강의계획서

(2022 학년도 2학기)

1. 강좌 및 담당교수

작성일 : 2022.07.28

교과목명		텍스트마이닝	학수번호	11023082	수강반	001
외국어강의구분			강의시간	화1,2[407-0202],목3[	강의실	[407-0202]
이수구분		전공선택	강좌구분		코티칭여부	
수강대상			학점구성	학점 : 3, 이론 및 실습 : 3, 설계 :		
담당교수	소속	항공우주및소프트웨어공학부	수업방법	대면수업		
	성명	이선아	연구실			
	전화번호	0557721377	E-mail	saleese@gnu.ac.kr		

2. 강의내용(목적)

항목	전공역량 세부목표	
	비율	연관성

본 과목에서는 대용량의 텍스트를 분석하여 유의미한 결론을 도출하는 텍스트 마이닝 기법에 대해 강의한다. 또한 자연어 처리 분야의 최근 기술 발전으로의 Transformer를 활용하는 프로젝트를 진행한다.

3. 교재 및 참고서적

구분	저자	도서명	출판사	비고
주교재	박상언, 강주영, 정석찬	파이썬텍스트마이닝완벽가이드	위키북스	

4. 과제

과제	과제명	참고사항

5. 평가방법

평가방법	출석	중간고사	기말고사	수시고사	과제물	기타	계
배점비율	10	35	45	0	10	0	100

6. 장애학생을 위한 지원사항

7. 주별 강의계획

주차	강의내용	강의방법	활용기자재	비고(상세수업방법)
1주차	01주: 텍스트 마이닝 기초 01-01: 강의소개와 기본 실습 환경 강의 소개 기본 실습 환경 소개 실습 환경 설치 및 사용 확인 01-02: 파이썬 기본 소개 파이썬 기초 파이토치 기초 파이썬과 파이토치 실습 01-03: 텍스트 마이닝 기초 텍스트 마이닝 정의 텍스트 마이닝의 주요 적용 분야 텍스트 마이닝을 위한 파이썬 라이브러리 소개			
2주차	02주: 텍스트 전처리 02-01: 토큰화 문장 토큰화 단어 토큰화 정규 표현식 활용 토큰화 02-02: 정규화 노이즈와 불용어 제거 어간 추출 표제어 추출 02-03: 품사 태깅 품사의 이해 nltk를 활용한 품사 태깅 한글 형태소 분석과 품사 태깅			

3주차	03주: 카운트 기반의 문서 표현 03-01: 카운트 기반의 문서 표현 카운트 기반의 문서 표현 - BOW 기반의 카운트 벡터 생성 - 사이킷런으로 카운트 벡터 생성 03-02: TF-IDF 활용 TF-IDF의 개념 BOW 기반으로 TF-IDF 사용 사이킷런으로 TF-IDF 사용 03-02: 한국어 텍스트의 처리 한국어 텍스트 처리 라이브러리 소개 한국어 텍스트의 카운트 벡터 변환 한국어 텍스트의 TF-IDF 활용			
4주차	04주: BOW 기반의 문서 분류 04-01: 뉴스그룹 데이터 준비 및 특성 추출 뉴스그룹 데이터 소개 데이터 셋 확인과 분리 카운트 기반 특성 추출 04-02: 나이브 베이즈 분류기를 이용한 문서 분류 나이브 베이스 분류기 소개 나이브 베이스 분류기의 분류 과정 나이브 베이스 분류 실습 04-03: 로지스틱 회귀분석을 이용한 문서 분류 로지스틱 회귀 분석 소개 로지스틱 회귀 분석을 이용한 분류 과정 로지스틱 분류 실습			
5주차	05주: BOW 기반 및 순서 기반의 문서 분류 05-01: 결정 트리를 이용한 문서 분류 - 결정트리 소개 결정트리를 이용한 분류 과정 결정트리 분류 실습 05-02: n-gram을 이용한 문서 분류 - n-gram 소개 - n-gram을 이용한 분류 과정 - n-gram 실습 05-03: 성능을 높이는 방법 토크나이저의 향상 특성의 수 증가			
6주차	06주: 차원 축소 06-01: 차원의 저주와 차원 축소의 이유 차원의 저주 차원 축소의 이유 06-02: PCA를 이용한 차원 축소 PCA 소개 PCA를 이용한 차원 축소 06-03: LSA를 이용한 차원 축소 LSA 소개 LSA를 이용한 차원 축소 06-04: tSNE를 이용한 차원 축소 tSNE 소개 tSNE를 이용한 차원 축소			
7주차	중간고사			
8주차	07주: 토픽 모델링으로 주제 찾기 07-01: 토픽 모델링과 LDA의 이해 토픽 모델링이란? LDA의 이해 모형의 평가와 적절한 토픽 수의 결정 07-02: 사이킷런을 이용한 토픽 모델링 데이터 준비 토픽 모델링 실행 최적의 토픽 수 선택 07-03: Gensim을 이용한 토픽 모델링 Gensim 사용법 Gensim 시각화 혼란도와 토픽 응직도를 이용한 최적값 선택			
9주차	08주: 감성 분석 08-01: 감성분석의 이해 어휘 기반의 감성분석 머신러닝 기반의 감성 분석 08-02: NLTK 영화 리뷰 데이터 감성 분석 1 영화 리뷰 데이터 준비 - TextBlob을 이용한 감성 분석 - AFINN을 이용한 감성 분석 08-02: NLTK 영화 리뷰 데이터 감성 분석 2 - Vader를 이용한 감성 분석 한글 감성 사전 NLTK 영화 리뷰에 대한 머신러닝 기반 감성 분석			

10주차	09주: 워드 임베딩 기법 이해 09-01: Word2Vec 이해 Word2Vec 소개 Word2Vec 원리 Word2Vec 활용 09-02: ELMo 이해 - ELMo 소개 - ELMo 원리 - ELMo 활용 09-03: FastText 이해 - FastText 소개 - FastText 원리 - FastText 활용			
11주차	10주: RNN 기반 문서 분류 10-01: RNN의 문서 분류 이해 RNN의 이해 RNN이 문서 분류에 적합한 이유 RNN의 문서 분류 적용 방안 10-02: RNN을 이용한 NLTK 영화 리뷰 감성 분석 워드 임베딩을 위한 데이터 준비 일반적인 신경망 모형을 이용한 분류 문서의 순서정보를 활용하는 RNN 분류 10-03: LSTM, Bi-LSTM, GRU를 이용한 성능 개선 LSTM을 이용한 성능 개선 - Bi-LSTM을 이용한 성능 개선 GRU를 이용한 성능 개선			
12주차	11주: CNN 기반 문서 분류 11-01: CNN의 작동 원리 - CNN의 이해 - CNN 원리 - CNN 활용 11-02: CNN의 문서 분류 이해 CNN을 이용한 문서 분류의 원리 CNN의 문서 분류 적용 방안 11-03: CNN을 이용한 NLTK 영화 리뷰 분류 CNN을 이용한 NLTK 영화 리뷰 분류 CNN과 RNN의 성능 비교 CNN과 RNN의 차이점 비교			
13주차	12주: 어텐션과 트랜스포머 12-01: Sequence 2 Sequence 딥러닝 기법 - Sequence2Sequence의 이해 - Sequence2Sequence의 원리 - Sequence2Sequence의 활용 12-02: 어텐션을 이용한 성능 향상 Attention의 이해 Attention의 원리 Attention의 활용 12-03: 셀프 어텐션과 트랜스포머 트랜스포머의 이해 트랜스포머의 원리 인코더와 디코더의 작동 원리			
14주차	13주: BERT의 이해와 간단한 활용 13-01: 언어모델의 중요성 및 BERT 구조 언어모델의 중요성 BERT의 소개 BERT의 구조 13-02: 언어모델을 이용한 사전학습과 미세조정학습 언어모델을 이용한 사전학습 언어모델에서의 미세조정학습 언어모델의 활용 13-03: 사전학습된 BERT 모형의 직접 사용방법 사전학습된 BERT 모형 다운받기 사전학습된 BERT 모형의 언어 처리 확인 사전학습된 BERT 모형의 사용			
16주차	기말고사			