

2024 디지털인문학 겨울학교 강의소개, 목차

신청링크 : <https://bit.ly/2024DHwinterschool>

2024 디지털인문학 겨울학교

2024년 1월 15일(월) ~ 19일(금)
고려대학교 문과대학(월~금), 국립중앙도서관(목)

이번 겨울학교에서는 디지털전환 시대 인문학의 새로운 학문분과로 떠오르고 있는 디지털인문학에 대한 최신의 지식과 기술을 습득할 수 있는 강좌 라인업을 준비하였습니다. 디지털인문학을 처음 접하는 인문학도들을 위한 기초강좌부터 특정 인문콘텐츠의 디지털 분석 방법을 다루는 심화과정까지, 연구자들의 관심과 수요를 반영한 다양한 강좌들이 개설됩니다. 강사진도 인문학 연구자들뿐 아니라 기업 현장 및 공공부문의 전문가들을 섭외함으로써, 디지털인문학이 학계 안팎에서 갖는 활용 가치와 미래 발전의 잠재력을 실감할 수 있게 했습니다. 수강생들은 기존의 방법론을 디지털 분석으로 보완하고, 나아가선 각자의 연구분야에 대한 새로운 통찰력과 영감을 디지털인문학에서 얻을 수 있을 것입니다.

요일/시간	09:00~10:15	10:30~11:45	13:30~14:45	15:00~16:15
월	등록 및 OT	<기초강연> 인공지능과 인간사회 / 장정선(고려대)	tomotopy를 활용한 토픽 모델링 / 이민철(카카오브레인) Gephi를 활용한 네트워크 시각화 / 김은재(고려대)	
화	DH를 위한 Python 입문 / 김바로(한국학중앙연구원) ----- 디지털 문화유산 아카이브 구축 실습 / 이종욱(한국전통문화대학교)			
수	문학 텍스트 기반 네트워크 분석 / 김희진(경북대) ----- 입코딩에서 노코딩으로 DH 연구하기 / 김광림(NAVER Cloud)			
목	한국 현대문학 DH 논문 A to Z / 전성규(성균관대) ----- 인문학의 디지털 자산화-OCR 핸즈온 실습 / 전영훈(업스튜디오) / 김수정(국립중앙도서관)			
금	그래프, 네트워크, 인문학 / 류인태 (고려대) ----- 근대 한국어 서브워드와 시계열 워드 임베딩 / 김병준(KAIST) / 원유권(바이칼AI)			

* 점심시간(11:45~13:30), 기초 강연을 제외한 수업은 동시 진행하며, 사전에 원하는 수업을 선택해 수강

** 목요일은 국립중앙도서관과 고려대 동시 진행

주최: 고려대학교 인문사회 디지털융합인재양성사업단(HUSS), 국립중앙도서관, 서울대학교 디지털인문학센터, 성균관대학교 국어국문학과 BK21 교육연구단, KAIST 디지털인문사회과학센터, 한국디지털인문학학회

주관: 고려대학교 인문사회 디지털융합인재양성사업단(HUSS)

후원: 교육부, 한국연구재단



<http://bit.ly/2024DHwinterschool>

1. (장정선) 인공지능과 인간사회

a. 강의소개

최근 OpenAI의 GPT4V, 구글의 BARD와 같은 생성형 인공지능 기술이 빠르게 발전하며, 다양한 산업 분야에서 활용되며 새로운 가치를 만들어 가고 있습니다. 특히 검색, 금융, 미디어, 미국 변호사 시험 등 특정 영역에서 인공지능이 사람보다 높은 수준의 결과를 보이면서 사회 전반에 새로운 변화를 만들고 있습니다.

인공지능 기술의 주요 개념과 수준에 대해 살펴보고, 현재 인공지능 기술의 한계에 대해 설명하며, 인공지능 시대를 맞이하는 우리의 삶에서 중요한 가치와 방향성에 대해 이야기하고자 합니다.

b. 목차

- 인공지능 기술의 주요 개념과 발전
- 다양한 영역에서 인공지능 기술의 활용과 한계
- 인공지능 시대에서 생각해봐야할 가치와 방향성

2. (이민철) tomotopy를 활용한 토픽 모델링

a. 강의소개

본 강의는 토픽 모델링의 기본적인 개념에 대해 소개하고 Python용 토픽 모델링 패키지인 tomotopy를 사용하여 실제로 토픽 모델링을 실습해보는 데에 그 목적이 있다. 본 강의를 통해 수강생들은 자신이 풀고자 하는 문제에 알맞는 토픽 모델을 선택하는 방법과 이를 위해 데이터를 전처리하는 방법, 토픽 모델링의 결과를 시각화하는 방법 등을 배울 수 있을 것이다.

b. 목차

- 토픽 모델링의 기본 개념
- LDA를 비롯한 다양한 토픽 모델에 대한 이해
- Python을 이용한 데이터 전처리
- tomotopy를 이용한 토픽 모델링 실습
- 토픽 모델링 결과 해석과 시각화

c. 참고

- <https://github.com/bab2min/tomotopy>
- <https://bab2min.github.io/tomotopy/v0.12.3/kr/>

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Lee, M., & Song, M. (2020). Incorporating citation impact into analysis of research trends. *Scientometrics*, 124(2), 1191-1224.

3. (김은재) Gephi를 활용한 네트워크 시각화

a. 강의소개

이 강의에서는 연결망의 기본적인 개념을 소개하고 **Gephi**를 통해 시각화하는 실습을 진행합니다. 전통적인 연결망 설문 자료 뿐만 아니라 텍스트 자료 등을 통해 재구성된 다양한 연결망의 예시를 살펴봅니다. 연결망 시각화 도구인 **Gephi**를 소개하고 실제 데이터를 활용하여 연결망 시각화 후 시각화에 대한 간단한 분석 및 해석을 시행합니다.

b. 목차

1. 연결망의 기본 개념
2. 다양한 데이터를 활용한 연결망 예시 살펴보기
3. **Gephi** 프로그램 소개 및 설치
4. 실제 데이터를 활용하여 **Gephi** 실습하기
5. 시각화한 연결망에 대한 분석 및 해석하기

4. (김바로) DH를 위한 Python 입문

a. 강의소개

본 강의는 **Python**이라는 이름만 들어본 인문학 연구자를 위하여, 인문지식 데이터의 수집, 정리 및 연구 주제에 입각한 데이터 분석, 그리고 연구 결과의 표현 등의 인문지식 처리의 제작 과정에서 데이터 처리를 효과적으로 수행할 수 있게 하는 **Python** 프로그래밍의 핵심 개념과 실전 활용 방법을 습득할 수 있도록 한다.

b. 목차

1. **Python**(Colab)을 왜 배워야 하는가?
2. **Colab** 기초 : **Colab** 파일 만들기, **Colab** 기본 기능, **Colab** 공유하기
3. **Python** 기초 : 변수, 조건, 반복, 함수, 주석
4. **Python** 심화 : 인문학 연구를 위한 데이터 입력과 출력
(점심식사)
5. **Python** 응용 : 인문학 연구를 위한 라이브러리의 개념과 활용, 예러 대처법

6. Python으로 논문 쓰기 : 방법론 작성법, 인용 방법, 데이터 및 프로세스 공개(공유) 방법

c. 참고:

1. 인문 데이터 분석 - 디지털인문학 시리즈: <https://wikidocs.net/book/8451>

2. 점프 투 파이썬 : <https://wikidocs.net/book/1>

* 관련:

데이터 입력 방법은? : 구글드라이브? 마운트? github 불러오기?

사용 라이브러리는? :

(이민철) tomotopy를 활용한 토픽 모델링: python, tomotopy, kiwi or n-gram or nltk

(김희진) 문학 텍스트 기반 네트워크 분석: python,

(김병준) 근대 한국어 서브워드와 시계열 워드 임베딩: python, kiwi, kiwi subword, 다이나믹 word2vec(gensim)?

(원유권) 한국어 연구를 위한 음성 분석/인식, 바른 형태소 분석기: python,

5. (이종욱) 디지털 문화유산 아카이브 구축 실습

a. 강의소개

디지털 문화유산은 문화유산의 속성과 그것이 가지는 가치에 대한 접근성을 높이는 것을 목적으로 한다. 이 수업은 문화유산 데이터의 종류와 특성을 분석하고 이를 저장, 관리하기 위한 아카이브의 현황과 한계점을 논의한다. 디지털 기록화를 위한 2D/3D 형상획득과 메타데이터 설계 및 입력을 통해 데이터를 생성하고 디지털 헤리티지간의 의미적 연관관계를 **protege**를 통해 설계한다. 의미적 관계가 설정된 디지털 헤리티지의 활용을 위해 **HCI(Human Computer Interaction)** 관점에서 사용자를 분석하고 사용자 경험 및 인터페이스를 설계한다.

b. 목차

1. 문화유산 데이터의 종류와 특성
2. 디지털 문화유산 아카이브 현황과 한계
3. 디지털 기록화 : 데이터수집 및 메타데이터 디자인
4. 디지털 시각화 : **protege** 실습하기
5. 디지털 경험 디자인 : 가상박물관 전시기획
6. 프로젝트 발표

6. (김희진) 문학 텍스트 기반 네트워크 및 군집 분석

- a. 강의소개: 이 강의는 디지털인문학의 다양한 방법론 중 사회 관계망 분석 도구와 텍스트 군집 분석 도구를 활용해 문학 텍스트를 분석하는 방법을 살펴본다. 사회 관계망 분석 도구를 문학 텍스트에 맞게 등장인물 관계망 분석 도구로 발전시켜, 드라마 텍스트와 소설 텍스트에 각각 적용하는 방법을 실습한다. 또한 딥러닝의 텍스트 군집 분석 도구를 활용해 문학 텍스트의 장르, 문체, 연도, 등을 분석해 본다. **Google Colab**을 사용해 컴퓨터에 직접 **Python** 설치 없이 실습한다. 코딩에 대한 사전 지식 없이 수강 가능하다.
- b. 목차
 - 1. 네트워크 분석 개관
 - 2. 드라마 텍스트 네트워크 분석
 - 3. 소설 텍스트 네트워크 분석
 - 4. 네트워크 분석 실습
 - 5. 텍스트 군집 분석 개관
 - 6. 문학 텍스트 군집 분석
 - 7. 텍스트 군집 분석 실습

7. (김광림) 입코딩에서 노코딩으로 DH 연구하기

- a. 강의소개 : 디지털인문학 연구를 가로막는 가장 큰 장벽 중 하나는 코딩을 비롯한 디지털 도구 사용의 어려움입니다. 최근 **ChatGPT**를 비롯한 다양한 **AI** 언어모델들이 등장하며 놀라운 코드 작성 능력을 보여주고 있는데요. 이러한 **AI** 언어모델을 이용하여, 코딩을 모르는 인문학 연구자가 **AI**와 대화하면서 디지털인문학 연구를 수행하는 방법을 소개합니다. 본 강의는 수강자와 함께 의견을 주고받으며, 노코딩 분석 과정 전체를 라이브 세션으로 시연하면서 진행됩니다. 강의를 끝나면, 막연해 보였던 디지털인문학 연구가 보다 친숙하고 만만하게 느껴졌으면 합니다.
- b. 목차
 - i. 거대언어모형(LLM)과 **ChatGPT** 소개
 - ii. 코딩 없이 연구 데이터를 정량적으로 살펴보기
 - iii. 코딩 없이 다양한 연구 데이터 전처리
 - iv. 코딩 없이 관계성(네트워크), 공간정보(GIS) 분석하기

8. (전영훈) 인문학의 디지털 자산화(OCR 핸즈온 실습)

- a. 강의소개 : 인문학 자료를 디지털 자산으로 만들기 위한 **OCR** 기술에 대해 소개하며, 쉽게 사용할 수 있는 **OCR** 제품의 사용법을 가르쳐 드리는 핸즈온 실습형 강의입니다. 본 본 강의 수강을 통해 인문학 데이터 구축에 활용할 수 있는 방법을 익히고, 실행할 수 있는 능력을 갖출 수 있습니다.
- b. 목차
 - i. 디지털 자산화와 데이터 구축
 - ii. OCR이란?
 - iii. AskUp을 활용한 OCR(국한문 혼용, 세로쓰기)
 - iv. DocumentAI를 활용한 OCR(한국어, 영어, 기타 외국어 가능 여부 확인 필요)
 - 1. **사용 라이브러리 공유, 혹은 코드** 원본 김바로 강사님께 전달 필요
 - 2. 드라이브 마운트
 - 3. colab 작업 환경 사용 필요
 - 4. **output**도 드라이브로 이동
 - v. 질 좋은 데이터의 중요성

9. (김수정/국립중앙도서관)

a. 강의 소개

국립중앙도서관은 국가지식문화유산의 보고로써 디지털 시대에 국가 지식 정보자원의 활용을 위한 다양한 방안을 추진하고 있다. 정보화 시대에 자원의 디지털화를 우선으로 추진하였다면, 디지털 시대에서는 큐레이션 중심의 지식문화유산의 활용을 위한 데이터화가 필요하다. 이번 강의에서는 도서관의 대표적인 미래형 큐레이션 서비스인 코리안 메모리를 소개하고, 큐레이션을 위하여 기본적으로 전제되어야 하는 국가지식문화유산의 디지털화와 데이터 구축에 대하여 설명한다. 특히, 디지털 신기술을 도입한 디지털화 방안 중에서 AI OCR을 활용한 텍스트 데이터베이스의 구축 시연을 볼 수 있다. 마지막으로 디지털인문학의 일환으로 국립중앙도서관의 자원을 활용하는 연구자들을 우선적으로 발굴하고 지원하는 연구정보서비스를 소개한다.

b. 목차

1. 국립중앙도서관 견학
2. 코리안 메모리 소개 및 컬렉션 시연
3. 국가지식문화유산 디지털화 이해와 실습
4. AI OCR을 활용한 원문 텍스트 구축 시연
5. 연구정보서비스

10. (전성규) 한국현대문학 DH 논문 AtoZ

a. 강의소개

한국 근대 잡지의 파라텍스트들(paratext, 텍스트 바깥의 텍스트들, 판권장, 회원정보, 주소정보 등 부수적인 정보라고 간주되었던 정보들)을 데이터로 생산하는 방법에 대해 고민해 본다. 근대 잡지를 대상으로 이루어진 디지털인문학 연구는 어휘 분석 기술을 기반으로 한 주요 기사 분석에 치우쳐 있다. 그간 한국문학 연구에서 잡지 생산의 제도적 측면에 대한 연구 성과가 활발히 제출된 바 이것과 상호작용할 수 있는 디지털인문학 연구 또한 요청되는 상황이다. 이번 강의에서는 파라텍스트들 중 주소 정보를 초점화하여 이를 간단히 시각화할 수 있는 툴을 사용해 지리정보를 생산하고 이를 기반으로 논문을 작성한 경험을 공유하고자 한다.

b. 목차

1. 한국 근대 잡지의 전반적인 체제와 이를 대상으로 데이터를 생산한 연구성과에 대한 검토(text 분석 + paratext 분석의 사례)
2. 잡지 <<어린이>>//<<친목회회보>>와 <<태극학보>>를 비롯 재일본 유학생 잡지 주소정보 수집, 데이터베이스 구축, 시각자료 생산과정, 분석에 대한 전반적 설명

3. 실습

4. + 성균관대 국문과 강의(학부)에서 생산된 데이터 소개: 2013년부터 활발히 제출된 쿼어문학을 대상으로 쿼어 지리, 공간 정보에 초점을 맞춰 데이터 생산 및 시각화

참고:

전성규, <근대 잡지의 계량적 연구방법에 대한 논의>, <<민족문학사연구>>, 2023

전성규, 박지은, <지도작성 기술 기반 <<어린이>> 독자 커뮤니티 변화연구>,

<<한국현대문학연구>>, 2023

전성규, 허예슬, 이여진, 최장락, <근대계몽기 지식인 단체 네트워크 분석>, <<상허학보>>, 2022 등

11. (김병준) 근대 한국어 서브워드와 시계열 워드 임베딩

a. 강의소개

이 강의는 크게 두 가지를 다룬다. 첫째, 현대 한국어 형태소 분석기로는 분절화(토큰나이징) 할 수 없는 근대 한국어 텍스트(국한문 혼용체 등)를 서브워드(subword)와 한국어 형태소 분석기를 혼용해서 분절화하는 방법을 배운다. 이때 수강생들이 각자 형태소 분석하고 싶은 근대 한국어 텍스트 자료를 가져와 직접 적용해 볼 것이다. 둘째, 워드 임베딩(word embedding)의 개념을 이해하고, 개념사 연구에 유용한 시계열 워드 임베딩(dynamic word embedding)을 실습해본다. 시간의 흐름에 따른 어휘 문맥의 변화를 살펴보는 실습이 될 것이다.

b. 목차

i. 근대 한국어 서브워드(1.5H)

1. 토큰나이징과 자연어 처리
2. 현대 한국어 형태소 분석기의 한계
3. 서브워드(subword)의 개념
4. kiwi를 활용한 서브워드 토큰나이저 만들기

ii. 시계열 워드 임베딩 (1.5H)

1. 워드 임베딩의 세 가지 철학
2. 개념사 연구와 시계열 워드 임베딩
3. 시계열 워드 임베딩 실습

a. 1920~40년 조선/동아일보 기사

c. 참고문헌

<https://github.com/bab2min/kiwipiepy>

<https://github.com/ByungjunKim/ModernKoreanSubword>

<https://github.com/ByungjunKim/NationStateChosun>

김병준 and 전봉관. (2023). 민족, 국민, 국가 — 시계열 워드 임베딩을 활용한 조선일보 기사의 민족 담론 의미 변동 추적(1920~40). 현대소설연구, 90, 5-38. <https://byungjunkim.com/publication/2022-06-31-paper-title-number-14>

12. (원유권) 한국어 연구를 위한 음성 분석/인식, 바른 형태소 분석기

a. 강의소개

음성인식 과정을 소개하고 한국어 음성 인식 분야에서 활용할만한 다양한 도구(whisper, pyannote, praat-parselmouth)를 실습하여 한국어 음성 분석 연구를 위한 기초를 다집니다. 바른 형태소 분석기 사용을 익혀 텍스트 분석에 사용합니다. 텍스트를 형태소 분석하여 word2vec 워드 임베딩, t-SNE 등 클러스터링 시각화를 사용해봅니다.

b. 목차

i. 한국어 음성인식 및 분석

1. 음성인식 소개

2. 음성분석 도구 소개

a. whisper

b. pyannote

c. praat-parselmouth

3. 실습

a. STT

b. 화자분리

c. 음성 feature 추출(모음 포먼트, 발화속도, 휴지 구간 측정)

ii. 바른 형태소 분석기 소개 및 활용

1. 바른 형태소 분석기란?

a. 바른의 분석 원리 및 특징

2. 바른 설치 및 api 사용하기

a. 바른 라이브러리(bareunpy)

b. 사용자사전 기능

c. 토크나이징 기능

3. 활용

c. 참고문헌:

i. 음성

<https://github.com/openai/whisper>

<https://github.com/pyannote/pyannote-audio>

<https://parselmouth.readthedocs.io/en/stable/>

<https://docs.bareun.ai/intro/>

[원유권\(2022\), 발화 속도와 말차례 교체 빈도에 따른 운율 단위 변화에 관한 연구, 말소리와 음성과학 14-2, 한국음성학회, 29-38p.](#)

[나종환,이보원\(2023\), 발화 속도와 휴지 구간 길이를 사용한 방언 분류, 말소리와 음성과학 15-2, 한국음성학회, 43-51p.](#)

ii. 바른 관련

박진호(2019), 딥러닝 기반 자연어 처리에서 도메인 지식의 역할, 인공지능인문학연구 4.

[임한열\(2023\), 조사 분포의 계량적 측정을 통한 정보 탐색 연구 — 구어성 텍스트와 문어성 텍스트의 특성 비교 —, 어문론집 95](#)

13. (류인태) 그래프, 네트워크, 인문학

a. 강의소개

네트워크 그래프(network graph)의 개념과 이해에 관한 전달을 기본 바탕으로 네트워크 그래프 형식을 매개로 디지털 인문학에서 다루어지는 데이터 분석, 시각화, 큐레이션 방법론의 구체적 사례 및 그 학술적 의미와 맥락을 전반적으로 정리해보고자 합니다. 특히, 디지털 인문학 영역에서 네트워크 그래프 형식을 활용한 대표적 데이터 처리 방법론이라 할 수 있는 네트워크 분석(network analysis)과 지식그래프(knowledge graph)의 차이가 무엇인지를 조망하는 가운데, 그래프데이터베이스(Neo4j)를 다루는 간단한 맛보기 실습 또한 진행할 예정입니다.

b. 목차

1. 네트워크(network)와 그래프(graph)란 무엇인가? (50M)
2. 디지털 인문학 연구에서 네트워크, 그래프의 활용 사례와 의미 (50M)
3. 네트워크 분석(network analysis)과 지식그래프(knowledge graph) (50M)
(점심 시간 12:00-13:30)
4. 온톨로지(ontology)와 스몰데이터(small data) (50M)
5. 그래프데이터베이스(Neo4j)의 이해와 실습: 기초 (50M)
6. 그래프데이터베이스(Neo4j)의 이해와 실습: 활용 (50M)

※상기 목차는 추후 약간의 수정이 있을 수 있습니다.

c. 참고문헌

- [네트워크 분석: 소셜 미디어에서 신경망까지](#)
- [네트워크 분석 방법론](#)
- [네트워크 분석방법의 활용과 한계](#)
- [그래프 알고리즘](#)
- [그래프 머신러닝](#)
- [온톨로지 과학자를 위한 지식그래프](#)
- [Neo4j로 시작하는 그래프 데이터베이스 2/e](#)