NCS(National Competency Standards)

파이썬과 R을 활용한 빅데이터 플랫폼 전문가 과정 수업계획안

2020년 02월 19일(수)

ITWILL

Contents

1. 수업 계획안
2. 포트폴리오 & 프로젝트

1. 수업 계획안

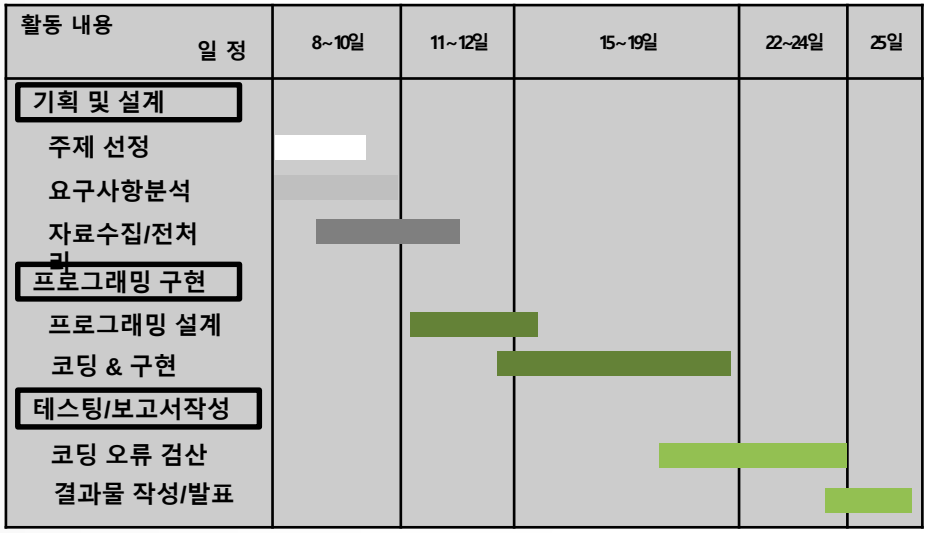
※ 교육 환경 및 여건에 따라 조정될 수 있습니다.

훈련교과	학습목표(내용)	시수
1. 데이터 분석을 위한 SQL	◦ 기본 SQL 작성하기 ◦ 고급 SQL 작성하기 ◦ 응용 SQL 작성하기	40H
2. R을 활용한 빅데이터 분석 및 시각화	◦ R 데이터 분석 환경 구축 ◦ 데이터 처리 및 가공 ◦ 데이터 분석 및 시각화 ◦ 기술 통계 및 통계분석(상관 및 회귀분석) ◦ 기계학습 알고리즘의 이해(지도학습과 비지도학습) ◦ **R을 활용한 세미 프로젝트**	180H (20H)
3. 파이썬을 활용한 머신러닝 기반 빅데이터 분석	◦ 파이썬 개발환경 구축 및 이해 ◦ 파이썬 라이브러리를 활용한 데이터 분석 ◦ 선형모델과 머신러닝의 이해	240H
4. 하둡을 활용한 빅데이터 저장 및 처리 환경구축	◦ 하둡을 이용한 빅데이터 플랫폼 개념 및 구축 ◦ 리눅스 운영체제에 대한 이해 및 설치 ◦ 하둡 Ecosystem(HIVE & SPARK)	120H
5. 머신러닝과 딥러닝 (인공신경망의 구축)	◦ 텐서플로 설치 및 기본코드 구현 ◦ 딥러닝 알고리즘과 신경망 구축 ◦ 인공지능 알고리즘 코드의 이해 ◦ 텐서플로 프로그래밍 ◦ CNN & RNN	120H
최종 프로젝트	◦ 빅데이터 개발 및 분석 프로젝트 진행 및 시연	100H

2. 포트폴리오 & 프로젝트

1. R 오픈소스 기반 통계분석
2. Python 기반 머신러닝 프로젝트
3. Tensorflow 기반 딥러닝 프로젝트

Example) 프로젝트 기간 : 2019년 07월 08일 ~ 07월 25일



R Portfolio

프로젝트 제목

서울 어디서 살아야 할까?

프로젝트 내용

SNS & Daum News 크롤링 기반

1인 가구 최적 환경 지역 선정

특.장점

- 데이터 크롤링(data crawling) 기법
  - ✓ SNS, Daum News 데이터 크롤링
- 지도 공간 시각화
  - ✓ 레이어 기법 적용 지도 공간 시각화

R Portfolio

프로젝트 제목

대장암 유무 예측

프로젝트 내용

공공데이터 7만 건을 이용한 대장암 유무 예측 분석과 중요변수

특.장점

- ➢ 변수 선택법 적용
  - ✓ Importance 이용
- ➢ 분류모델 적용
  - ✓ Knn, NB, DT, RF, SVM

알고리즘	정확도	F1 Score
knn	0.752232	0.750382
Naive Bayes	0.714857	0.703404
Decision Tree	0.753567	0.749391
Random Forest	0.754902	0.751595
Support Vector Machine	0.755270	0.750574

R Portfolio

프로젝트 제목

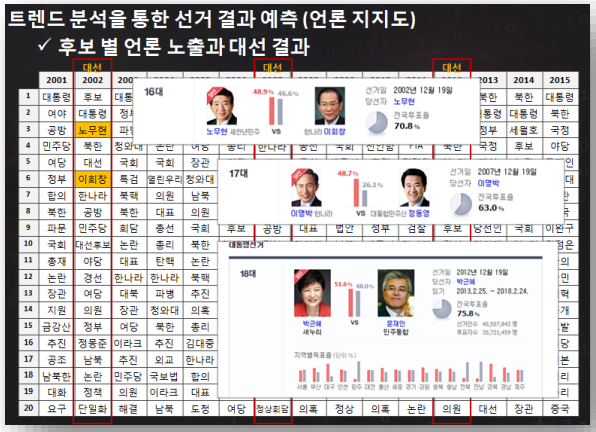
미디어 트랜드 분석

특.장점

- 데이터 수집과 전처리
  - ✓ 뉴스기사 크롤링&전처리
- 키워드 분석
  - ✓ 분야별(문화,국제,경제,정치, 스포츠) 키워드
- 특정 분야 연관분석

프로젝트 내용

15년 간의 "헤드라인 뉴스" 기사를 통해서 한국 사회 시대별 트랜드 분석



R Portfolio

프로젝트 제목

서울시 유기견 현황분석

특.장점

- Web 데이터 크롤링(Crawling)
  - Python code 이용 유기견 정보 수집
- 5년 vs 3년 기준 추이 분석
  - ✓ 최적의 ARIAM 모형 탐색

프로젝트 내용

서울시 유기견 현황 분석을 위해서 7년간 유기견 정보를 웹에서 크롤링 하여 시계열 분석

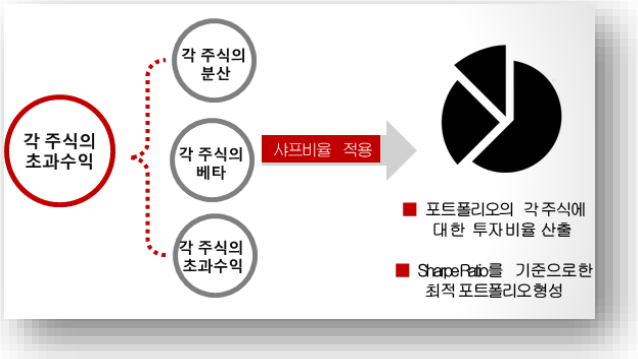
R Portfolio

프로젝트 제목

샤프비율 (Sharpe Ratio)를
기준으로 한 최적 포트폴리오 구성

프로젝트 내용

최적 포트폴리오 배분률에 따른
분산 투자 효율성 평가 & 분석



특.장점

- 분산 투자 관련 모형 및 용어
  - ✓CAMP, Beta, Sharpe Ratio
- 최적 포트폴리오 구성과 평가/분석
  - ✓ 효율성 평가/Monte Carlo Simulation

Python Project

프로젝트 제목

보험 사기자 예측 알고리즘 개발

프로젝트 내용

보험 사기 실제 데이터 기반
사기여부와 유의미한 데이터
프레임 생성 및 모델링

특.장점

- 변수 간 상관관계 탐색/시각화
- 유의미한 파생변수 생성
  - ✓ 사기여부와 유의미한 데이터 프레임 생성
- 교차검증 : 과적합 문제 검증

탐색 → 파생변수 → 모델링 → 검증

Python Project

프로젝트 제목

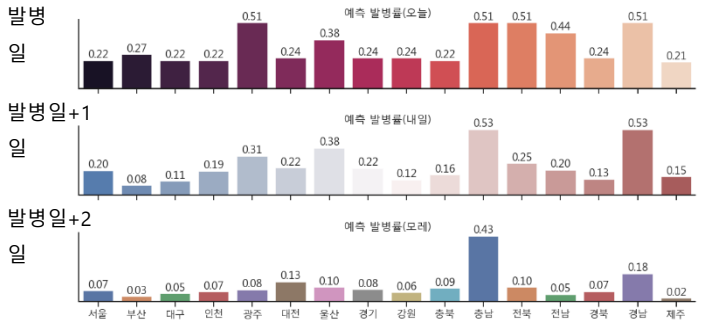
날씨에 따른 매개 감염병 확산 지역 예측

특.장점

- Selenium : 발병률 데이터 크롤링
- Keras API 이용
- 질병관리 본부 공모전
- R(EDA) + Python(Model)

프로젝트 내용

감염병을 옮기는 매개체(모기, 진드기, 쥐 등)의 생태계 특징을 파악하여 기상 기후 변화에 증식 지역 예측



Python Project

프로젝트 제목

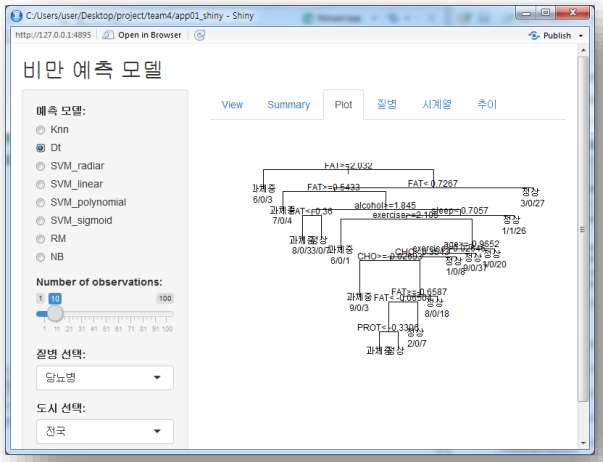
비만에 영향을 미치는 요인

특.장점

- 비만 예측 모델 생성/비교
  - kNN,DT,SVM,RM,NB
- Shiny 프로젝트
  - ✓ 예측 모델$인프라 예측 현황

프로젝트 내용

비만의 원인과 예측, 비만 합병증에 따른 인프라 예측



Tensorflow Project

프로젝트 제목

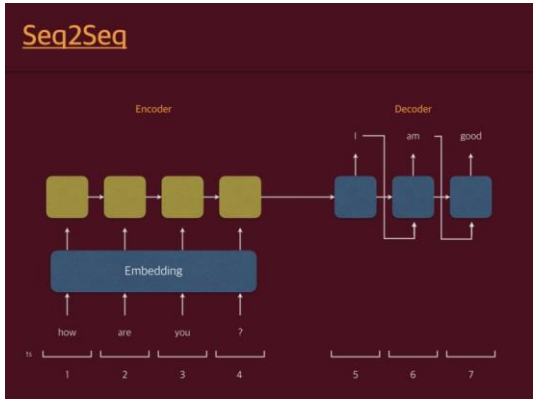
헬로, 해리포터-쳇봇

특.장점

- 영화대본 자료수집/전처리
- Tokenize, Tagging
- Word2Vec, Seq2Seq
- RNN

프로젝트 내용

사용자와 등장 인물 간에 직접적인 의사 소통을 위한 쳇봇 구현



Tensorflow Project

프로젝트 제목

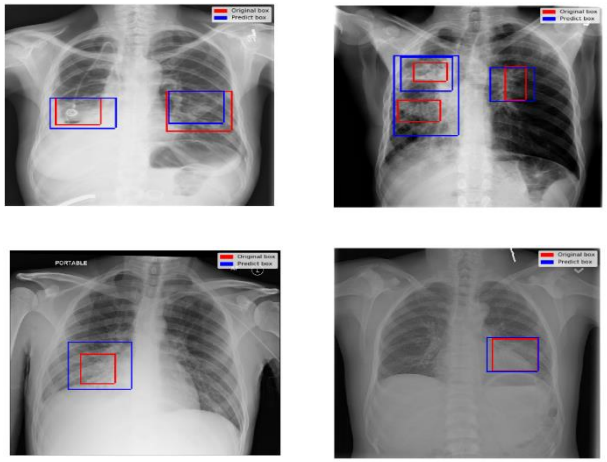
Hospit A.I

특.장점

- YOLO v3 Deep Learning Network 적용
- CNN
- 의학영상진단 예측력 향상

프로젝트 내용

딥러닝 기반 의학영상진단 구현



Q & A