

데이터과학



타슈데이터 - Python

지난주..

- 추가 수강신청 학생
 - 2일뒤(**3월 11일 오후 6시**)까지 저번주 과제 제출
- 제출 기한 : 수업 전날 오후 6시까지
- 제출 형태 맞추기

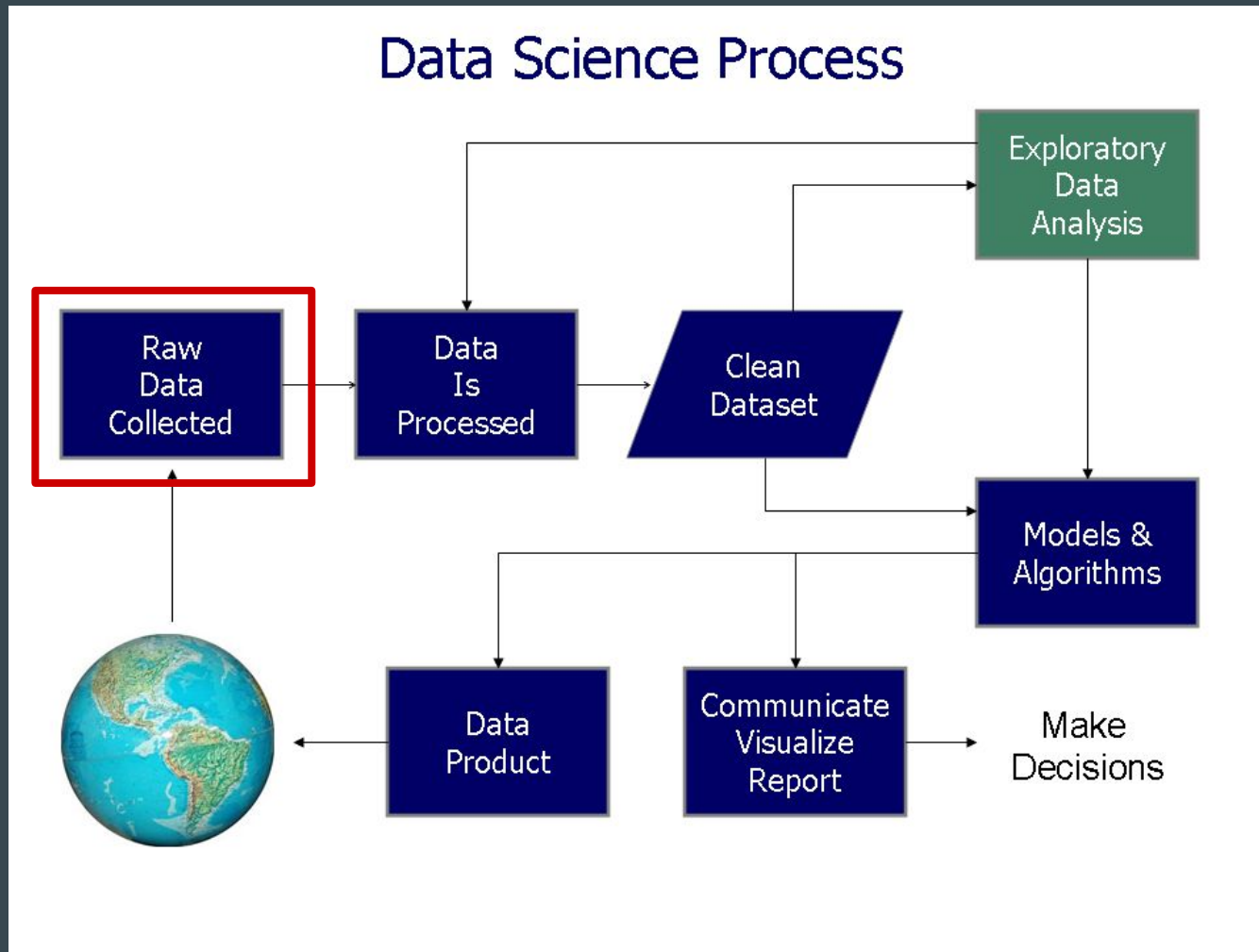
	<u>링크</u> https://www.datacamp.com/statement-of-accomplishment/course/a9d2ded7779da6d0cf34c6335ca025434337c380
	<u>링크</u> https://www.datacamp.com/statement-of-accomplishment/course/587f573e1db6397d53d172c79f3ff60cb676d416

목표

- Data Analysis Process 따라하기
- 타슈데이터 받기
 - 공공데이터 포털
- 파이썬을 사용한 타슈데이터 분석
 - 인기 정류장 TOP 10
 - 인기 경로 TOP 10

Data Analysis Process

- https://en.wikipedia.org/wiki/Data_analysis



1. Collecting Raw Data - 공공데이터

- 정부에서 공개하는 공공데이터
 - CSV, JSON, API 다양한 형태로 제공됨

정부 **3.0** DATA 공공데이터포털
GO . KR

데이터셋

활용사례

참여마당

정보공유

검색어를 입력하세요.



 FILE DATA

 OPEN API

 STANDARD DATA



교육



국토관리



공공행정



재정금융



산업고용



사회복지



식품건강



문화관광



보건의료



재난안전



교통물류



환경기상



과학기술



농축수산



통일외교안보



법률

1. Collecting Raw Data - 타슈 데이터

- “타슈”로 검색했을 때 가장 처음에 나오는 데이터
- 타슈 대여정보 (사용자가 빌린날, 빌린정류장, 반납한날, 반납정류장)
- 정류장 정보(정류장이름, 정류장번호, 위치)

홈 / 데이터셋 / 통합검색

Q 상세검색

연관 검색어 무인공공자전거

전체(1)

파일데이터(1)

오픈API(0)

표준데이터(0)

전체 1건을 찾았습니다.

📄 파일데이터 [1건]

날짜순 ▾

제목순 ▾

조회순 ▾

다운로드순 ▾

타슈(무인공공자전거) 운영정보 조회수 : 2,493 다운로드수 : 2,774
수정일 : 2016.05.25 기관 : 대전광역시시설관리공단 서비스유형 : 다운로드
타슈(무인공공자전거) 운영정보 데이터 제공

CSV XLS XLSX

보통파일

인기검색어

1

부동산

2

날씨

3

음식점

4

광명시

5

서대문구

국가중점데이터

서비스유형별데이터

1. Collecting Raw Data - 타슈 데이터(cont'd)

- 2013년, 2014년, 2015년 총 3년의 데이터 존재 (모두 다운로드)
- 형식은 모두 다름

타슈(무인공공자전거) 운영정보

타슈(무인공공자전거) 운영정보 데이터 제공

매체유형 : 텍스트 파일, 링크 건수 : 6 전체 행 수 : N/A 확장자 : CSV / XLS / XLSX 다운로드 횟수(바로가기 횟수) : 2774

☐ 전체 **선택 다운로드**

※ 서비스 오류가 있을시 오류신고 버튼을 이용해주세요.

<input type="checkbox"/> CSV 2015년도 상/하반기 타슈대여현황 이력 ...	<input type="checkbox"/> XLSX 2014년도 하반기 타슈대여현황 이력 데...
다운로드 상세정보 오류신고	다운로드 상세정보 오류신고
<input type="checkbox"/> XLSX 2014년도 상반기 타슈대여현황 이력 데...	<input type="checkbox"/> CSV 2015년 3월 타슈(무인공공자전거) 스테...
다운로드 상세정보 오류신고	다운로드 상세정보 오류신고
<input type="checkbox"/> XLSX 2013년도 하반기 타슈대여현황 이력 데...	<input type="checkbox"/> XLSX 2013년도 상반기 타슈대여현황 이력 데...
다운로드 상세정보 오류신고	다운로드 상세정보 오류신고

2. Data is Processed?

- Data initially obtained must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (i.e., structured data) for further analysis, such as within a spreadsheet or statistical software.

- 데이터가 모두 가공처리되었는지?
 - 가공처리 필요
 - 원하는 column만 선택

	A	B	C	D
1	대여 스테이션번호	대여일시	반납 스테이션번호	반납일시
2	16	20150100000000	16	20150100000000
3	16	20150100000000	110	20150100000000
4	35	20150100000000	20	20150100000000
5	102	20150100000000	95	20150100000000
6	34	20150100000000	50	20150100000000
7	56	20150100000000	32	20150100000000
8	11	20150100000000	18	20150100000000
9	11	20150100000000	41	20150100000000
10	30	20150100000000	13	20150100000000
11	4	20150100000000	7	20150100000000

2015년 데이터

	A	B	C	D	E	
1	회원구분	대여 스테이션 정보	대여일시	반납 스테이션 정보	반납일시	총대여시간
2	No	105	20140101000005	105	20140101000304	0
3	Yes	71	20140101000411	45	20140101002733	0
4	Yes	73	20140101000916	80	20140101001447	0
5	Yes	14	20140101001203	43	20140101002139	0
6	Yes	14	20140101001205	25	20140101011345	0
7	Yes	14	20140101001219	17	20140101001453	0
8	Yes	14	20140101001221	25	20140101011335	0
9	Yes	60	20140101001401	32	20140101003953	0
10	Yes	60	20140101001417	56	20140101003751	0
11	Yes	60	20140101001433	56	20140101003736	0
12	Yes	60	20140101001435	56	20140101003023	0
13	Yes	60	20140101001517	61	20140101002101	0
14	Yes	60	20140101001543	61	20140101002127	0
15	Yes	78	20140101001749	45	20140101002749	0
16	Yes	34	20140101002217	21	20140101004553	0
17	Yes	34	20140101002221	21	20140101004621	0
18	Yes	34	20140101002249	21	20140101004609	0
19	Yes	57	20140101002427	55	20140101002951	0

2014년 데이터

	A	B	C	D	E
1	IS_MEMBER	RENT_STATION	RENT_DATE	RETURN_STATION	RETURN_DATE
2	No	43	20130101055603	34	20130101060217
3	No	97	20130101060400		20130101102037
4	No	2	20130101060406	10	20130101061859
5	No	106	20130101105305	105	20130101105743
6	Yes	4	20130101112223	4	201301011121753
7	No	21	20130101113953	105	20130101114943
8	No	90	20130101120833	91	20130101125136
9	No	13	20130101131429	30	20130101133039
10	Yes	1	20130101133743	1	20130101133815
11	Yes	1	20130101133847	2	20130101150958
12	Yes	1	20130101133847	2	20130101151014
13	No	9	20130101134253	23	20130101142012
14	Yes	27	20130101134328	27	20130101134356
15	Yes	30	20130101134850	7	20130101145651
16	Yes	30	20130101134907	30	20130101135526
17	Yes	30	20130101134909	18	20130101140929
18	Yes	30	20130101134922	8	20130101141746

2013년 데이터

2. Data is Processed?

- Station
- 인코딩문제
 - 한글이 보이지 않음
 - 파일의 인코딩 방식 변경시켜야 함(euc-kr -> utf-8)
 - 리브레오피스 에서 인코딩 변경
 - iconv
 - <http://linuxfortj.blogspot.kr/2011/12/iconv.html>

```
1 'ëÄü ¹«ÄÏ°ø°øÄÜÄü°Ä(Ä,½') ½°Ä×ÄÏ½Ç Ä=°, , , , , , , ^M
2 , , , , , , , ^M
3 ¹øÉÉ,Ä°¿Ä½°Ä°¹øÉÉ,±,°, , íÄ°,Ä$Äí,ÄÖ½Ö,°ÄÄí'ë,ÄÄÇ¥^M
4 1,1,Ä-½°±, , ¹«¿ÄÜ½Ä°ÜÄÖ± (ÄÄ½Ä½Ä° Ää ¾Ö),¿Ç½°Æ÷'Ü,® ,ÄÄ°Æí, Ä-½°±, µµ·æµ¿ 3-8,14,"36.374325,127.387462"^M
5 2,2,Ä-½°±, , 'ëÄÜÄÄ°¥½Ç ¾¾ÄÍ ¾Ö,µÐ»ë'ë±³ ,ÄÄ°Æí, Ä-½°±, µµ·æµ¿ 4-19,20,"36.374472,127.392241"^M
6 3,3,¾ ±, , ÇÑ¹Ç¾ö ñ¿ø(Ä=¹ÄÖ±),ÇÑ¹Ç¾ö ñ¿ø ³», ¾ ±, , ³âµ¿ 396,19,"36.369855,127.388749"^M
7 4,4,¾ ±, , ÄË¿ø¾ÆÄÄ®104µ¿°Ï±Ü(¹ö½°Ä=·ÜÄä),ÄË¿ø¾ÆÄÄ® 104µ¿¾Ö ÄË¹® Ä°±³ ¹ö½°Ä=·ÜÄä ¾Ö, ¾ ±, , ³âµ¿ 401,12,"36.36819
2,127.379281"^M
```

2. Data is Processed?

- 13, 14, 15 년도의 파일 형식이 다름
 - csv, xlsx
- 파일 내부의 형식도 다름
 - column, column명(같은 역할이지만 이름이 다름)
- 하나의 파일로 합쳐야한다!
 - 합칠 때 위에 컬럼명 삭제
 - 리브레오피스
 - Linux Command

■ **cat**

■ **2015-1.csv ,2015-2.csv2014-1.csv**

■ **cat 201* > tashu.csv**

3. Clean Dataset

- tashu.csv (13,14,15년도 합친 파일)
 - 13, 14, 15년도 모두 하나의 파일로 합치면 3,404,663 line
- station.csv (정류장 정보 파일)

tashu.csv

```
2447598 18,20131231235637,12,20140101001205
2447599 4,20131231235654,11,20140101000853
2447600 4,20131231235655,9,20140101000743
2447601 18,20131231235713,44,20140101000805
2447602 47,20131231235717,112,20140101004828
2447603 105,20131231235816,105,20131231235941
2447604 70,20131231235834,108,20140101000425
2447605 29,20131231235907,30,20140101000503
```

station.csv

```
1 번호,키오스크번호,구별,명칭,위치,주소,거치대,좌표
2 1,1,유성구,무역전시관입구(택시승강장 앞),엑스포다리 맞은편,유성구 도룡동 3-8,14,"36.374325,127.387462"
3 2,2,유성구,대전컨벤션 센터 앞,둔산대교 맞은편,유성구 도룡동 4-19,20,"36.374472,127.392241"
4 3,3,서구,한밭수목원(정문입구),한밭수목원 내,서구 만년동 396,19,"36.369855,127.388749"
5 4,4,서구,초원아파트104동부근(버스정류장),초원아파트 104동앞 쪽문 육교 버스정류장 앞,서구 만년동 401,12,"36.368192,127.379281"
6 5,5,서구,둔산대공원 입구(버스정류장),한밭수목원에서 평송수련원 가는길 버스정류장 앞,서구 둔산동 1521-10,13,"36.365034,127.389361"
```

4. Modeling and algorithms

- 반환 정류소 출력 예제

```
1 import csv
2
3 tashu_file = open('tashu.csv','r')
4 tashu = csv.DictReader(tashu_file)
5
6 for rent in tashu:
7     print (rent['RENT_STATION'])
```

```
20
3
139
64
64
3
24
67
3
17
3
3
47
112
43
143
3
3
19
55
69
3
145
147
55
90
118
59
4
135
60
33
102
31
31
```

실습

- tashu.csv(13,14,15년도 합친 데이터) 를 사용
- 대여 정류장 Top10 출력

```
macgongmon-2:code macgongmon$ python3 test.py
Station : 3 Count : 348977
Station : 56 Count : 182114
Station : 31 Count : 166866
Station : 17 Count : 165778
Station : 32 Count : 147063
Station : 33 Count : 142310
Station : 14 Count : 114878
Station : 105 Count : 112921
Station : 21 Count : 111715
Station : 55 Count : 110045
```

과제

1. [필수]가장 인기있는 정류장 Top 10 (정류장 이름 포함)
2. [필수]가장 인기있는 경로 Top 10 (정류장 이름 포함)
3. [선택] 1,2번 문제 이외의 문제
 - 많은 문제, 어려운 문제를 푼 사람은 추가 점수

과제 문제

- 가장 인기있는 정류장 Top 10 (이름 포함)
- 가장 인기있는 경로 Top 10 (이름 포함)
- 각 구별 정류장 개수 비교 (+차트)
- 각 구별 이용 횟수 비교(+차트)
- 요일별 이용 횟수 비교 (+차트)
- 시간별 이용 횟수 비교 (+차트)

과제 진행방법

- [과제 파일]
 - homework_tashu.py
 - test.py
- 과제 1,2번은 필수이므로 스켈레톤 제공
- 코드주석에 적혀있는 대로 input, output을 맞춰야함
- 결과값 테스트
 - `python3 test.py -v` : 1,2번 과제 결과값 일치 테스트
 - `python3 -m unittest test.Test.test_get_top10_station` : 1번문제 테스트
 - `python3 -m unittest test.Test.test_get_top10_trace` : 2번문제 테스트
- 이외의 문제는 코드를 자율적으로 추가하여 진행

과제 제출 방법

- 과제 제출 기한 : **2017년 3월 15일 오후 6시까지!**
- Google Classroom에 제출!
 - 제출마감이후부터 24시간 경과시마다 만점의 20%씩 추가감점
 - 예) 10점 만점에 24시간 경과시 8점 만점, 48시간 경과시 6점 만점, 5일 경과시 제출점수 1점만 있음
- 파일 제목 : **DS_학번_이름_주차.pdf, DS_학번_이름_주차.zip**
 - 보고서(PDF형태) : **HWP, DOC**일 경우 채점 안함
 - 코드(zip형태)
 - homework_tashu.py
 - 그외 추가구현문제 있다면 추가구현파일
 - **코드zip과 보고서를 하나로 압축하지 말 것!**
 - **파일 제목 및 형태 틀리면 -1점**

채점 기준

- test.py 통과 여부
 - 답이 일치하는가?
- 코드
 - 정상적으로 데이터 가공을하였는가?
 - 답을 static하게 고정하였는지?
- 보고서
 - 과제를 진행하는 과정 설명
 - 데이터 가공 및 합치기
 - 과제문제 해결과정
 - 결과 화면

참고문서

- 데이터 타입
 - <https://docs.python.org/2/tutorial/datastructures.html>
- 정렬
 - <https://wiki.python.org/moin/HowTo/Sorting>
- 정렬된 데이터 출력 (about tuple)
 - <http://www.dotnetperls.com/tuple-python>
- Python matplotlib 이용
 - http://matplotlib.org/users/pyplot_tutorial.html

질문 사항

- 방문
 - 606호 (데이터네트워크 연구실)
- 메일
 - dbgustlr92@cs-cnu.org
- Google Class Room
 - Good!

기타

- 과제 채점 점수
- https://docs.google.com/a/cs-cnu.org/spreadsheets/d/1qR_Alрма9eIqFlvpNW8zMRgoWB7RzoLXtIoRP32ELbI/edit?usp=sharing