교육
과정
소개서.

R을 활용한 텍스트마이닝 CAMP

Fast campus

코스요약

코스명	R을 활용한 텍스트마이닝 CAMP
기간	2020. 9. 15 - 2020. 11. 10 (총 8주, 9월 29일 추석연휴로 인한 휴강)
일정	매주 화요일 19:30 - 22:30 (주 1회 3시간, 총 24시간)
장소	패스트캠퍼스 강남강의장
준비물	개인 노트북, 필기구
담당자	02-568-9886 / help-ds@fastcampus.co.kr
수강료	1,500,000
상세페이지 url	fastcampus.co.kr/data_camp_rtm/

코스목표

기초적인 텍스트 분석인 워드클라우드 및 클러스터링, 문서 분류 등의 고급 텍스트 분석이 가능하며, 문서를 일일이 보지 않아도, 비슷한 주제를 자동으로 군집화 하여 이를 시각적으로 나타낼 수 있습니다. 수준 높은 보고서 작성 등 텍스트마이닝을 사용하여 실무에서 다양한 활용이 가능해집니다.

코스정보

R 핵심 문법부터 텍스트 마이닝 이론, 실습, 비즈니스 활용법까지! 실무에서 활용범위가 높은 진짜 분석이 필요할 때, 현직 실무자가 말하는 진짜 써먹을 수 있는 분석 스킬을 알려드립니다.

코스특징

텍스트 분석 + 시각화 + 보고서 ALL-IN-ONE.

본 캠프에서는 키워드 연관 분석부터, 군집화, 분류, 감성 분석 등 다양한 텍스트 분석을 실제 데이터를 활용해 실습합니다. 뿐만 아니라 간단한 워드 클라우드가 아닌 Shiny를 활용한 인터렉티브한 결과물 시각화 등 보다 심화된 시각화 방법을 배웁니다. 이를 기반으로 수강생의 목적에 맞는 전략 보고서를 작성하는 것까지 가르쳐드리는 All-in-ONE 강의입니다.

독학용 스크립트 및 코드, 형태소 분석기 제공

텍스트 분석은 수업 이외에도 꾸준한 예습/복습이 중요합니다. 초보자들이 코드를 한 줄 한 줄 꼼꼼히 이해하고, 복습에 용이하도록 매 실습에 맞춘 R 스크립트를 별도로 제공합니다. 또한, 공개된 형태소 분석기를 사용하는 것이 아니라, 강사님께서 직접 개발하신 고성능 형태소 분석기까지 제공하여 드립니다. 이를 통해 수업시간에 배운 내용을 바로 실무에서 활용하실 수 있습니다.

현업 전문가의 1:1 피드백 및 조교를 통한 케어

텍스트마이닝은 단순 분석보다 그 결과를 어떻게 해석하고, 전략을 도출하느냐가 핵심입니다. 2010년부터 금융업, 제조업, 서비스업 등의 분석 프로젝트는 물론 비즈니스와 연계한 VOC데이터나 소셜데이터의 활용까지 풍부한 실무 경험을 가진 강사님의 실전 노하우를 전합니다. 또한, 강의를 못 따라갈까봐 두려우신 초보자 분들을 위하여 이론과 실습을 보조하여 주실 조교님께서 강의에 상주해 도움을 드립니다.

커리큘럼

〈PART 1. 텍스트 분석을 위한 데이터 전처리〉

1회차 자연어 처리와 형태소 분석 이해

자연어 처리의 기본 개념을 이해하고, 한국어 처리방법에 대하여 알아봅니다.
텍스트 분석을 위한 R과 형태소 분석기 설치 및 사용법을 알려드립니다.

[이론]
- 자연어처리에 대한 기본 개념 이해
- 텍스트 분석 사례
- 검색엔진 작동원리를 통한 형태소 분석에 대한 이해
- 오픈소스 형태소 분석기에 설명

[실습]
- R/Rstudio 설치 및 사용법 설명
- 형태소분석기 설치

2회차 데이터 핸들링을 위한 기초 R 문법 학습

R이라는 분석툴과 친숙해져 봅시다.
R에서 데이터전처리 및 분석을 진행하기 위한 기본적인 함수 및 시각화 방법을 배웁니다.

[이론]
- R data handling (apply, dplyr 등)
- R 문자 data handling
- 데이터 시각화 방법 (ggplot)

[실습]
- R에서의 패키지 설치 및 불러오기
- 데이터 불러오기 및 저장하기
- 여러 개의 데이터를 결합하기
- 데이터 재구조화 하기
- 제어문 / 반복문 및 간단한 함수 만들기

3회차 R에서의 텍스트데이터 전처리

R에서 텍스트 데이터를 처리하기 위한 기본적인 함수를 익히고,
R 텍스트 마이닝 패키지 tm과 형태소분석 패키지 NLP4kec 사용법에 대해 알아봅니다.

[이론]
- NLP4kec package를 통한 형태소 분석
- 텍스트데이터 전처리
- DTM 생성 및 TF-IDF 이해
- tm, konlp package 기본 사용법

[실습]
- 텍스트마이닝을 위한 패키지 설치 및 활용
- 문자열 전처리(특수문자 제거, 특정 단어 삭제, 동의어 처리, 숫자 삭제 등)
- 텍스트 분석의 기본인 단어 빈도수 시각화 하기
- 워드클라우드 및 트리맵 그리기

커리큘럼

〈PART 2. 본격! R을 활용한 텍스트마이닝〉

4회차 문서내 핵심 키워드 추출 및 연관 키워드 분석

뉴스나 블로그 글을 일일이 읽지 않아도 텍스트데이터 내에서 중요한 키워드를 선별할 수 있고,
키워드간의 관계를 네트워크 맵으로 시각화하는 방법을 배웁니다.

[이론]
- 단어빈도와 문서빈도 관계에 대한 이해
- 연관 키워드 네트워크 분석
- R shiny와 ggplot을 이용한 시각화
- word2vec를 이용한 단어간 유사도 측정

[실습]
DATA : 가전제품에 대한 카페 댓글
- 연관 키워드 추출하기 (냉장고와 가장 연관도가 높은 단어는 무엇일까?)
- 단어간 상관관계 구하기 (냉장고와 에어컨은 얼마나 상관성이 있을까?)
- 전체 단어간의 관계 시각화하기
- 연관 키워드 네트워크 맵 그리기
- 특정 키워드만 선택하여 네트워크 맵 그리기

5회차 WORD EMBEDDING 및 문서 분류기법

Word embedding에 대해 알아보고, word embedding을 활용하여 단어간 유사도를 구하고, 문서를 분류하는 방법에 대해 배웁니다.

[이론]
- word embedding에 대한 이해
- 벡터 간 거리 및 유사도 측정 방법 이해 (consine, 유클리드)

[실습]
- word2vec을 이용한 단어간 cosine 유사도 측정
- cosine 유사도를 활용한 키워드 네트워크 분석
- word2vec을 이용한 문서 분류

6회차 비슷한 주제의 문서를 자동으로 군집화 해주는 토픽 clustering의 이해

비슷한 주제의 문서를 자동으로 군집화 해주는 토픽 클러스터링 기법에 대해 배웁니다.
토픽 클러스터링 기법은 사람이 직접 다 읽어볼 수 없는 대량의 문서에서
주제를 나타내는 핵심 키워드를 찾아 쉽고 빠르게 문의 내용 분류를 가능케합니다.

[이론]
- LDA기반의 토픽클러스터링 이해
- topicmodel 패키지를 활용한 토픽클러스터링 분석 및 시각화

[실습]
DATA :콜센터 데이터
- 문서를 읽지 않고도, 문서 내에서 주요한 토픽을 추출하여 보기
- 각 토픽을 대표하는 단어를 추출하기
- 클러스터링 결과 및 단어의 분포를 동적으로 시각화하기
- 토픽 클러스터링 시각화 결과 해석

커리큘럼

〈PART 2. 본격! R을 활용한 텍스트마이닝〉

7회차 머신러닝 기법을 통한 문서 Classification (감성분류)

문서를 특정 카테고리로 분류하기 위한 지도학습 머신러닝 알고리즘을 학습하고, 직접 감성분류 모델을 개발합니다.

[이론]
- 분류모델을 이해하기 위한 조건부 확률 학습
- 나이브베이즈 기법을 통한 문서 분류
- SVM 을 통한 문서분류

[실습]
DATA :쇼핑몰 댓글 데이터
- 나이브베이지안 모델링을 통한 스팸 문서 예측하기
- SVM 모델링을 통한 스팸 문서 예측하기
- 새로운 댓글이 들어왔을 때 분류 예측 결과 확인하기

8회차 분류 모델 평가 및 성능 높이기

R에서 만든 분류 모델의 성능을 평가해보고, 성능을 높일 수 있는 방법에 대하여 알아봅니다.

[이론]
- 분류 모델에 대한 성능 평가 방안
- 분류 모델 성능 향상을 위한 caret 패키지 사용법

[실습]
DATA :상품 구매 만족도 데이터
- 다수의 약한 학습기를 결합하여 하나의 강한 학습기를 만드는 앙상블 기법 학습하기
- 모델링을 쉽게 할 수 있게 해주는 caret 패키지를 사용하여 샘플링, 모델평가 과정을 자동화기

〈PART 3. 실제 텍스트마이닝을 활용한 마케팅 전략 수립 사례 소개〉

텍스트 마이닝 비즈니스 적용 사례 소개

앞에서 배운 텍스트 분석 기법을 활용하여 VOC, 소셜, 뉴스 등의 텍스트 데이터를 분석하여 실제 비즈니스에 적용한 사례에 대해 소개합니다.

[이론]
- 여러 업무분야에서 텍스트 분석을 적용한 비즈니스 사례 소개
- 텍스트 분석을 활용한 마케팅 전략 수립 사례 소개

강사소개

김남윤

現 K사 데이터 분석 업무 담당
前 아모레퍼시픽 데이터 분석 및 마케팅 퍼포먼스 담당
前 LG CNS 빅데이터 센터 근무 (텍스트마이닝 연구 및 분석컨설팅)
前 LIG System 근무 (분석솔루션 개발)
前 SAS Korea 근무

김남윤 강사님은,
현재 사내에서 데이터 분석 업무를 담당하고 있습니다. 2010년 부터 다년간 SAS, LG CNS, 아모레퍼시픽, 금융권에서 다양한 분석 실무 경험을 쌓았습니다. 특히 소셜 분석 서비스를 개발하면서 현업의 비즈니스와 연계하여 인사이트를 도출 할 수 있는 텍스트마이닝 활용 기술을 연구/개발 하였습니다.

수강환경

강남강의장

❖ 강의에 따라 강의장이 변경될 수 있습니다.