Jaram Summer Workshop 대체 과제

• Before You Begin!

Language: Python 3.5+

Due date: 2020 / 09 / 01 Tuesday 23:59:59

PART A: 웹 크롤링

PART B: 머신 러닝

PART A 와 PART B 모두 템플릿이 있으니 사용하여 진행해 주시면 됩니다.

PIP 사용법

• PIP이란 Python 패키지 관리 도구이다.

• Python 3.5+ 이상을 사용하면 기본적으로 같이 설치가 된다.

ex. pip install 패키지명

폴더 안에 requirements.txt를 사용하여 필요한 패키지를 설치할 수 있다.

pip install -r requirements.txt

과제 진행 중 질문 사항이 있으면 임원진에게 연락하시길 바랍니다. >.<

제출 방법: 학술 부장한테 완성된 파일을 주시면 됩니다.

1. PART A (웹 크롤링)

목적: 네이버에 특정 단어에 대한 자료를 웹 크롤링을 한다.

1) 웹 크롤링에 사용할 모듈을 가져온다.

hint. requirements.txt를 살펴보면 된다.

2) ‘urllib’을 사용해서 ‘어몽 어스’ 키워드를 가진 뉴스를 모든 페이지에 대하여 네이버에 검색하기

3) 검색된 뉴스들에 대하여 제목을 ‘news_title’ 에 저장한다.

4) 검색된 뉴스들 중에서 제목에 ‘폴 가이즈’를 가진 제목을 found 에 저장한다.

5) 검색된 뉴스들의 개수와 found을 출력한다.

2. PART B (간단한 머신 러닝) 주택 가격 예측

1) 머신 러닝에 필요한 모듈을 가져온다.

2) 폴더 안에 있는 ‘boston.csv’를 가져온다.

hint. ‘pandas’에 있는 ‘read_csv’ 활용한다.

3) boston.csv 안에 없는 값들은 열의 평균으로 대체한다.

hint. ‘sklearn.impute’에 있는 ‘SimpleImputer’활용

4) boston.csv 에 있는 값에 대하여 요약 통계를 구성한다.

i. describe 메소드를 사용하여 요약 통계 하나를 구성한다.

ii. ‘heatmap’를 출력한다. hint. ‘seaborn’, ‘matplotlib.pyplot’

5) ‘heatmap’을 활용하여 관계가 높는 값들을 이용해 데이터 셋을 구성한다.

6) 머신 러닝 알고리즘 중에서 선형 회귀를 이용하여 머신 러닝을 구현한다.

i. ‘train_set’과 ‘test_set’으로 데이터를 나눈다.

ii. LinearRegression을 활용하여 예측 값을 출력한다.

iii. ‘mean_squared_error’을 활용하여 오류에 대하여 값 출력한다.

hint. ‘sklearn.metrics’안에 정의되어 있다.