국외출장 결과보고서

1 출장 개요

□ 출장목적

○ 비정형 빅데이터 국가 전문기관의 역할 논의 및 기계학습과 데이터 윤리 이슈 논의

□ 과제명

○ [일반19-013-00]빅데이터기반 보건복지정책과 기술간 융합체계 구축

□ 출장기간

○ 2019.08.18.(일) ~ 2019.08.24.(토)

□ 출장국가(도시)

○ 영국(런던, 맨체스터)

□ 출장자

○ 오미애 연구위원

□ 일정요약

일자	국가(도시)	방문기관	면담자	주요 활동상황
08.18. (일요일)	한국-런던			출국
08.19. (월요일)	런던	The London School of Economics and Political Science, Statisitcs	Dr. Yining Chen Assistant Professor	Machine Learning 활용사례논의
08.20. (화요일)	런던	The London School of Economics and Political Science, Statisitcs	Dr. Yining Chen Assistant Professor	알고리즘의 편향성 및 데이터 윤리 이슈 논의
08.21. (수요일)	맨체스터	The National Centre for Text Mining(NaCTeM)	John McNaught (Deputy Director) Professor	최신동향을 파악하고, 비정형 빅데이터 분석이 실질적인 정책연구에 활용될 수 있는 방안에 대한 토론과 아이디어 공유
08.22. (목요일)	맨체스터	The National Centre for Text Mining(NaCTeM)	John McNaught (Deputy Director) Professor	비정형 빅데이터 software tool 개발, 세미나 및 컨퍼런스, 워크숍 개최, 텍스트 마이닝 관련 자료 출판 등의 역할 및 체계 논의
08.23.- 08.24.	런던-한국			입국(한국 08.24일 도착)

2 출장 주요내용

①	기계학습 활용사례 논의
일 시	08.19(월요일) 11시~3시
장 소	런던, LSE연구실
참석자	Dr. Yining Chen(LSE) , 오미애(KIHASA)

* 요즘들어 Pin-tech technique 가 중요해지고 있음.

* 5년전 network 분석에서는 topic model 이 대중적이지 않았지만, 요즘엔 topic model을 활용한 분석결과를 많이 볼 수 있음. clustering, textmining 에서도 기계학습 방법을 많이 활용하고 있음

* Computation time 은 점점 중요해지고 있지 않음. 요즘은 R 패키지를 개발할 때 C나 포트란 기반으로 코드를 짜고 효율적으로 패키지를 개발하기 때문에 예전처럼 각 방법론마다 computation time 이 중요하지 않음.

* LSE의 경우, capstone project 가 있음. 여기에서 기계학습을 포함한 통계적 방법론을 적용하여 민간, 공공기관과 같이 데이터 분석을 하고 있음. 현재 10개의 프로젝트가 진행되고 있으며, 각 프로젝트 당 2명이 담당

- TASCO의 프로젝트에서는 self scan 데이터셋으로 분석을 한 경험이 있음. 이 데이터셋으로 recommendation system을 개발하고자 하였으며, shopping item을 분석하고, 어떤상품을 사고 그 뒤에 뭐를 스캔하는지에 대한 심층분석을 하였음.

- MS 프로젝트에서는 다른 company와 주고받은 e-mail 의 communication issue를 분석하였음. 이 분석은 업무의 효율성을 높여주는 작업임.

- airlines 프로젝트에서는 best time promotion을 찾기 위해 데이터 분석. travel pattern도 데이터로 심층 분석

* 데이터 분석 시, 기밀문서들이 많기 때문에 프로젝트에서 학생들이 활용할 수 있는 데이터로 만드는 것이 중요하고 PAPER에 쓸 수 있게끔 조율

* 프로젝트 진행 시 학생과 회사를 연결해주고, check point presentaion을 할 때 관여.

* 프로젝트 결과물에서 회사는 기존의 모델보다 더 나은 모델이면 만족해 하는 반면, 학교 입장에서는 paper를 고려할 때 new model을 개발해야 하는 고민이 있음.mismatching 발생

* 우리나라의 기계학습을 이용한 정책활용 사례 공유

- 복지사각지대 사례 Data
  * social security information system (Social security information service)
  * households that had experienced power disconnections (Korea Electric Power Corporation)
  * households that had experienced water disconnections (The office of waterworks)
  * unemployment information (Korea Employment Information Service)
  * people who attempt suicide (Emergency medical center)
  * and so on.

=> Link by unique ID(like social security number)

- 위기아동발굴 사례 Data
  * absences from schools (Ministry of Education)
  * lack of medical service records on children (National Health Insurance service)
  * children with abused experience. (National child protection agency)
  *and so on.

=> Link by unique ID(like social security number)

②	**알고리즘의 편향성 및 데이터 윤리 이슈 논의**
일 시	08.20(화요일) 10시~4시
장 소	런던, LSE연구실
참석자	Dr. Yining Chen(LSE) , 오미애(KIHASA)

* 데이터 윤리의 대표적인 사례가 automatic driving 일 것임. 여기에서 가장 중요한 것은 data collection일 것임. 사고가 날 경우, 데이터 수집과정의 문제인지, 분석의 문제인지, 기계결함의 문제인지. 책임은 누가 져야 할지 생각해보아야 할 것임. 또한, 운전자의 데이터도 수집이 될텐데 정보수집에 대한 동의도 고려해보아야 함. 이런 부분에서 개인정보 및 윤리적 이슈. 풀어야 할 문제가 많음

* 특히 보건복지분야의 경우, human being을 생각해야 함

* health application에서 데이터 수집단계를 보면 high level, low level 로 나뉘어질 수 있는데, 제대로 masking안하면 누군지 알 수 있음. 따라서 수집된 데이터에서 한번 더 개인정보가 노출되는지를 확인

* 통계학자, 수학자, AI 프로그래머가 윤리적 이슈를 고민해야 하는지 논의

* 빅데이터 분석·활용에서 알고리즘은 중요한 부분을 차지하는데, 알고리즘의 불공정성, 편향성 등의 윤리적인 문제에 대한 고려는 모든 사람에게 공정하고 포용적인 시스템을 만들기 위해 꼭 필요한 일임을 공유

* 모델링을 하는 프로그래머들은 알고리즘의 투명성은 담보할 수 있어야 하며 공정하게 분석과정이 이루어질 수 있도록 노력해야 함

* 합리적으로 생각하며 모델링 할 수 있어야 함

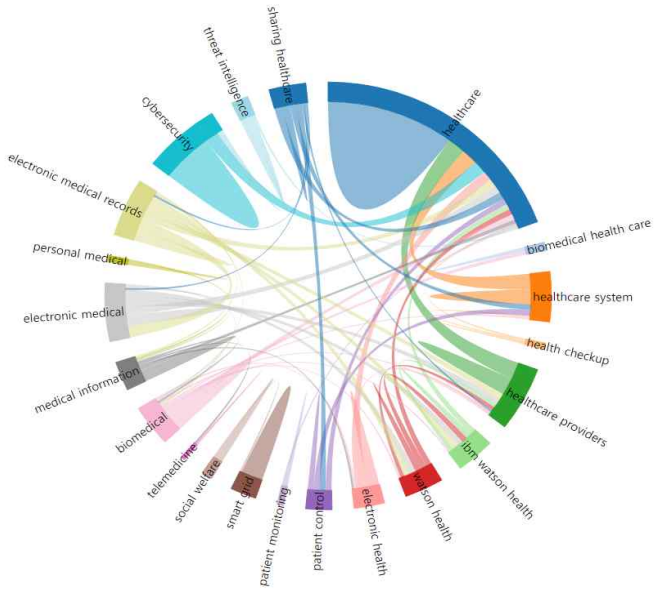
- 알고리즘의 편향성 결과는 데이터 수집, 분류, 이용 과정에서 일어날 수 있지만 어느단계에서 발생하는지는 각 단계별로 검증해보아야 함

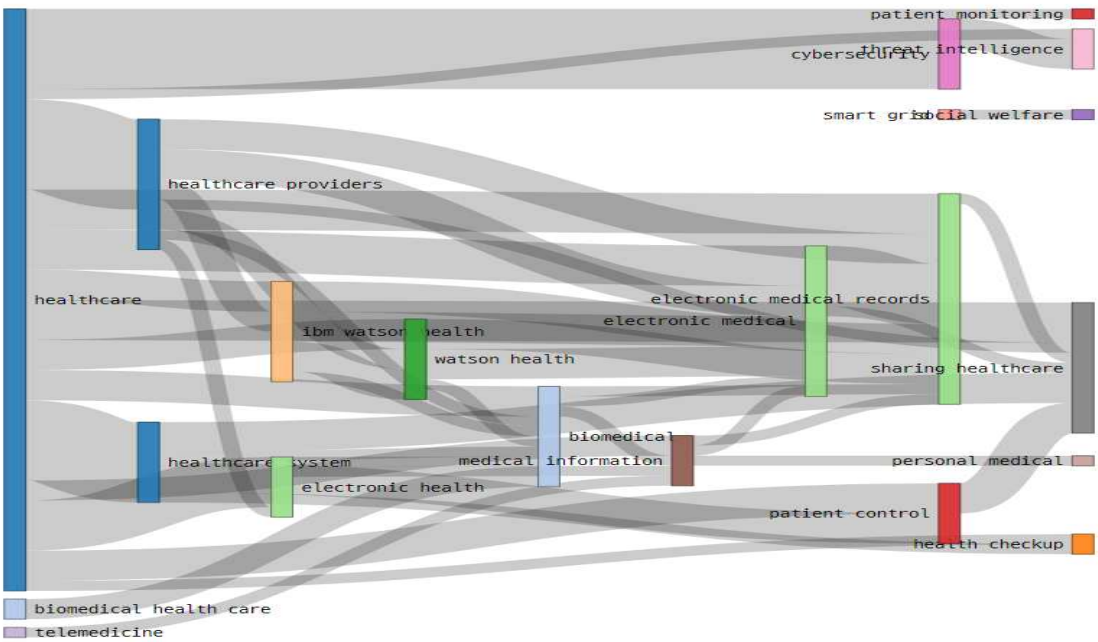
* 영국 정부는 '데이터윤리혁신센터(Centre for Data Ethics and Innovation)가 설치되어 있음. 이 센터에서 관련 이슈들에 대한 논의들이 이루어지고 있음

③	**텍스트자료의 정책연구에서의 활용방안과 아이디어 공유**
일 시	08.21(수요일) 10시-5시
장 소	맨체스터, 맨체스터대학 연구실
참석자	John McNaught(맨체스터대학), 오미애(KIHASA)

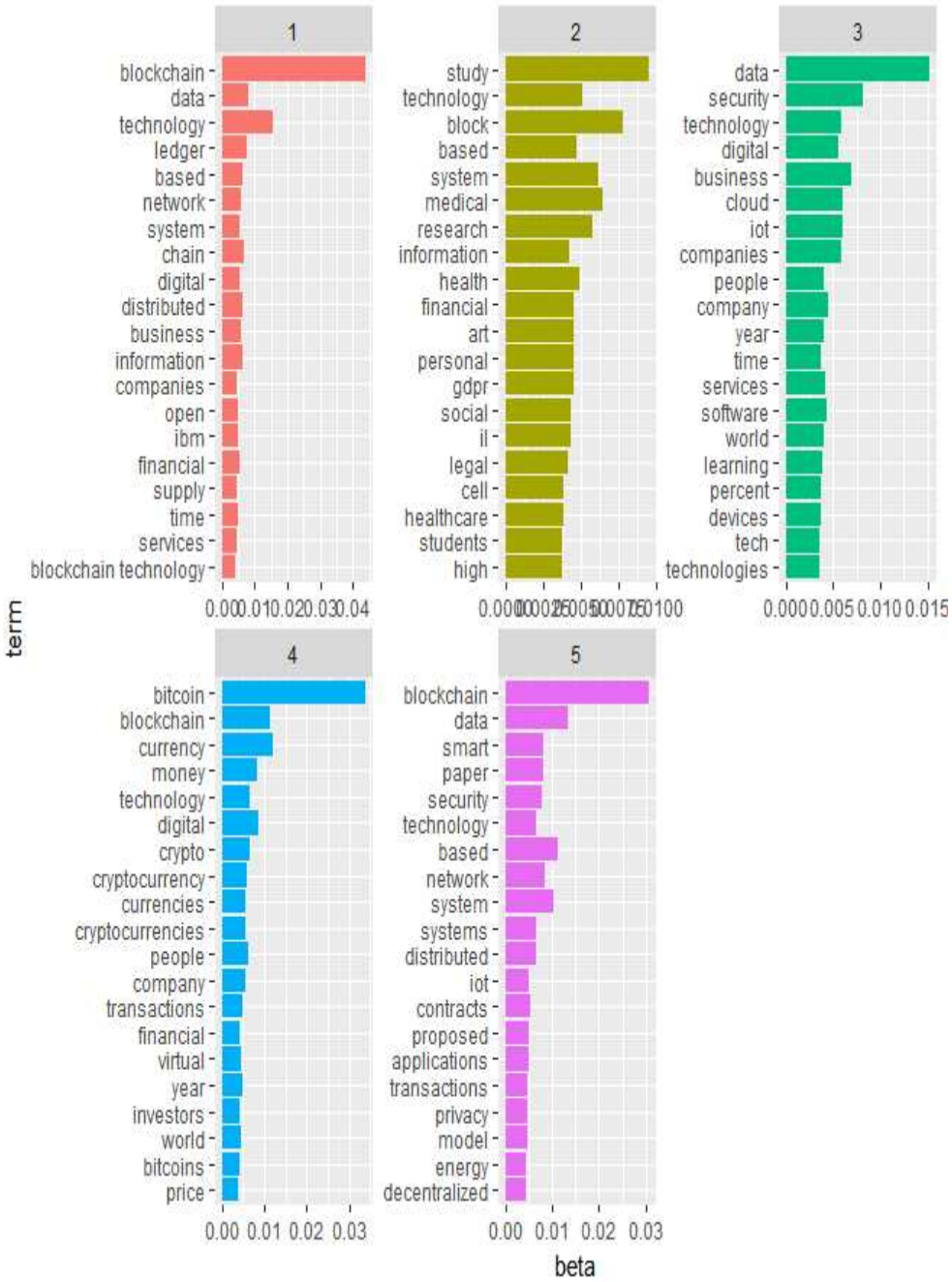
* 작년 소셜 빅데이터 기반 보건복지동향분석 연구결과 공유

블록체인과 보건복지 이슈(영문 문서) 결과를 공유하며, 한글문서 수집과 함께 영문문서 수집의 필요성 논의





영문문서의 topic model 결과



* NaCTeM(the national centre for text mining)의 설립 과정 설명

처음에는 bio 쪽에서 시작. 2003년에 textmining에 Biomedicine funder 들이 관심을 가졌고 natural language processing 에 대한 연구가 본격적으로 시작됨
textmining center가 많았는데, 서로 경쟁이 있었음. 뜻있는 몇몇 사람들이 의견을 모아 경쟁이 아니라 협력할 수 있는 체계를 만들어나가려고 함. 처음엔 학계(academy)쪽에서 만듬

나라에서 fund가 들어온 것은 2012년임. 이때부터 재정이 안정적

유럽 안의 research project 들을 함께 수행했으나, 브랙시트로 EU research project 에 영국의 연구자가 참여하는 것이 어려운 상황임

* Textmining 의 특징 논의
- textmining system 이 모든걸 다 할 수는 없음. 특정 문제를 풀 수 있는 시스템이고, customization 이 필요
- 우선적으로 practical system을 구축하고 센터의 학생들 및 연구원들이 먼저 사용. 수정되어야 할 점들을 우선적으로 반영
- 새로운 tool 만들 때 그 결과에 대해 specific question에 대답할 수 있는 전문가 필요
- Infrastructure의 경우 complex system으로 software 엔지니어링 전문가도 필요.
- sentence split 은 domain마다 다르기 때문에 매우 어려운 작업임.
- dictionary는 old item 으로, most common name을 찾는 것이 중요함. 예를들어 사전에는 인플루엔자로 나오지만 뉴스나 기사에는 인플루로 검색. 이 두 단어를 같은걸로 일치시키는 작업들이 중요함.

* project 관련
- health risk 관련 프로젝트에서 health insurance 회사에서 health record를 분석하여 cost를 분석하고자 하는데, health record 정보(데이터)를 얻는데 비용과 시간 듬
- health area 에서의 ethics 매우 중요. 정보를 익명화 해서 분석. 3년의 프로젝트에서 데이터 access하는데 2년이 걸림

* scientific article의 경우 data 얻기가 힘듬
- copyright 이 중요한 문제임. 주요 저널에 유료(closed) 저널들이 많음. text 를 수집하기 힘듬.
- 2014년에 copyright 법을 개정하여 제목, abstract 는 수집될 수 있도록 함. 상업적 목적은 안됨.
- 정책영역에서 textmining활용하기 위해서는 legal team에 자문을 받아야 함. 우리나라 상황은 어떠한지 살펴보아야 함.
- image는 art 라서 figure쪽에서 보아야 하고 text는 copyright쪽에서 봐야 함.

* textminig에서의 향후 이슈
- copyright 문제는 풀어야할 숙제임
- 기술이 어디로 흐르고 있는 지 알아야 함. 최근 크게 shift 됐음. 머신러닝, NN, word embedding 등 빠르게 변하고 있는 흐름을 알아야 함.

* textminig에서 중요한 것은 언어(language)임. 결과가 나왔으면 왜 결과가 좋게 나왔는지 설명할 수 있어야 함. computational linguistic team 이 중요하고 필요함

④	비정형 빅데이터의 software tool 개발 및 센터 역할/ 체계 논의
일 시	08.22(목요일) 11시-3시
장 소	맨체스터, 맨체스터대학 연구실
참석자	John McNaught(맨체스터대학), 오미애(KIHASA)

* NaCTeM에서의 역할
- good documentation 필요. 많은 사람들이 센터에 있기 때문에 이직이 발생할 경우 이전에 있던 사람이 무슨일을 어떻게 했는지에 대한 자세한 설명 필요. 프로그래머의 경우 code에 대한 설명 필요.
- NaCTeM 사람들에게 그들이 무엇을 하고 있는지에 대한 확신을 심어주어야 함
- 센터는 새로운 tool 개발을 하고 있으며 potential user들과 팀을 이루어 많은 대화를 함. 원하는 게 무엇인지, 어떻게 리서치하기를 원하는지. 이런것들이 requirement가 됨. user를 팀에 속하게 하여 funding을 주면 협업할 수 있는 구조로 만듬.
- user는 NaCTeM에 아이디어와 데이터를 제공하고, 결과를 평가(evaluate)함

* 센터의 plan은 plan of work 이라고 볼 수 있음. 누가 어떤일을 할 것인지 각 구성원들의 직무를 잘 정의하고 하는 일을 보아야 장기적인 플랜을 짤 수 있음

* 하는 연구에 대해 check point만들고, 하고 있는 일에 대해 평가도 함. evaluation measure 가지고있어서 user들에게 평가해달라고 함. recall이 들어오기도 함.

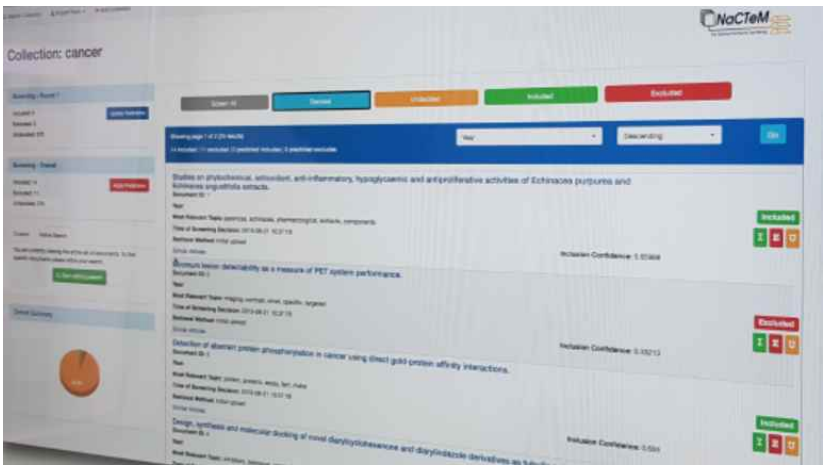
* 좋은 체계는 user community와 academic community 둘 다 가지고 있는 것임. 그래야 confidence 줄 수 있음.

* NaCTeM을 잘 홍보하는 루트는 많음. 센터에 속하는 모든 연구자 및 직원들이 다양한 방법으로 홍보를 함
- 가장 좋은 방법은 학회 발표, paper임. network 도 매우 중요함. 많은 학회에서 invitation talk 을 함.
- committiee나 자문위원으로 활동하며 센터를 알리기도 함.
- 유럽에서는 textmining이 왜 중요한지 알고 있음
- 트위터 등을 통해서도 홍보

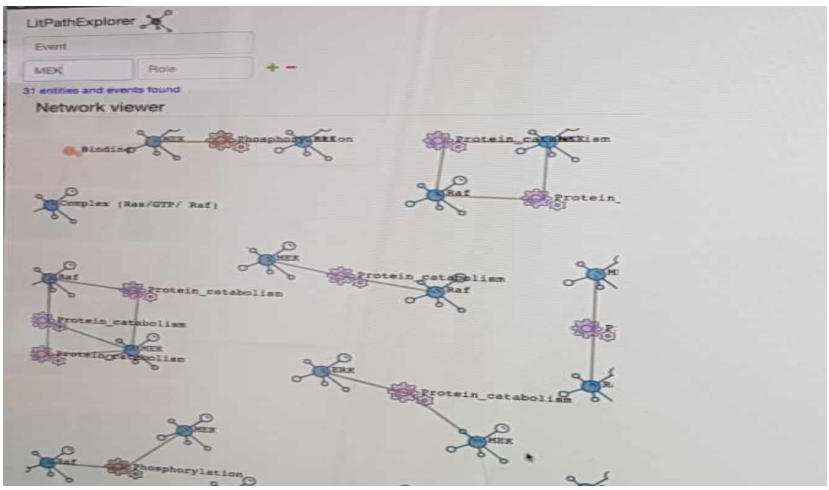
* center에서 많은 tool 들 개발하고 있음

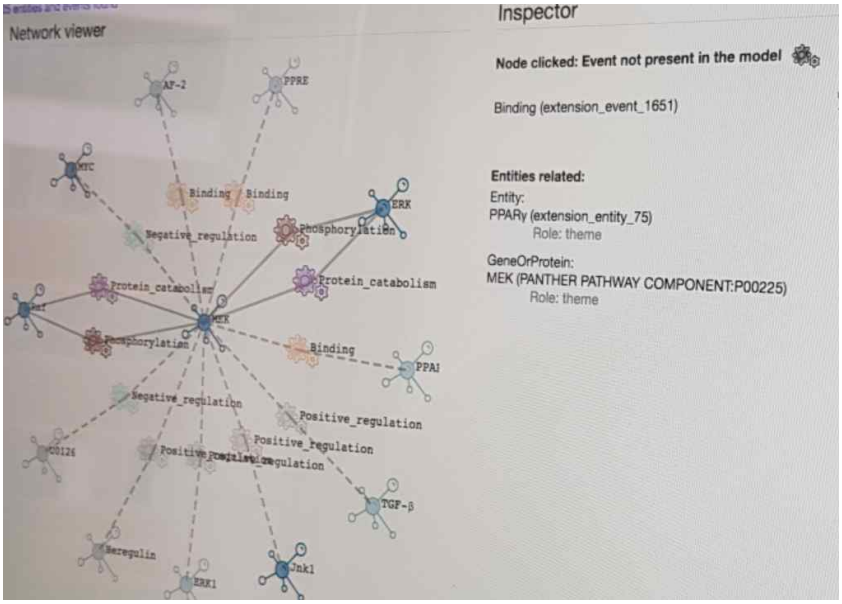
* Thalia 프로젝트
  (Talia, Text mining for Highlighting, Aggregate and Linking Information in articles)는 Pubmed에 색인된 biomedical abstract에서 발생하는 개념을 인식할 수 있는 검색 엔진으로 chemicals, diseases, drugs, genes, metabolites, proteins, species and anatomical entities 등 8가지 유형으로 나눔
  caffeine치면 관련 결과 나오는데 meta data가 잘되어 있음. 관심 있는 것과 엮어야 함

* supporting evidence based public health 관련하여
- overall summarize선택하고 나면 updata predictions나옴. inclusion confidence 0.659 이런식으로 값이 나오는데, user 가 사용하면 할수록 inclusion confidence 높아짐
- active learning 사용됨

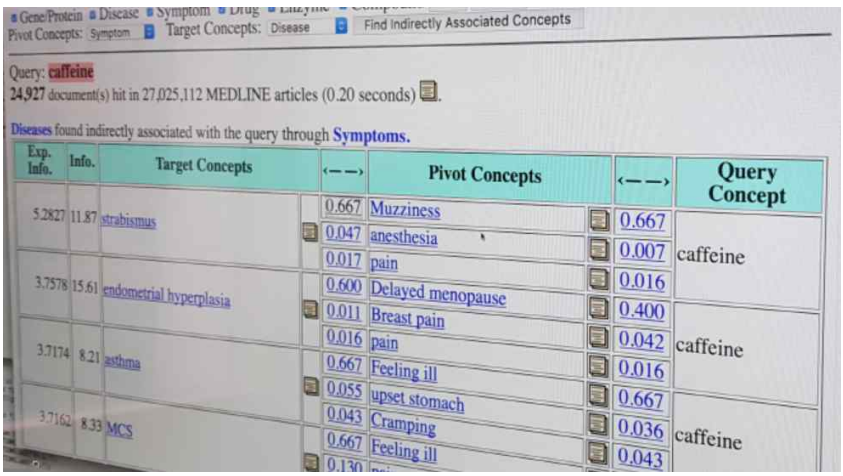


* LitPathExplorer
- network viewer 가 보여지는데 다른 툴도 마찬가지지만 모델이 만들어지면 로봇 과학자가 이런 작업들을 함. 로봇 과학자도 작업에 필요





* FACTA+
  caffeine 쿼리를 검색하면 target concept- pivot concept- Query concept의 결과가 나옴.



Exp. Info.	Info.	Target Concepts	‹—›	Pivot Concepts	‹—›	Query Concept
5.2827	11.87	strabismus	0.667	Muzziness	0.667	caffeine
			0.047	anesthesia	0.007	
			0.017	pain	0.016	
3.7578	15.61	endometrial hyperplasia	0.600	Delayed menopause	0.400	caffeine
			0.011	Breast pain	0.042	
3.7174	8.21	asthma	0.016	pain	0.016	
			0.667	Feeling ill	0.667	caffeine
3.7162	8.33	MCS	0.055	upset stomach	0.036	
			0.043	Cramping	0.043	
			0.667	Feeling ill		

* 향후 cowork 할 수 있는 방법 찾아보기로 함