[K-Digital Training]
인공지능 통합과정

Semi-project
for Part 1 ~ Part 4-1

Daeyeon Jo
repositivator@gmail.com

Course Overview

* 아래 커리큘럼의 세부 사항은 변동될 수 있습니다.
* 진도 상황에 따라 1~2일 정도 차이가 발생할 수 있습니다.

파이썬 프로그래밍 기초 (프로그래밍의 3가지 축 / 파이썬 자료구조 등)
파이썬 정형 데이터 분석 (데이터 탐색 / 데이터 전처리 / 데이터 시각화)

파이썬을 활용한 데이터 수집 & 웹 스크레이핑 (+ 자동화 프로그램 개발)
파이썬 기반 텍스트 데이터 분석

Python 기초 수학 & 통계분석 (빈도분석 / 기술통계 / 교차검정 / 평균차이검정)

1차 세미 프로젝트 (데이터 수집 / 탐색 및 전처리 / 통계 분석 / 시각화 / 팀별 발표)

SQL 기초 프로그래밍 (Data Modeling / SQL CRUD / Adv. Techniques)

머신러닝 핵심 이론 & 주요 알고리즘 이론
파이썬 기반 머신러닝 알고리즘 실습 (Scikit-learn)
+ 데이터 분석 관련 직무 & 학습 리소스 소개

2차 세미 프로젝트 (Feature engineering & applying ML algorithms)

Mid-Project (텍스트 데이터 수집 / 데이터 전처리 / 각종 텍스트 기반 분석기법 적용)

딥러닝 핵심 이론 & 인공신경망 최적화 이론
파이썬 기반 딥러닝 알고리즘 실습 (Tensorflow & Keras)
+ 분야별 머신러닝 & 딥러닝 활용 사례 소개 + 각종 자동화 도구 실습

3차 세미 프로젝트 (데이터 수집 / 탐색 / 전처리 / 시각화 + ML&DL model tuning)

* 매 정시마다 메인 회의실에서 출석 채팅 진행합니다.

수업 관련 공지사항

* **데이터 선정 / 분석 범위 선정 / 분석 방법 선정 모두 자유입니다.** (배운 내용의 복습에 Focus!)

* Part 1~4-1 에서 **배운 지식들을 최대한 빠짐없이 활용**하는데 초점을 맞춰주세요.
  크롤링 필수 X & 추후 포트폴리오로 활용될 프로젝트라고 여기고 최대한 상세히 & 꼼꼼하게 발전시켜보세요.

* **발표 시 포함할 사항 : 분석 주제 및 목표 / 데이터 수집 방법 / 분석 프로세스 / 분석 결과**
  발표 시 제출할 사항 : 발표 자료 (ppt or pdf) / 전체 코드 with 주석 (.ipynb) / 원본 데이터

* **6/23 수요일 14:30 : 팀별 발표 및 질의응답 (15~20분 내외/팀, 최대 20분)**
  : 6/23 (수) 14:20 전까지 발표 자료 & Jupyter notebook(+원본 데이터) 제출 @ 슬랙 DM

* 발표 시작 시간은 일정에 따라 변동될 수 있습니다 & **도움이 필요할 경우 슬랙 채널에서 호출**

수업 관련 공지사항

1팀 : 박용민, 김현정, 민정현, 박건우, 이준동

2팀 : 이규호, 권산하, 이용석, 임주란

3팀 : 박민수, 조윤정, 주리아, 한창환

4팀 : 이소연, 강민정, 박정재, 황준우

5팀 : 김주연, 강원석, 안성훈, 이병준, 전가은

6팀 : 류범상, 이혜민, 정소연, 조성곤, 최용수

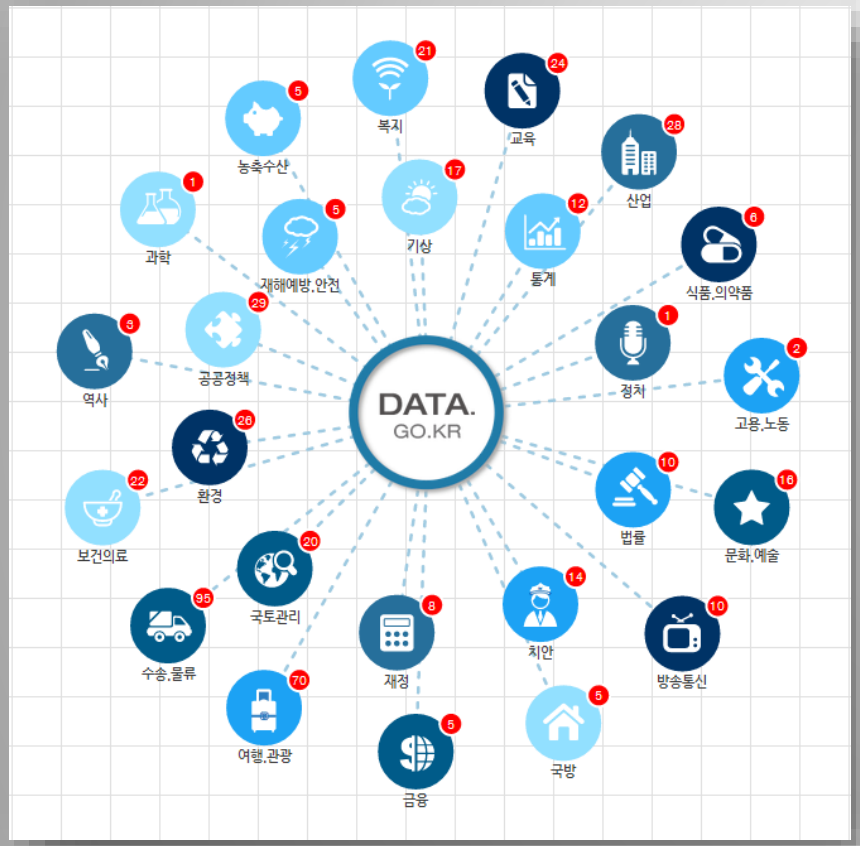
Various data collection - etc (Datasets / Data repository)

Awesome Public Datasets @ https://github.com/awesomedata/awesome-public-datasets
Google AI Datasets @ https://ai.google/tools/datasets
Google Dataset Search @ https://toolbox.google.com/datasetsearch
SKT BigData Hub @ https://www.bigdatahub.co.kr
Kaggle competition datasets @ https://www.kaggle.com/datasets
(ex. **Google Play Store Apps data** @ http://j.mp/2PDhbKR)

https://data-on.co.kr - **데이터온 (대한민국의 모든 데이터를 한 곳에서, 누구나 쉽게 찾고 활용하는 데이터플랫폼)**
http://www.aihub.or.kr - **AI 오픈이노베이션 허브 (한국어 음성 & 대화, 한국인 안면, 법률/특허/헬스케어/관광/농업/이미지 데이터)**
https://openapi.kftc.or.kr & https://developers.kftc.or.kr/dev - **금융결제원 오픈API 통합포털 (오픈뱅킹 & 금융인증)**
https://golmok.seoul.go.kr - **서울시 우리마을가게 상권분석 서비스**
http://data.seoul.go.kr - **서울 열린 데이터 광장**
https://www.dataquest.io/blog/free-datasets-for-projects - **19 Places to Find Free Data Sets for Data Science Projects**
http://dataportals.org - **A Comprehensive List of Open Data Portals from Around the World**
https://www.kdnuggets.com/datasets/index.html - **Datasets for Data Mining/Science**

* 각종 데이터분석 관련 공모전/대회/프로젝트사례 모음 @ http://j.mp/2MPDfON
* 해외 기업의 인공지능 데이터 개방과 활용 현황 (구글 사례를 중심으로) @ http://j.mp/2paKt7j
* 딥러닝 학습을 위한 국내외 데이터셋 현황 (이미지/동영상/음성) @ http://j.mp/2BSShNy / http://j.mp/2roFj8i / http://j.mp/2VbHXeG
* 2020 **문화체육관광** 빅데이터플랫폼 데이터 설명회 (한국문화정보원, 국립중앙도서관, 국민체육진흥공단 등) https://j.mp/30OnHkM & https://j.mp/30KFuJH
* KLUE: Korean Language Understanding Evaluation (**한국어 NLP 데이터셋** & pretrained-LM 베이스라인) @ https://j.mp/3woY9Iy & https://j.mp/3f7p0mU

Various data collection – Public data & Open data (APIs & files)



- 공공 데이터 포털 : https://www.data.go.kr
- 국가 통계 포털 : http://kosis.kr
- MDIS(MicroData Integrated Service) : https://mdis.kostat.go.kr

* 오픈 API를 통한 공공데이터 수집 (서울열린데이터 광장) @ http://j.mp/2AWRA5g
* 코로나 확진자 동향 데이터 수집 및 시각화 (공공데이터포털 코로나19 감염 현황 OpenAPI 활용) @ https://j.mp/3bHa8dC

Various data collection – Unowned data

Use APIs & Web scraper

- APIs (Twitter, Facebook, Instagram, etc)
- Bots (Web crawler, Web scraper)

* 여기어때, 야놀자 DB 무단수집 위법 판결 @ https://j.mp/3fgxi9Q + 크롤링과 저작권 침해 고소 진행 일대기 @ https://j.mp/3k5vbbl
* **Listly (크롬 확장프로그램 for 웹크롤링)** @ https://j.mp/2LSb8kh / 네이버 크롤링 라이브러리 Kocrawl (날씨/미세먼지/지도/맛집/맞춤법) @ https://j.mp/2CbdRA8
* Web Scraping Tool & Web Data Extractor : **ScrapeStorm** @ http://j.mp/2Y4porj / **Octoparse** @ https://j.mp/3o5i23q / **Automatio** @ https://automatio.co

The process of web scraping (detailed)

URL 분석 (패턴 존재 여부, query 종류)
ex. https://alldic.daum.net/search.do?q=happiness

↓

URL 구성 (str, 추후 자동화 고려)

HTTP Response 얻기 : urlopen(URL) **or** request.get(URL).content

HTML source 얻기 : BeautifulSoup(HTTP Response, 'html.parser')

↓

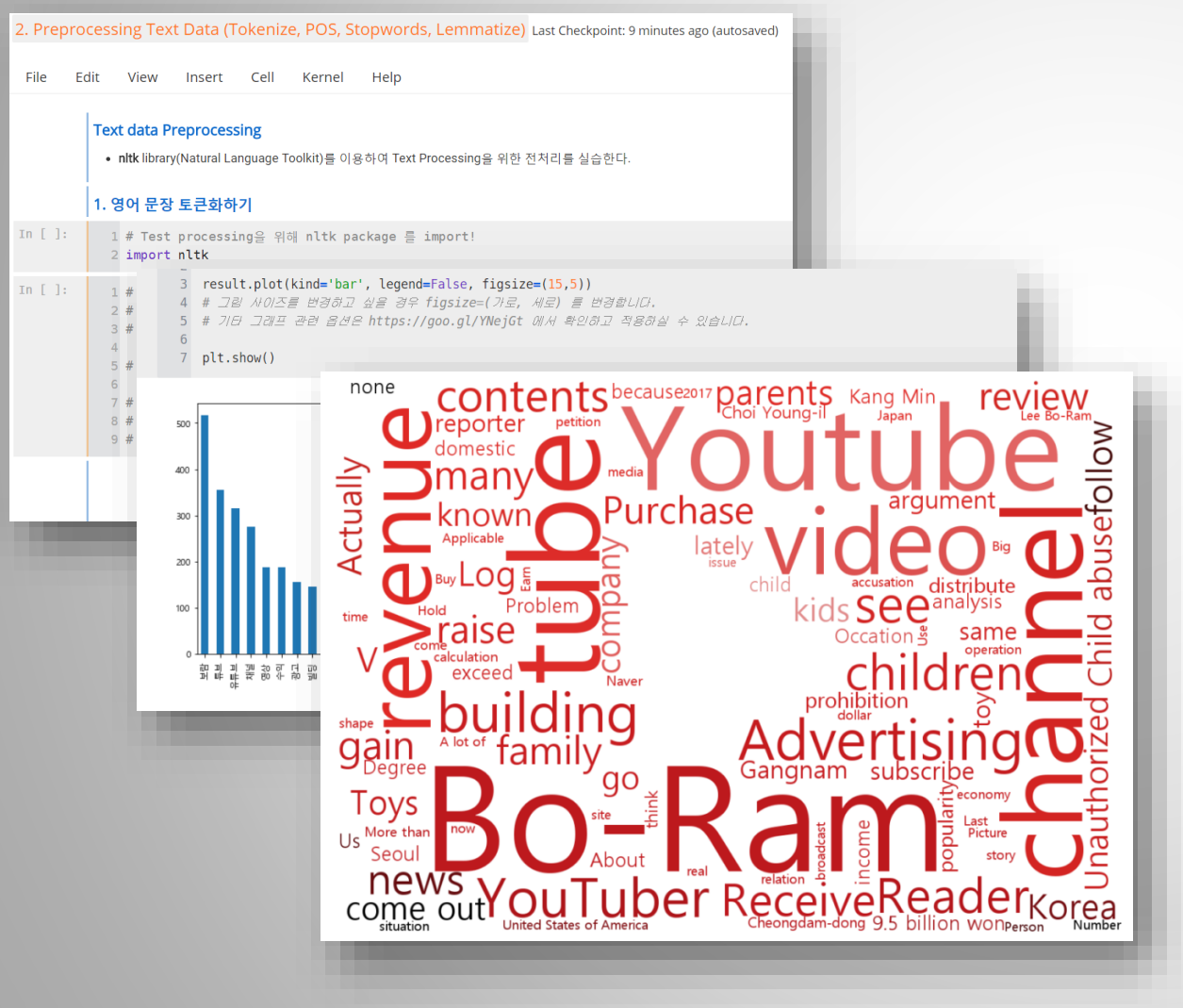
HTML Tag 꺼내기 (Tag 이름, Attr 이름, Attr 값 지정)

.find(~) : 1개의 Tag (조건이 동일한 Tag가 여러 개일 경우에는 첫번째 Tag)
.find_all(~) : 여러 개의 Tag를 찾은 다음 for 문으로 Tag 단위로 꺼내어 활용

↓

Tag 로부터 텍스트 혹은 Attribute values 꺼내기 : Tag.get_text() **or** Tag.attrs

The process of data analysis for text data



텍스트 데이터를 str 자료형으로 준비

↓

Tokenize (형태소 분석)

POS Tagging (Part-of-speech, 품사 표시)

Stopwords 제거 (불용어 제거)

↓

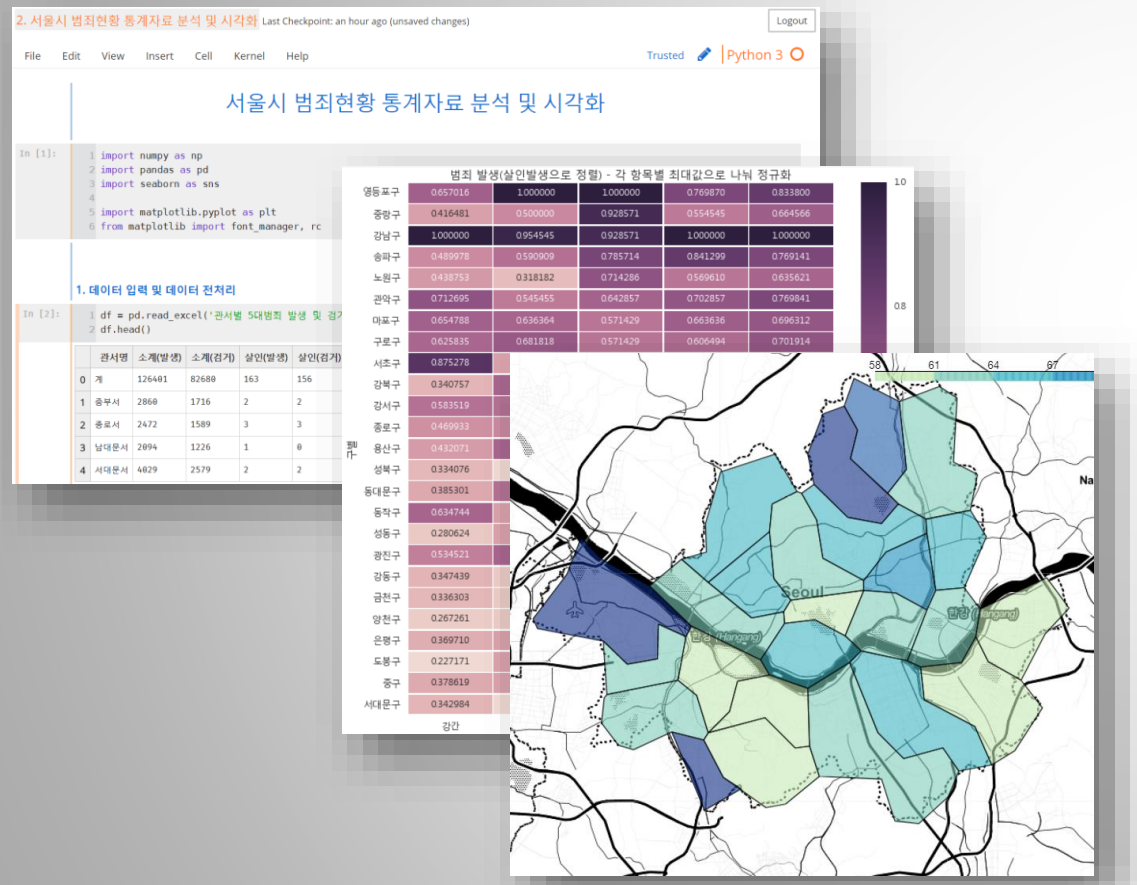
단어 갯수 카운팅 & 단어 사전 생성

단어 사전 기반 데이터 시각화

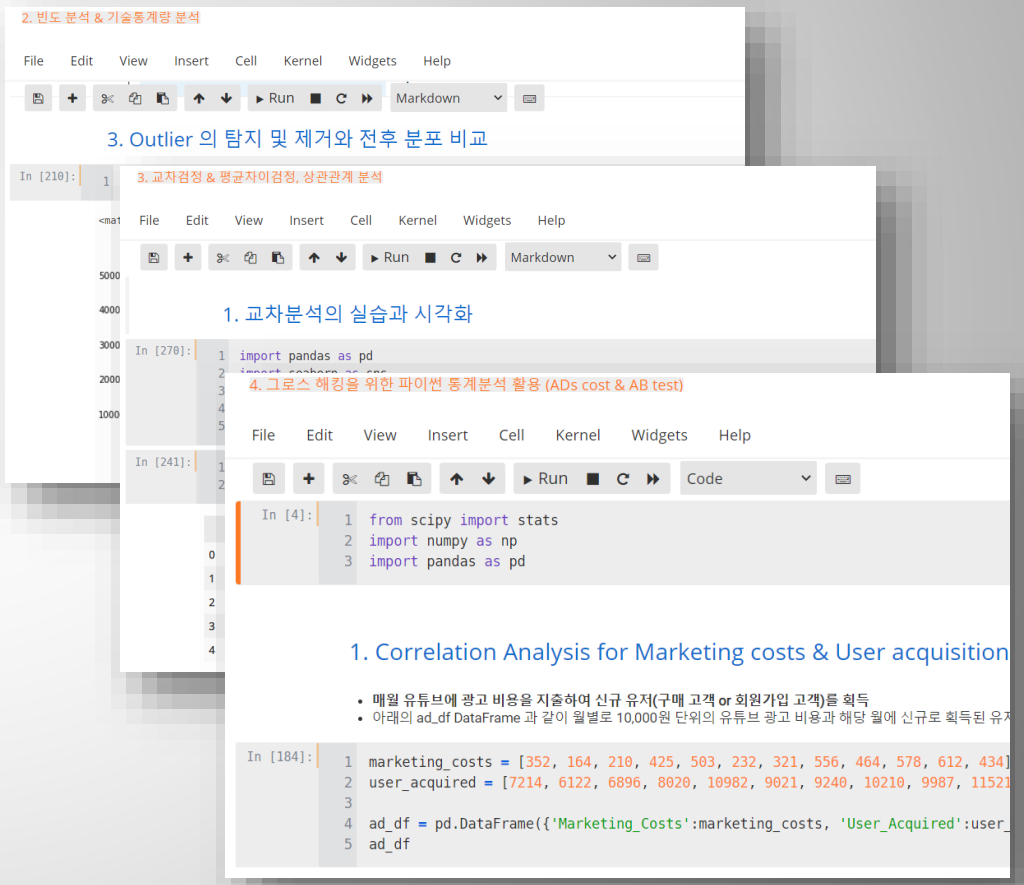
(+ 머신러닝/딥러닝 모델 적용)

The process of structured data analysis & statistical analysis

Part 2. 파이썬을 활용한 정형데이터 분석과 시각화



Part 4-1. 파이썬을 활용한 기초 통계분석



* Top 6 Python libs for Visualization: Matplotlib/Seaborn/Plotly/Bokeh/Altair/Folium (장단점) @ https://j.mp/30772sU
* 데이터의 종류와 시각화 목적에 따른 다양한 차트 (유의점 및 소스코드 포함) @ https://goo.gl/ErLHCY

[K-Digital Training]
인공지능 통합과정

End of Document