

6주차 정규세션

To Big's 10기

강 인 구 정 윤 호



Algorithm

알고리즘은 데이터 전처리에 반드시 필요하다!

'빅데이터' 동아리 TOBIG'S

진짜로 **빅** 데이터를 다뤄본 적이 있나요?

'빅데이터' 동아리 TOBIG'S

 groupcall.csv	2019-02-17 오전...	Microsoft Excel ...	2,319,206KB
 groupcall.ipynb	유형: Microsoft Excel 싼표로 구분된 값 파일 PYNB 파일 크기: 2.21GB 수정한 날짜: 2019-02-17 오전 12:01		10KB

아래와 같이 수많은 사람들이 가족과 통화하거나 팀프로젝트를 하는 등 그룹 통화를 한 내역이 있습니다.

	call_start_day	call_start_time	call_end_day	call_end_time	hashed
0	1	09:14:58.558	1	09:41:30.200	967393e81d99ce8e577ee130b7ce8e4fd45e3e9cecb560...
1	17	11:05:05.176	17	13:07:42.515	02181a0c962f34f019bc9d5b582fb0ec79b1441f96aa4d...
2	20	02:18:43.172	20	02:28:58.177	86022904c5cf72a54978479c94041f4256d6c3c2a1f71c...
3	22	자정으로부터 절대 시간		92	aafb40d212fe18ff4eafb82fdcf3b53f2161cb3ce59de4...
4	26	06:29:21.182	26	06:50:55.004	c87c2fad141edf323f3787335b54be22945a02fe052448...
5	36	09:12:30.447	36	09:31:51.871	f31a5e8feee0aedfa66378cc35f1663623634563f2d977...
6	집계일로부터 통화 시작/종료 날짜			05.309	52f4a6a555803e8b239e8b69288d4787d39dd40c2a126e...
7				54.300	d4843247de5b8a0f34d04b418b55bbde84fe7d31dc2192...
8	50	09:13:55.044	50	09:16:06.351	e2e030d3c933fde97b5484aad91969aa5479540f5b27b0...
9	54	01:52:02.512	54	09:28:25.278	f16d8b891f4d1d52f4c298d20c4a4fb4e63fd9d024629e...

그런데 사람들의 정보가 있는 hashed의 0번 데이터를 보니 아래와 같이 암호화 되어있어 갖고 있는 정보량이 쓸데없이 많습니다.

hashed[0] = 967393e81d99ce8e577ee130b7ce8e4fd45e3e9cecb560de427ede6ea49e024f,
a0b6ecbec654b18fe36ebe6230e25a653fb12125733583d012741572134447f4,
3193ab18168bcadbcb8342c06c4a35fa0d6e58d9619fe805fb811fc4e6562fef

3명!

각 사람들이 집계일로부터 몇 회 통화했는지 등 다양한 정보를 얻고 싶다면 각 hash에 대해서 정리를 해야 합니다. 어떻게 하면 될까요?

```
In [4]: df['hashed'].apply(lambda x: x.split(','))
```

```
MemoryError                                Traceback (most recent call last)
```

```
<ipython-input-4-bde400bb9e44> in <module>()
```

```
----> 1 df['hashed'].apply(lambda x: x.split(','))
```

```
C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python36_64\lib\site-packages\
```

```
3192         else:
```

```
3193             values = self.astype(object).values
```

```
-> 3194             mapped = lib.map_infer(values, f, convert=convert_dtype)
```

```
3195
```

```
3196         if len(mapped) and isinstance(mapped[0], Series):
```

```
pandas/_libs/src/inference.pyx in pandas._libs.lib.map_infer()
```

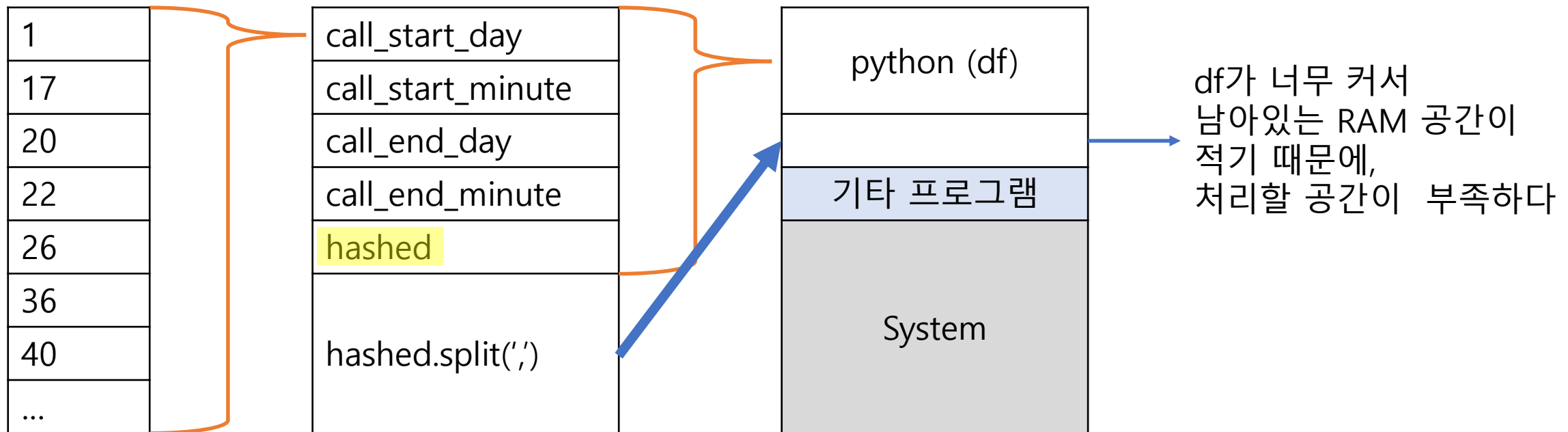
```
<ipython-input-4-bde400bb9e44> in <lambda>(x)
```

```
----> 1 df['hashed'].apply(lambda x: x.split(','))
```

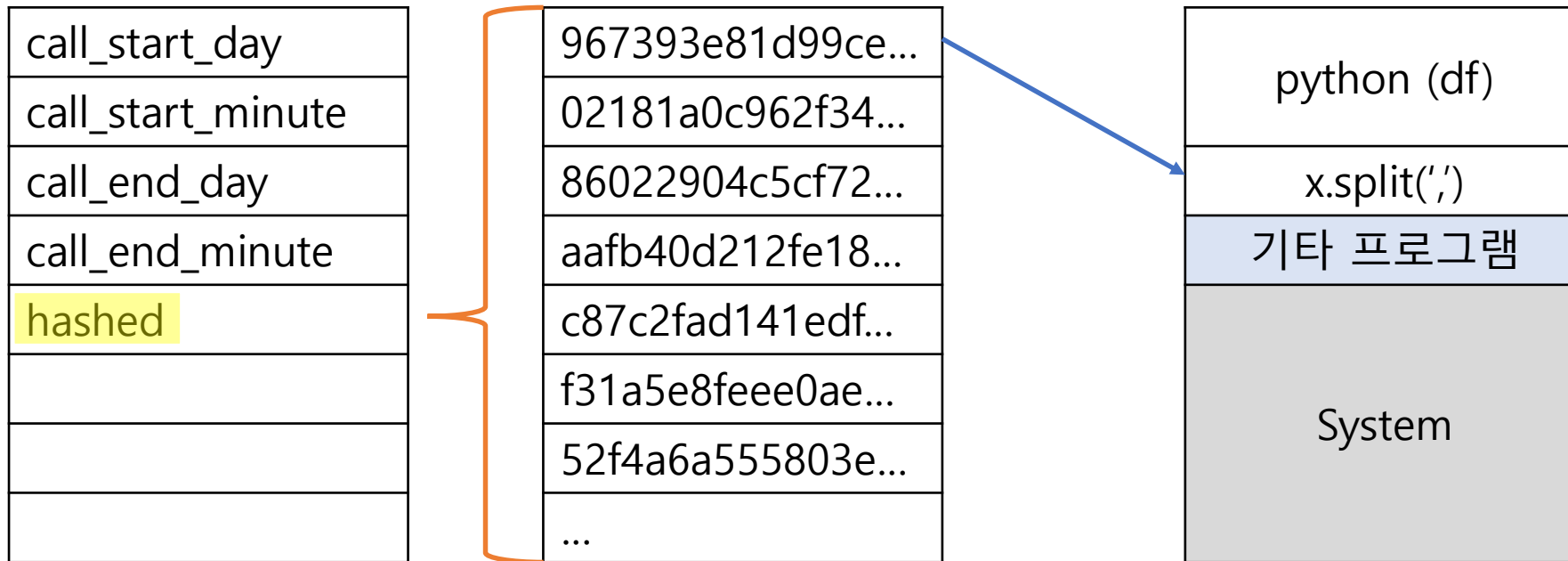
```
MemoryError:
```

split만 했는데 오류가 발생한다

메모리 오류가 뭐지?



한번에 전체를 메모리에 참조 -> 비효율적
한번에 하나의 데이터만 하나의 공간에 가져와 처리 -> 효율적



한번에 전체를 메모리에 참조 -> 비효율적

한번에 하나의 데이터만 하나의 공간에 가져와 처리 -> 효율적

Python Generator

Generator Iterator를 생성해주는 함수

Iterator next()함수를 통해 순차적으로 값을 가져오는 object

yield 함수가 멈추는 지점, 즉시 next()를 호출한 곳으로 값을 반환함
일반 함수와 다르게 지역변수가 휘발되지 않고 그 상태로 유지됨

동시성 / 병렬성 프로그래밍

Multithreading Multiprocessing

한 코어에 여러 개의 작업

쓰레드간 메모리(코드, 데이터, 힙) 공유

계산보다 I/O가 훨씬 중요한 작업에만 유용

파이썬은 **GIL** 때문에 오히려 계산 성능 저하

한 코어당 하나의 작업

코어간 메모리 공유 X (코어간 통신 필요)

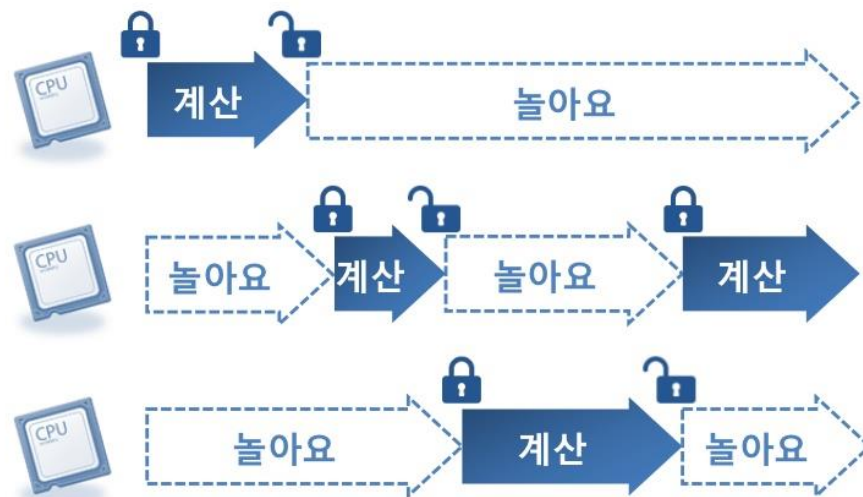
계산, I/O 모두 고효율

우리가 바란 것



GIL때문에 일어난 일

자물쇠는 하나뿐 → 한 CPU만 일(계산)한다!



Multiprocessing을 하자!
(array 연산 = 계산 \neq I/O)

실습

과제

	call_start_day	call_start_time	call_end_day	call_end_time	hashed
0	1	09:14:58.558	1	09:41:30.200	967393e81d99ce8e577ee130b7ce8e4fd45e3e9cecb560...
1	17	11:05:05.176	17	13:07:42.515	02181a0c962f34f019bc9d5b582fb0ec79b1441f96aa4d...
2	20	02:18:43.172	20	02:28:58.177	86022904c5cf72a54978479c94041f4256d6c3c2a1f71c...
3	22	09:22:01.936	22	09:47:40.192	aafb40d212fe18ff4eafb82fdcf3b53f2161cb3ce59de4...
4	26	06:29:21.182	26	06:50:55.004	c87c2fad141edf323f3787335b54be22945a02fe052448...
5	36	09:12:30.447	36	09:31:51.871	f31a5e8feee0aedfa66378cc35f1663623634563f2d977...
6	40	10:58:28.822	40	11:01:05.309	52f4a6a555803e8b239e8b69288d4787d39dd40c2a126e...
7	47	16:30:21.582	47	16:44:54.300	d4843247de5b8a0f34d04b418b55bbde84fe7d31dc2192...
8	50	09:13:55.044	50	09:16:06.351	e2e030d3c933fde97b5484aad91969aa5479540f5b27b0...
9	54	01:52:02.512	54	09:28:25.278	f16d8b891f4d1d52f4c298d20c4a4fb4e63fd9d024629e...

group_call_HMS.csv 를 이용해서 각 사람별로 파생변수들 만들기

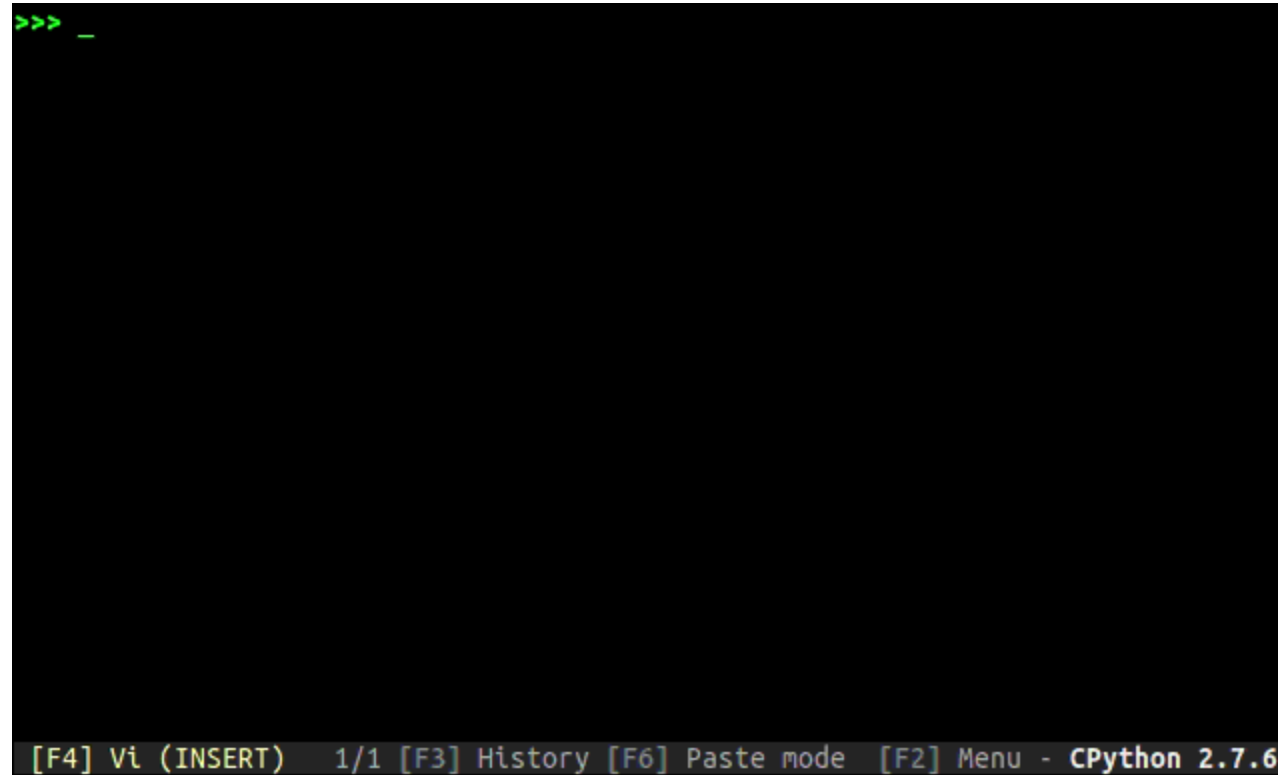
Ex) 총 통화 횟수, 총 통화 시간, 평균 통화 시간,
총 통화한 인원수, 1회 통화에 참여한 평균 인원수
인맥지도 만들기 등



단순 for문 -> 20시간
얼마나 오래 걸릴지도 모르고 돌아가고 있는지도 모르겠어요 $\pi\pi$

tqdm

```
>>> _
```



[F4] Vi (INSERT) 1/1 [F3] History [F6] Paste mode [F2] Menu - CPython 2.7.6

Reference

- Iterator 설명 - <https://kkamikoon.tistory.com/91>
- Yield 설명 - <https://kkamikoon.tistory.com/90>
- GIL 설명
 - <https://wangan9.tistory.com/entry/pythonthreadGIL>
 - <http://highthroughput.org/wp/cb-1136/>
- tqdm - <https://pypi.org/project/tqdm/>

Q & A

들어주셔서 감사합니다.