학술연구에서의 텍스트마이닝 활용 현황과 주요분석기법, 한국엔터프라이즈아키텍처학회, 2016년

4차 산업 혁명 기반 기술 이미지 출처 - /www.embedded-computing/white-papers/white-0-challenges-solutions-storage-devices/