데이터과학 실습 보고서

2017.3.15

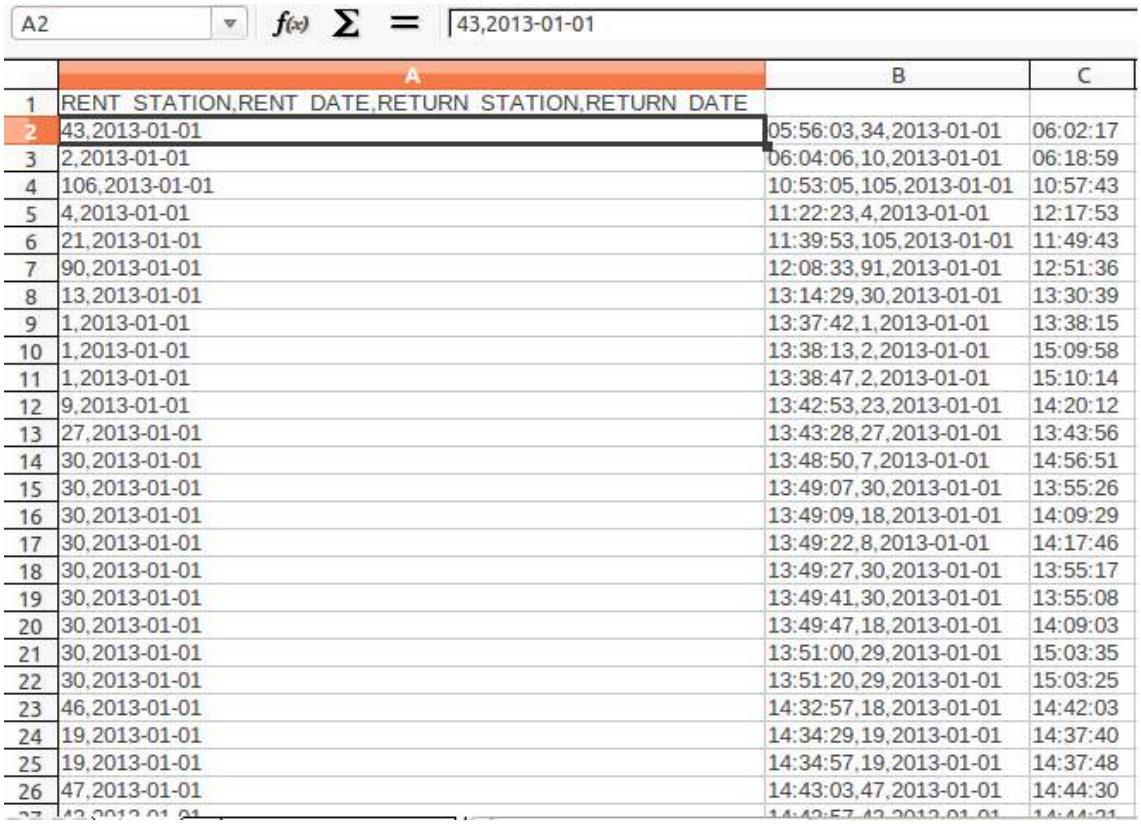
201201185 장진우

1.데이터 가공 및 합치기

우선 데이터를 RENT_STATION, RENT_TIME, RETURN_STATION, RETURN_TIME 4가지의 컬럼으로 정리를 하고, 상단의 컬럼을 제외한 모든 컬럼을 제거한다. 그리고 cat 명령어를 이용하여 하나의 파일로 .csv파일을 합친다. 2015년도의 데이터에서 날짜는 배열에 넣은 후 0번째에 '가 들어있으니 한 칸 당겨 데이터를 다시 저장하도록 합니다.
station은 중복되는 키오스번호를 지우고
iconv –f euc-kr -t utf-8 station1.csv > station.csv
명령어를 통해 인코딩을 해줍니다.

	A	B	C
1	번호,구별,명칭,위치,주소,거치대,좌표		
2	1,유성구,무역전시관입구(택시승강장	앞),엑스포다리	맞은편,
3	2,유성구,대전컨벤션	센터	앞,둔산대교
4	3,서구,	한밭수목원(정문입구),한밭수목원	내,
5	4,서구,초원아파트104동부근(버스정류장),초원아파트	104동앞	쪽문
6	5,서구,	둔산대공원	입구(버스정류장),한밭수목원에서
7	6,서구,백합4가	앞(농협앞),"백합아파트	상가
8	농협	버스정류장	앞",
9	7,서구,정부청사	입구(대덕대로),둔산	시외버스터미널
10	8,서구,	정부청사	입구(샘머리),둔산
11	9,서구,황실아파트앞(성룡초교	앞),"성룡초교	정문
12	버스정류장앞",	서구	월평동
13	10,서구,만년동	KBS	부근(기업은행
14	육교	건너편",	서구
15	11,서구,누리아파트앞(후문버스정류장),"누리아파트	후문과	
16	무지개아파트	사이	버스정류장
17	12,서구,	정부청사역	앞(4번
18	13,서구,	삼천중학교	앞,"수정타운
19	1동	버스정류장	앞",
20	14,서구,둔산	하이마트	앞,둔산
21	15,서구,	둔산	홈플러스
22	16,서구,	국화아파트앞(501동	앞),"국화아파트
23	버스정류장	앞",	서구
24	17,서구,타임월드	앞	,"타임월드
25	버스정류장	앞/우리은행	앞",
26	18,서구,	대전시청	앞,대전

<station 정보>



	A	B	C
1	RENT_STATION,RENT_DATE,RETURN_STATION,RETURN_DATE		
2	43,2013-01-01	05:56:03,34,2013-01-01	06:02:17
3	2,2013-01-01	06:04:06,10,2013-01-01	06:18:59
4	106,2013-01-01	10:53:05,105,2013-01-01	10:57:43
5	4,2013-01-01	11:22:23,4,2013-01-01	12:17:53
6	21,2013-01-01	11:39:53,105,2013-01-01	11:49:43
7	90,2013-01-01	12:08:33,91,2013-01-01	12:51:36
8	13,2013-01-01	13:14:29,30,2013-01-01	13:30:39
9	1,2013-01-01	13:37:42,1,2013-01-01	13:38:15
10	1,2013-01-01	13:38:13,2,2013-01-01	15:09:58
11	1,2013-01-01	13:38:47,2,2013-01-01	15:10:14
12	9,2013-01-01	13:42:53,23,2013-01-01	14:20:12
13	27,2013-01-01	13:43:28,27,2013-01-01	13:43:56
14	30,2013-01-01	13:48:50,7,2013-01-01	14:56:51
15	30,2013-01-01	13:49:07,30,2013-01-01	13:55:26
16	30,2013-01-01	13:49:09,18,2013-01-01	14:09:29
17	30,2013-01-01	13:49:22,8,2013-01-01	14:17:46
18	30,2013-01-01	13:49:27,30,2013-01-01	13:55:17
19	30,2013-01-01	13:49:41,30,2013-01-01	13:55:08
20	30,2013-01-01	13:49:47,18,2013-01-01	14:09:03
21	30,2013-01-01	13:51:00,29,2013-01-01	15:03:35
22	30,2013-01-01	13:51:20,29,2013-01-01	15:03:25
23	46,2013-01-01	14:32:57,18,2013-01-01	14:42:03
24	19,2013-01-01	14:34:29,19,2013-01-01	14:37:40
25	19,2013-01-01	14:34:57,19,2013-01-01	14:37:48
26	47,2013-01-01	14:43:03,47,2013-01-01	14:44:30

<정리된 Tashu 데이터>

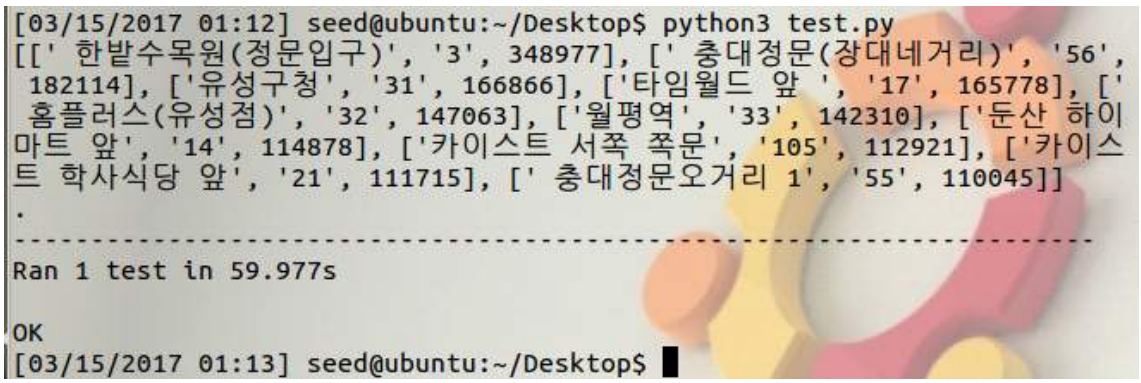
2.과제문제 해결과정

1)가장 인기 있는 정류장 Top_10

우선 tashu 데이터와 station데이터가 들어있는 파일을 읽어 들입니다. 그리고 matrix와 place라는 1차원 배열을 여유 있게 250의 길이로 만든 후, matrix에는 RENT_STATION과 RETURN_STATION 의 STATION 번호에 맞는 matrix의 좌표에 각각 +1씩 누적을 시켜줍니다. 그 후 place 배열에는 station 번호에 맞는 place의 좌표에 station 데이터의 명칭을 넣어 줍니다.

그리고 MAX값을 찾는 search를 통해 가장 큰 값의

matrix 배열의 원소 값과 위치, 명칭을 result에 append 해주고, 가장 큰 값의 matrix 원소 값은 0으로 만들어 줍니다.



<결과 화면>

2)가장 인기 있는 경로 Top_10

가장 인기 있는 정류장을 구하는 코드를 구현하려고 했으나 구현하지 못하였습니다. 하지만 구하려고 했던 방법으로는 위와 비슷하게 matrix[[rent,return]]와 같은 식으로 해당하는 배열에 +1씩 하여 누적을 시키고, rent,return station의 값을 이용하여 place 배열에서 station의 명칭을 가지고 오려고 하였습니다. 하지만 index out of range에 걸려 해결이 되지 않아 완성을 시키지 못하였습니다. 다음에는 파이썬 문법에 대해 조금 더 공부하여 열심히 하도록 하겠습니다.