Python & Spark 기반 분석 스터디

◎ 수업 개요

- 데이터 분석 기본 개념 학습

◎ 수업 목표

- 수업목표
  - 파이썬을 이용한 데이터분석 기초 개념 학습
  - 스파크를 이용한 데이터분석 기초 개념 학습

주차명		학습활동
1	**서울시 구별 CCTV 현황 분석**	1-1 CCTV 현황과 인구 현황 데이터 구하기 1-2 파이썬에서 텍스트 파일과 엑셀 파일을 읽기 pandas 1-3 pandas 기초 익히기 1-4 pandas를 이용해서 CCTV와 인구 현황 데이터 파악하기 1-5 pandas 고급 기능 두 DataFrame 병합하기 1-6. CCTV 데이터와 인구 현황 데이터를 합치고 분석하기 1-7 파이썬의 대표 시각화 도구 Matplotlib 1-8 CCTV 현황 그래프로 분석하기
2	**서울시 범죄 현황 분석**	2-1 데이터 획득하기 2-2 pandas를 이용하여 데이터 정리하기 2-3 지도 정보를 얻을 수 잇는 Google Maps 2-4 Google Maps를 이용해서 주소와 위도, 경도 정보 얻기 2-5 pandas의 pivot_table 학습하기 2-6 Pivot_table을 이용해서 데이터 정리하기 2-7 데이터 표현을 위해 다듬기 2-8 좀 더 편리한 시각화 도구 Seaborn 2-9 범죄 데이터 시각화하기 2-10 지도 시각화 도구 Folium 2-11 서울시 범죄율에 대한 지도 시각화 2-12 서울시 경찰서별 검거율과 구별 범죄 발생율을 동시에 시각화하기
3	**시카고 샌드위치 맛집 분석**	3-1. 웹 데이터를 가져오는 Beautiful Soup 익히기 3-2 크롬 개발자 도구를 이용해서 원하는 태그 찾기 3-3 실전: 시카고 샌드위치 맛집 소개 사이트에 접근하기 3-4. 접근한 웹 페이지에서 원하는 데이터 추출하고 정리하기 3-5 다수의 웹 페이지에 자동으로 접근해서 원하는 정보 가져오기 3-6 Jupyter Notebook에서 상태 진행바를 쉽게 만들어주는 tqdm 모듈

주차명		학습활동
		3-7 상태 진행바까지 적용하고 다시 샌드위치 페이지 50개에 접근하기 3-8 50개 웹 페이지에 대한 정보 가져오기 3-9 맛집 위치를 지도에 표기하기 3-10 네이버 영화 평점 기준 영화의 평점 변화 확인하기 3-11 영화별 날짜 변화에 따른 평점 변화 확인하기
4	**셀프 주유소는 정말 저렴할까**	4-1 Selenium 사용하기 4-2 서울시 구별 주유소 가격 정보 얻기 4-3. 구별 주유 가격에 대한 데이터의 정리 4-4. 셀프 주유소는 정말 저렴한지 boxplot으로 확인하기 4-5. 서울시 구별 주유 가격 확인하기 4-6. 서울시 주유 가격 상하위 10개 주유소 지도에 표기하기
5	**시계열 데이터를 다뤄보자**	5-1. Numpy의 polyfit으로 회귀(regression) 분석하기 5-2. Prophet 모듈을 이용한 forecast 예측 5-3. Seasonal 시계열 분석으로 주식 데이터 분석하기 5-4. Growth Model과 Holiday Forecast
6	**자연어 처리 시작하기**	6-1 한글 자연어 처리 기초 - KoNLPy 및 필요 모듈의 설치 6-2 한글 자연어 처리 기초 6-3 워드 클라우드 6-4 육아휴직 관련 법안에 대한 분석 6-5 Naive Bayes Classifier 의 이해 영문 6-6 Naive Bayes Classifier 의 이해 한글 6-7 문장의 유사도 측정하기 6-8 여자 친구 선물 고르기
7	**스칼라와 스파크를 활용한 데이터 분석**	7-1 데이터 과학자를 위한 스칼라 7-2 스파크 프로그래밍 모델 7-3 레코드 링크 7-4 스파크 셸과 SparkContext 시작하기 7-5 클러스터에서 클라이언트로 데이터 가져오기 7-6 클라이언트에서 클러스터로 코드 보내기 7-7 RDD에서 Data Frame으로 7-8 DataFrame API로 데이터 분석하기 7-9 데이터프레임에 대한 빠른 요약 통계 7-10 데이터프레임의 축 회전과 형태변환 7-11 데이터프레임을 결합하고 특징 선택하기 7-12 실제 환경을 위한 모델 준비하기 7-13 모델 평가 7-14 한 걸음 더 나아가기

주차	명	학습활동
8	**음악 추천과 Audioscrobbler 데이터셋**	8-1 데이터셋 8-2 교차 최소 제곱 추천 알고리즘 8-3 데이터 준비하기 8-4 첫 번째 모델 만들기 8-5 추천 결과 추출 검사하기 8-6 추천 품질 평가하기 8-7 AUC 계산하기 8-8 하이퍼파라미터 선택하기 8-9 추천 결과 만들기
9	**의사 결정 나무로 산림 식생 분포 예측하기**	9-1 회귀로 돌아와서 9-2 벡터와 특징 9-3 학습 예제 9-4 의사 결정 나무와 랜덤 포레스트 9-5 Covtype 데이터셋 9-6 데이터 준비하기 9-7 첫 번째 의사 결정 나무 9-8 의사 결정 나무 하이퍼파라미터 9-9 의사 결정 나무 튜닝하기 9-10 범주형 특징 다시 살펴보기 9-11 랜덤 포레스트 9-12 예측하기
10	**K-평균 군집화로 네트워크 이상 탐지하기**	10-.1 이상 탐지 10-.2 K-평균 군집화 10-.3 네트워크 침입 10-.4 KDD 컵 1999 데이터셋 10-.10- 첫 번째 군집화하기 10-.6 k 선정하기 10-.7 R에서 시각화하기 10-.8 특징 정규화 10-.9 범주형 변수 10-.10 엔트로피와 함께 레이블 활용하기 10-.11 군집화하기
11	**숨은 의미 분석으로 위키백과 이해하기**	11-1 문서-단어 행렬 11-2 데이터 구하기 11-3 파싱하여 데이터 준비하기 11-4 표제어 추출 11-5 단어빈도-역문서빈도(TF-IDF) 계산하기 11-6 특잇값 분해

주차명		학습활동
		11-7 중요한 의미 찾기 11-8 낮은 차원 표현에 대한 의문과 고찰 11-9 단어와 단어 사이의 연관도 11-10 문서와 문서 사이의 연관도 11-11 문서와 단어 사이의 연관도 11-12 여러 개의 단어로 질의하기 11-13 한 걸음 더 나아가기
12	**몬테카를로 시뮬레이션으로 금융 리스크 추정하기**	12-1 전문 용어 12-2 VaR 계산 방법 12-3 우리의 모델 12-4 데이터 구하기 12-5 전처리하기 12-6 요인 가중치 결정하기 12-7 표본추출 12-8 실험 실행하기 12-9 수익 분포 시각화하기 12-10 결과 평가하기 12-11 한 걸음 더 나아가기

◎ 교재

1. 파이썬으로 데이터 주무르기 24,750 원
(http://www.kyobobook.co.kr/cooper/redirect_over.jsp?LINK=NVB&next_url=http://www.kyobobook.co.kr/product/detailViewKor.laf?mallGb=KOR&ejkGb=KOR&linkClass=&barcode=9791186697474)
2. 9 가지 사례로 익히는 고급 스파크 분석 23,400 원
(http://www.kyobobook.co.kr/cooper/redirect_over.jsp?LINK=NVB&next_url=http://www.kyobobook.co.kr/product/detailViewKor.laf?mallGb=KOR&ejkGb=KOR&linkClass=&barcode=9791162240526)