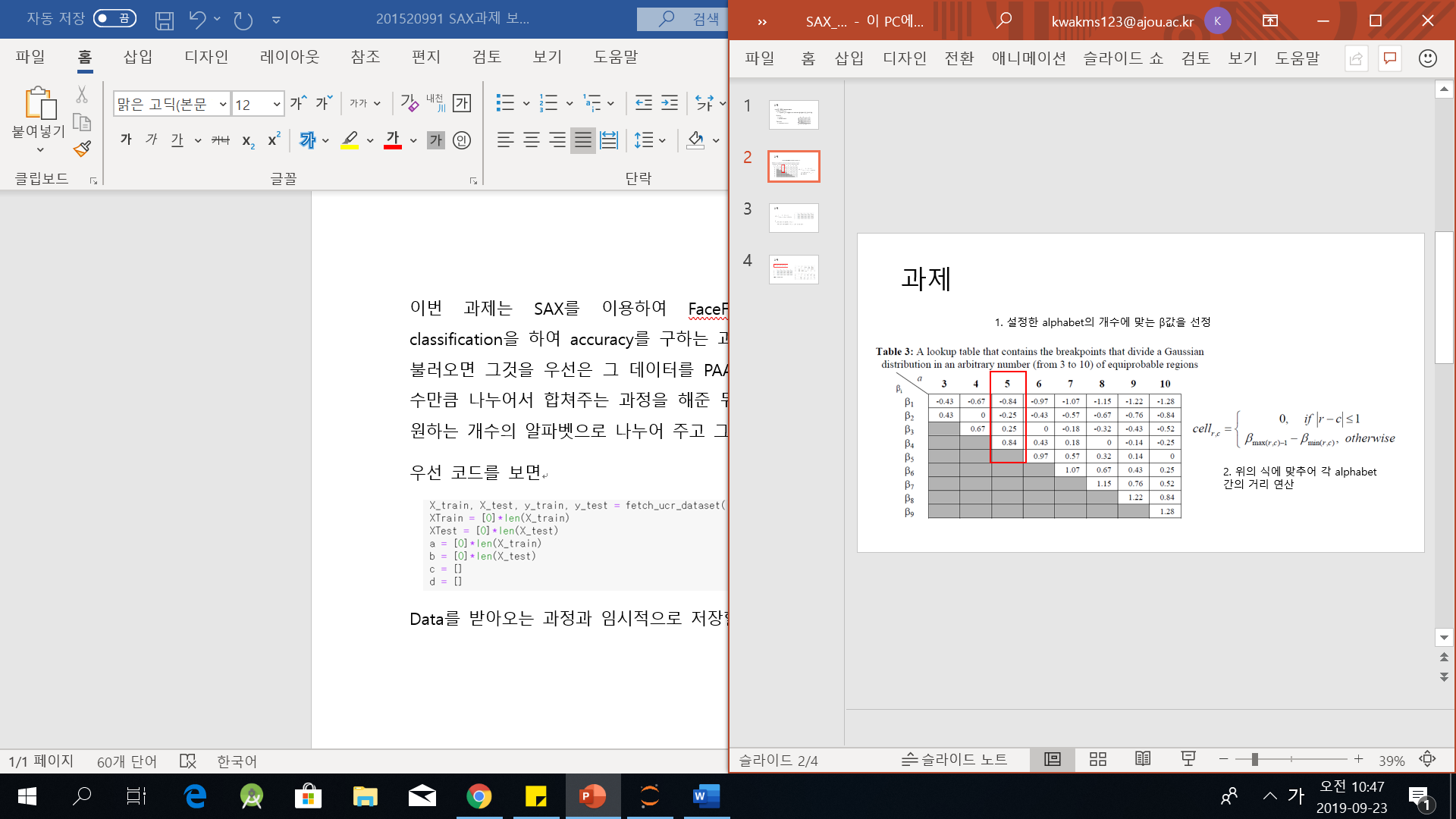
SAX를 이용한 classification

201520991

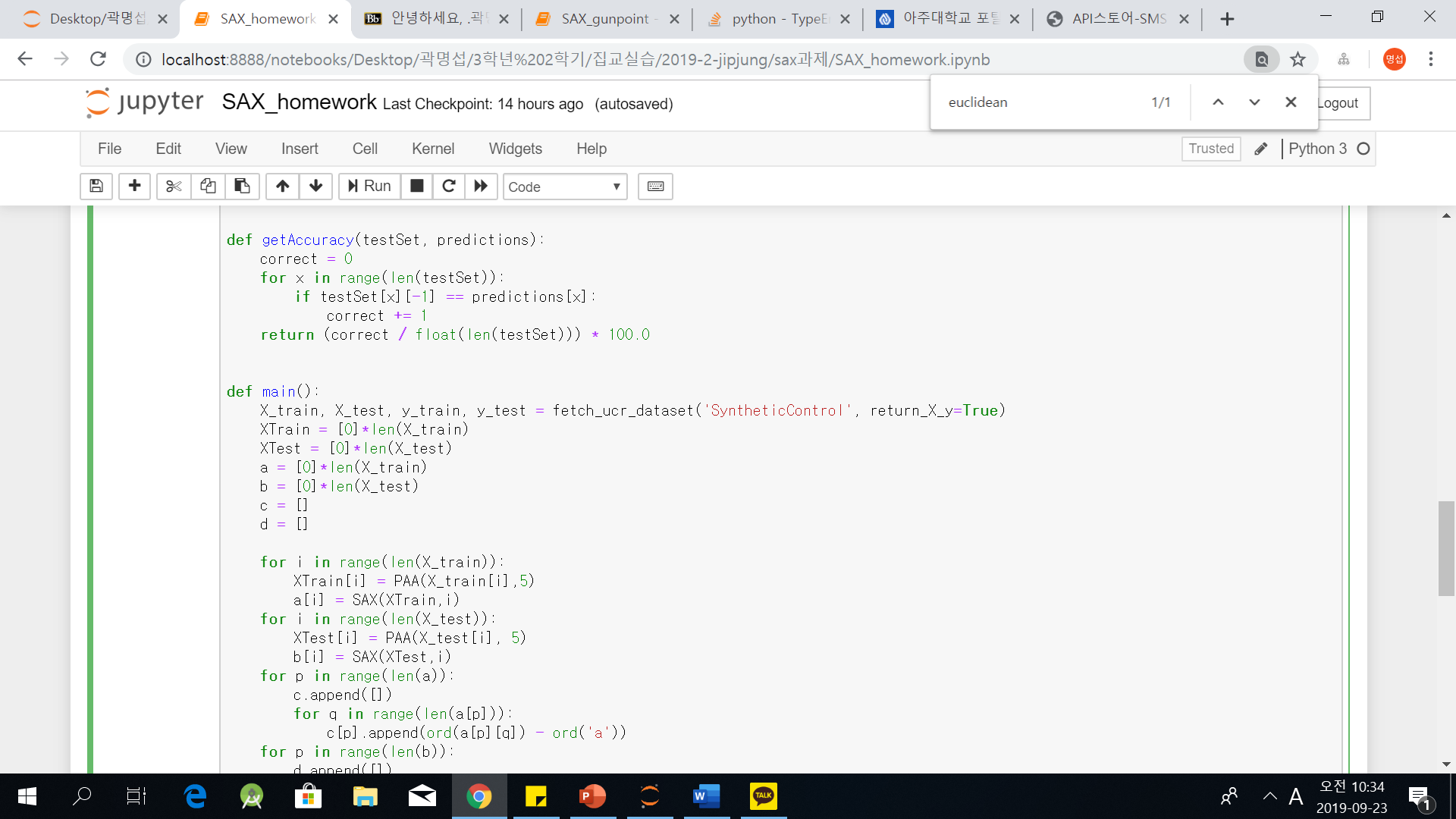
곽명섭

이번 과제는 SAX를 이용하여 FaceFour과 SyntheticControl data set의 classification을 하여 accuracy를 구하는 과제였다. 대략적인 메커니즘은 데이터를 불러오면 그것을 우선은 그 데이터를 PAA를 통해서 차원축소를 하여 원하는 개수만큼 나누어서 합쳐주는 과정을 해준 뒤에 그것을 SAX를 통해 a,b,c,d,e등으로 원하는 개수의 알파벳으로 나누어 주고 그것의 distance를 구하기위해서

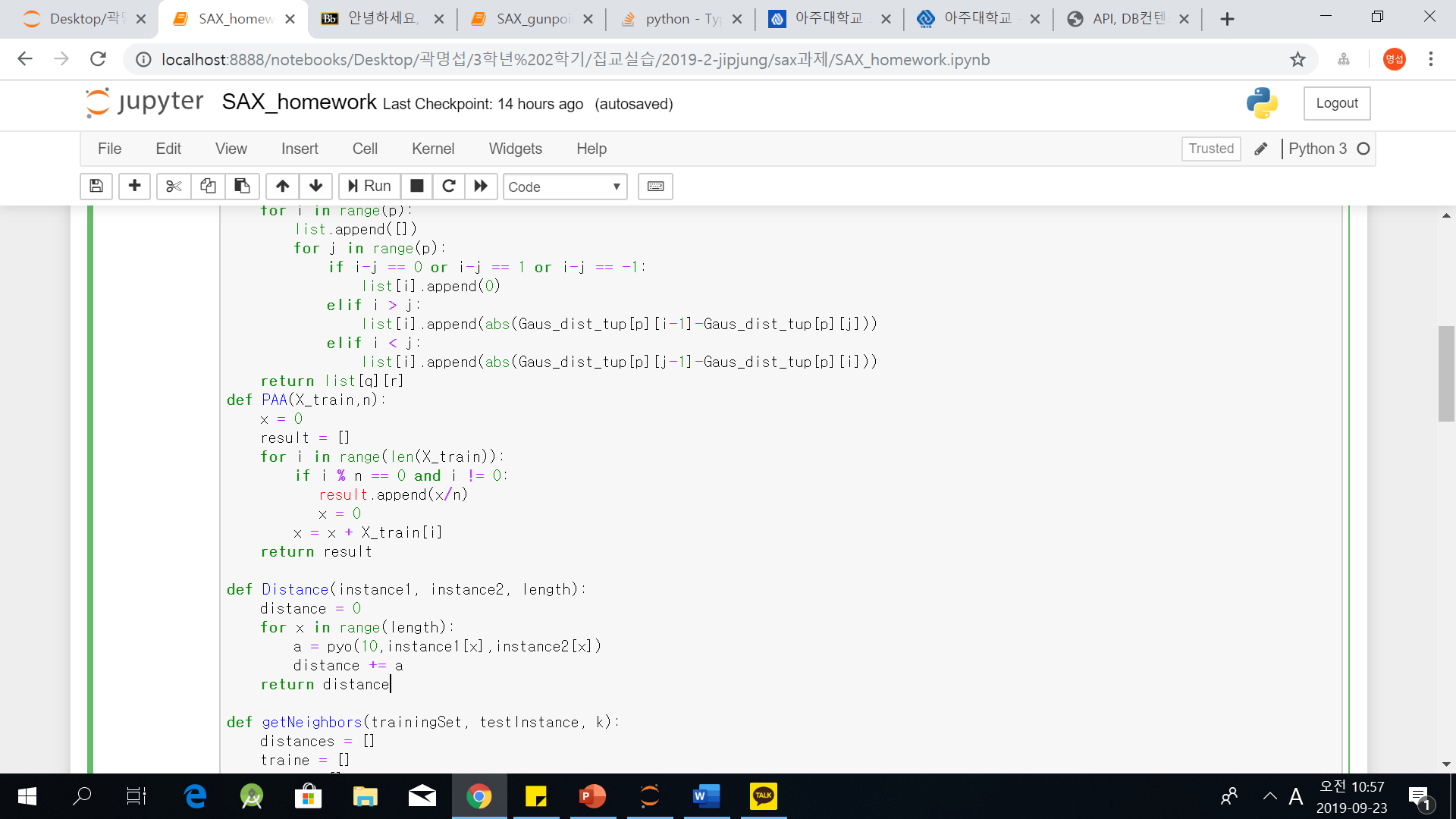


이런 식으로 거리 연산을 위해 표를 만들어 준 후에 그 표에 맞는 값을 distance로 설정을 해준 뒤 accuracy를 계산하는 방식이다.

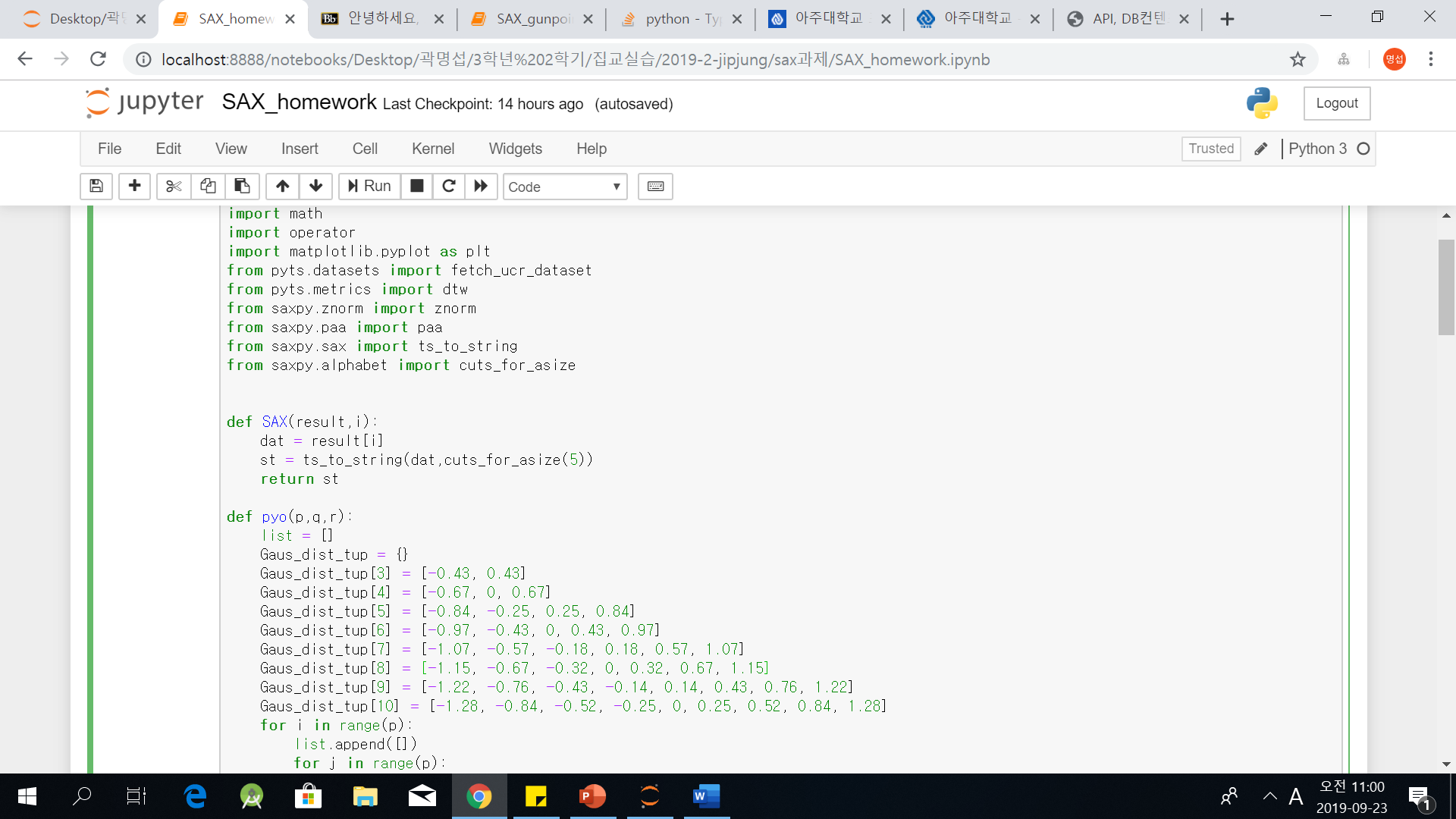
우선 코드를 보면



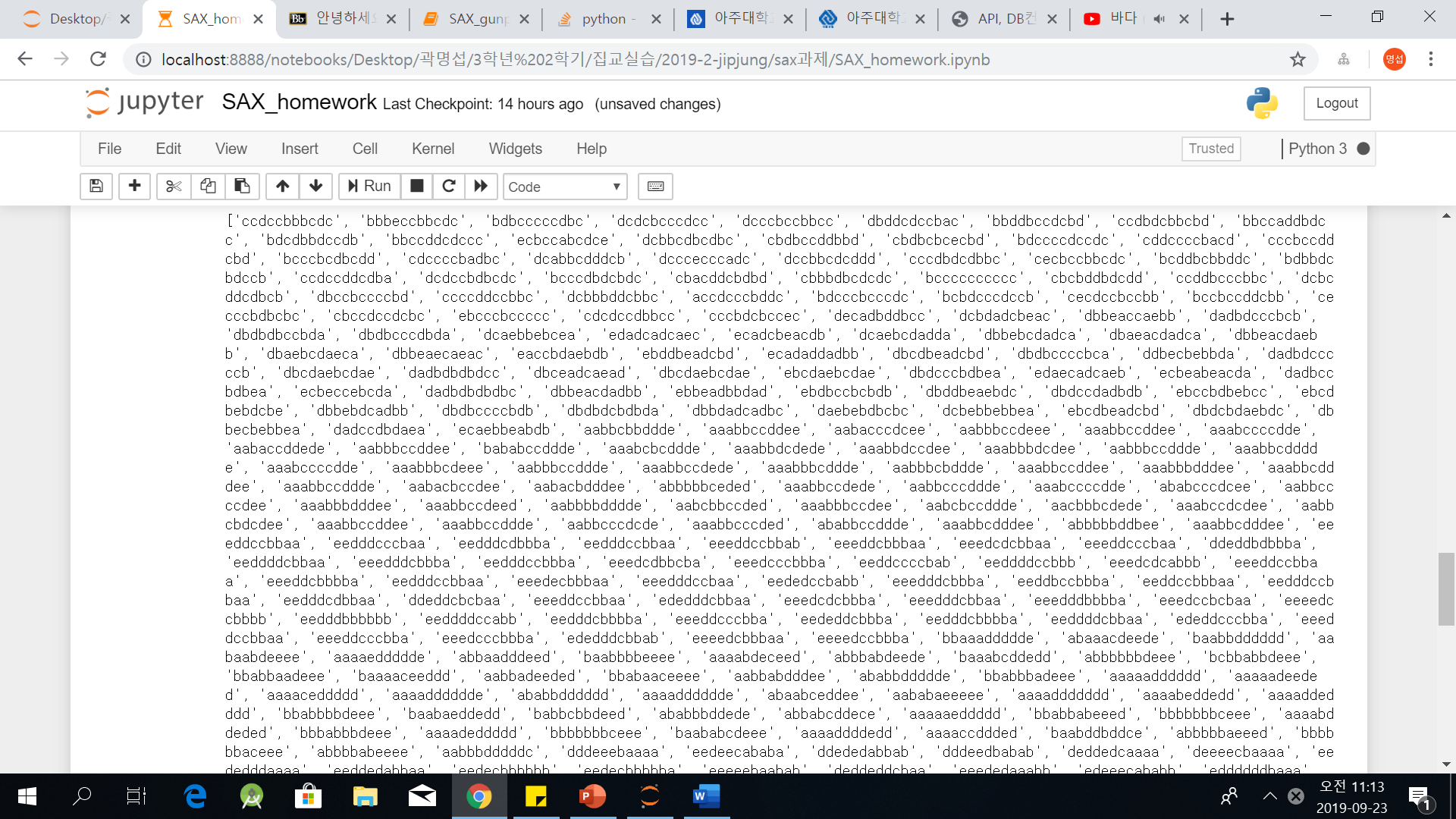
Data를 받아오는 과정과 임시적으로 저장할 list를 불러오는 과정이다. 여기서 SyntheticControl이나 FaceFour data를 불러온다. 그러고나서 training set과 test set을 분리해서 저장을 해주고 나중에 사용할 임시적인 저장용 list들을 선언한다.



Data를 불러오고 나면 그 이후에는 바로 trainingset과 dataset을 PAA를 통해서 차원축소를 해주는 과정이다. 몇 개로 해줄지는 n이라는 parameter로 받아 준다.

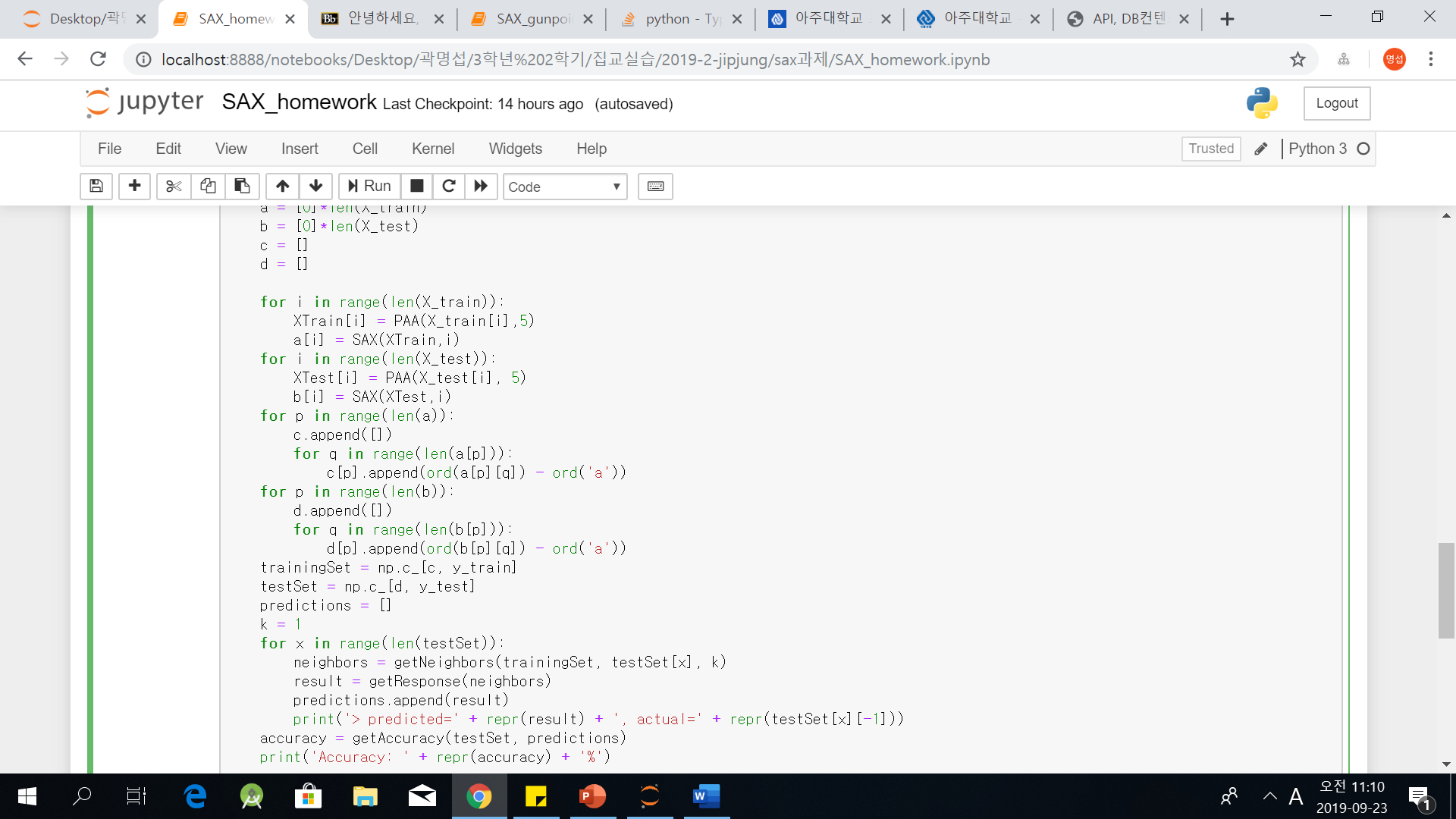


그리고 나면 PAA까지 하여 축소된 dataset을 또 한번 SAX를 통해서 a,b,c,d,e등으로 나타내어 주는 과정을 나타낸다. 이 과정에서는 그냥 SAX module을 사용하였다. cuts\_for\_asize(--)이 안에 넣어주는 parameter에 따라서 알파벳의 개수가 결정이 된다.

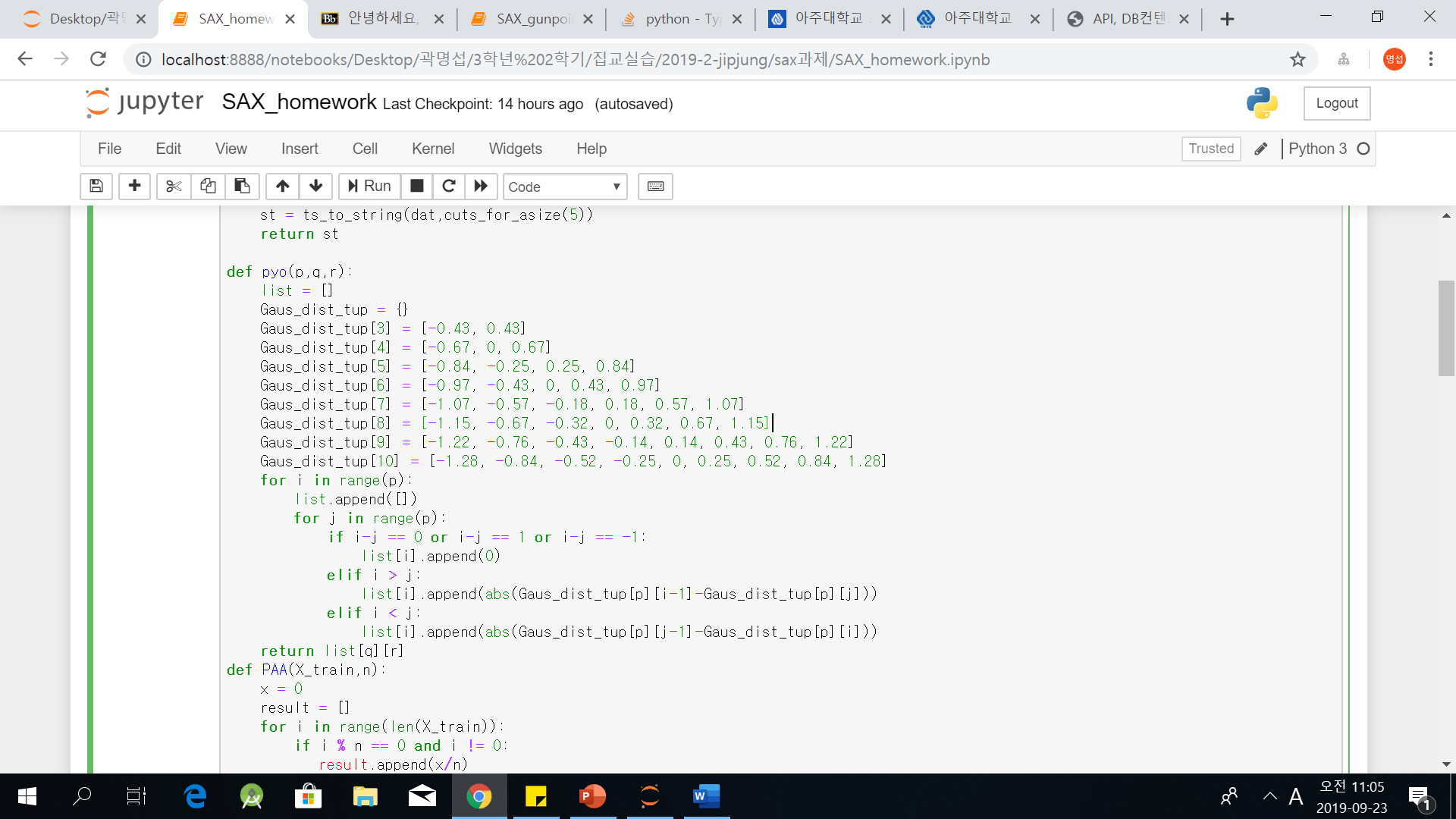


SAX를 해주고 난 후의 자료형이 어떻게 저장되었는지를 보기위해서 print로 나타낸 것이다.

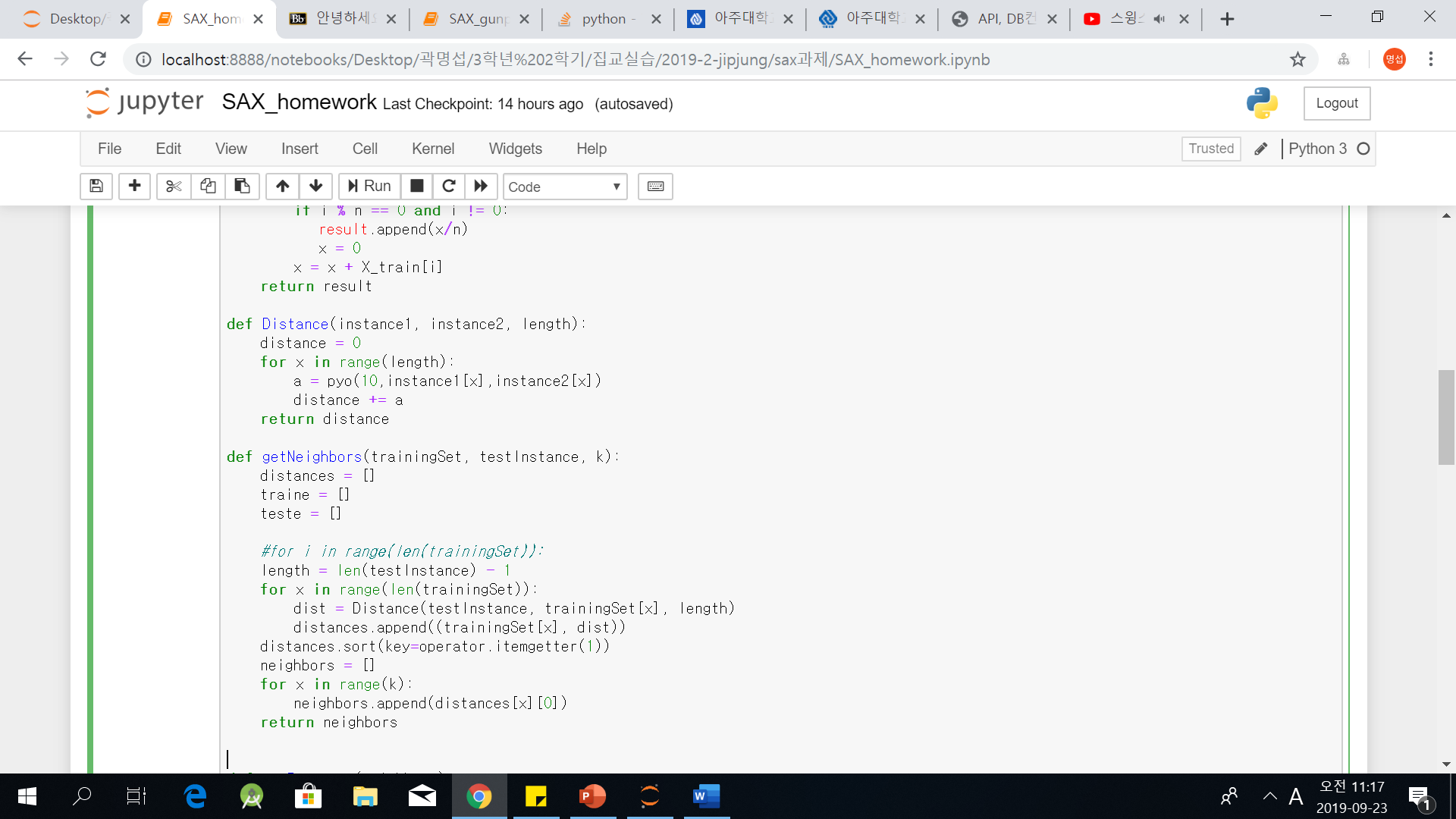
SAX로 알파벳으로 만들어 주고나면 각 알파벳간의 거리를 gaussian distance를 통해서 구해주어야 한다. 그런데 자료형이 char형이다보니 그냥 gaussian distance를 구해줄 방법이 딱히 생각이 나지 않아서 나는 a -> 0 b -> 1 c -> 2와 같은식으로 ascii코드를 이용하여 치환을 우선 하기로 마음을 먹었다.



이런 식으로 ord()함수를 사용하여 알파벳들을 전부 치환을 해주었다.

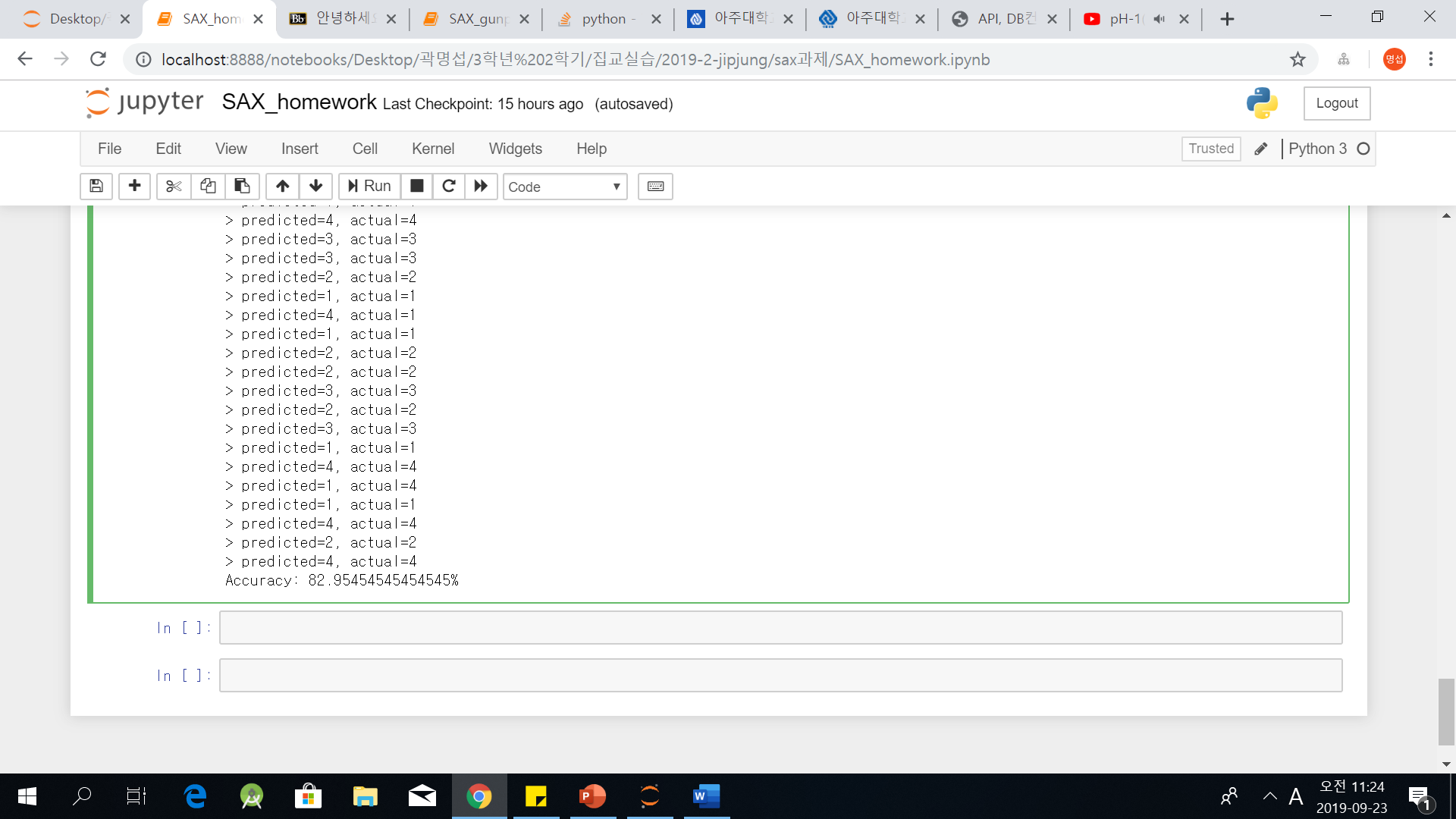


숫자로 치환된 데이터를 이제 gaussian distance로 나타내기 위해 표를 만들어준 함수이다.

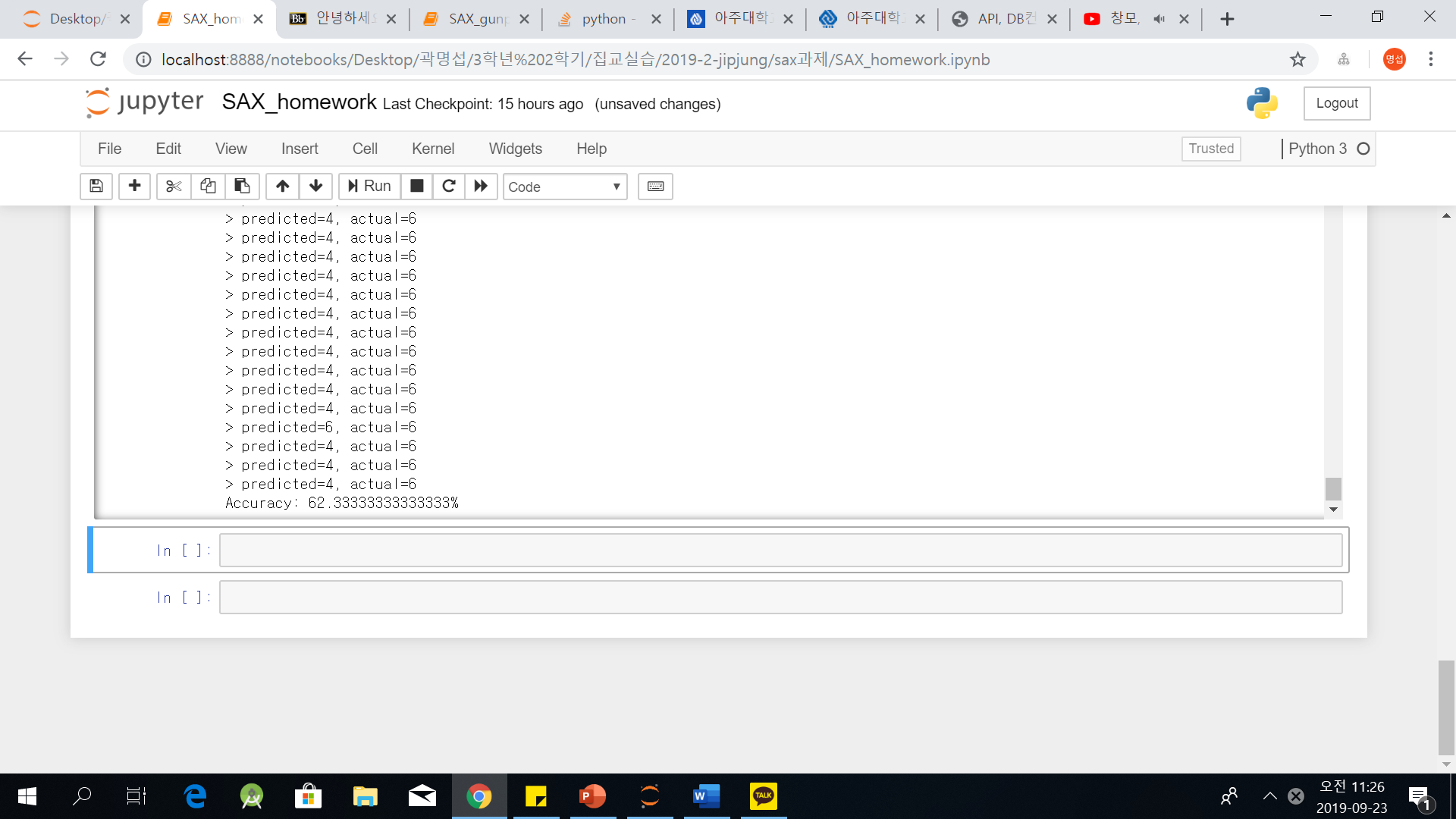


이제 여기서 데이터들을 gaussian distance를 구해주는 과정이다.

이제 결과 값들을 살펴보면 우선 PAA의 단위는 5, SAX의 알파벳 개수 = 5, gaussian distance 표도 5로 설정해준 결과 값을 보면

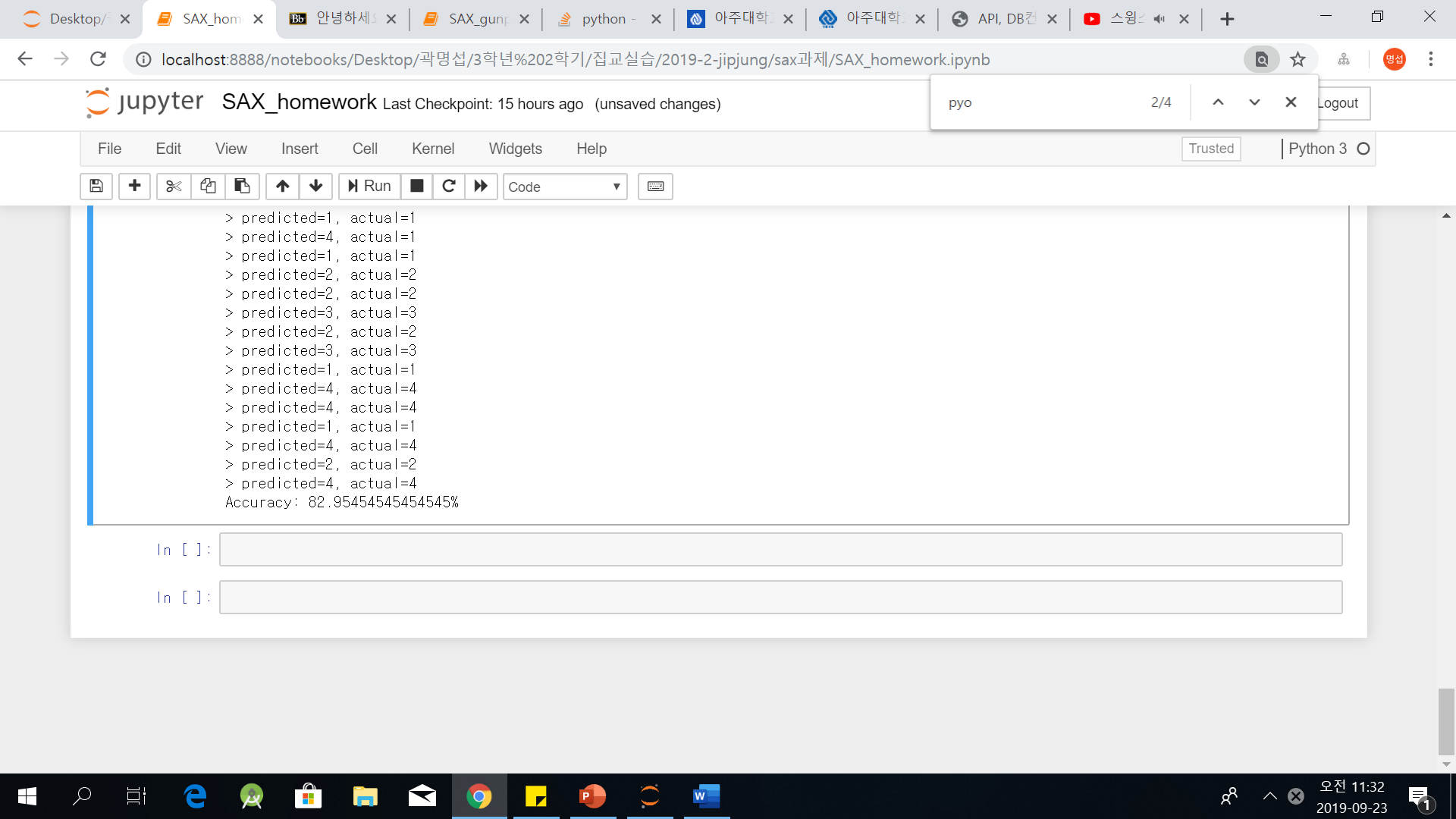


facefour기준으로 83%정도의 accuracy가 나오는 것을 알 수 있었다.

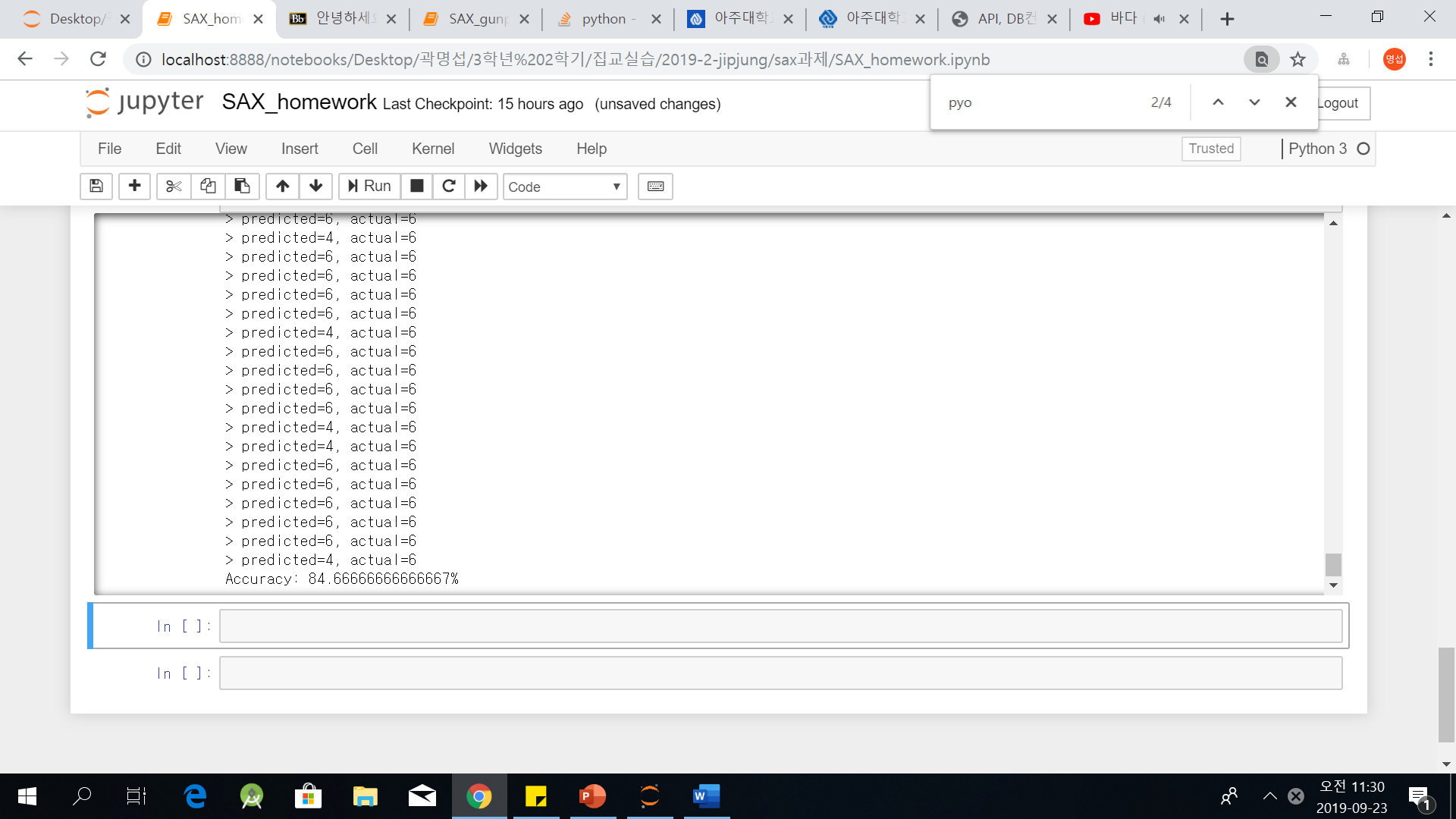


SyntheticControl기준으로 accuracy는 62.3%정도가 나왔다.

SyntheticControl이 특히 생각 한 것 보다 accuracy가 낮게 나오는 것 같아 SAX의 알파벳 개수를 8로 늘리고 GAUSSIAN DISTANCE표도 8로 바꾸어본 결과

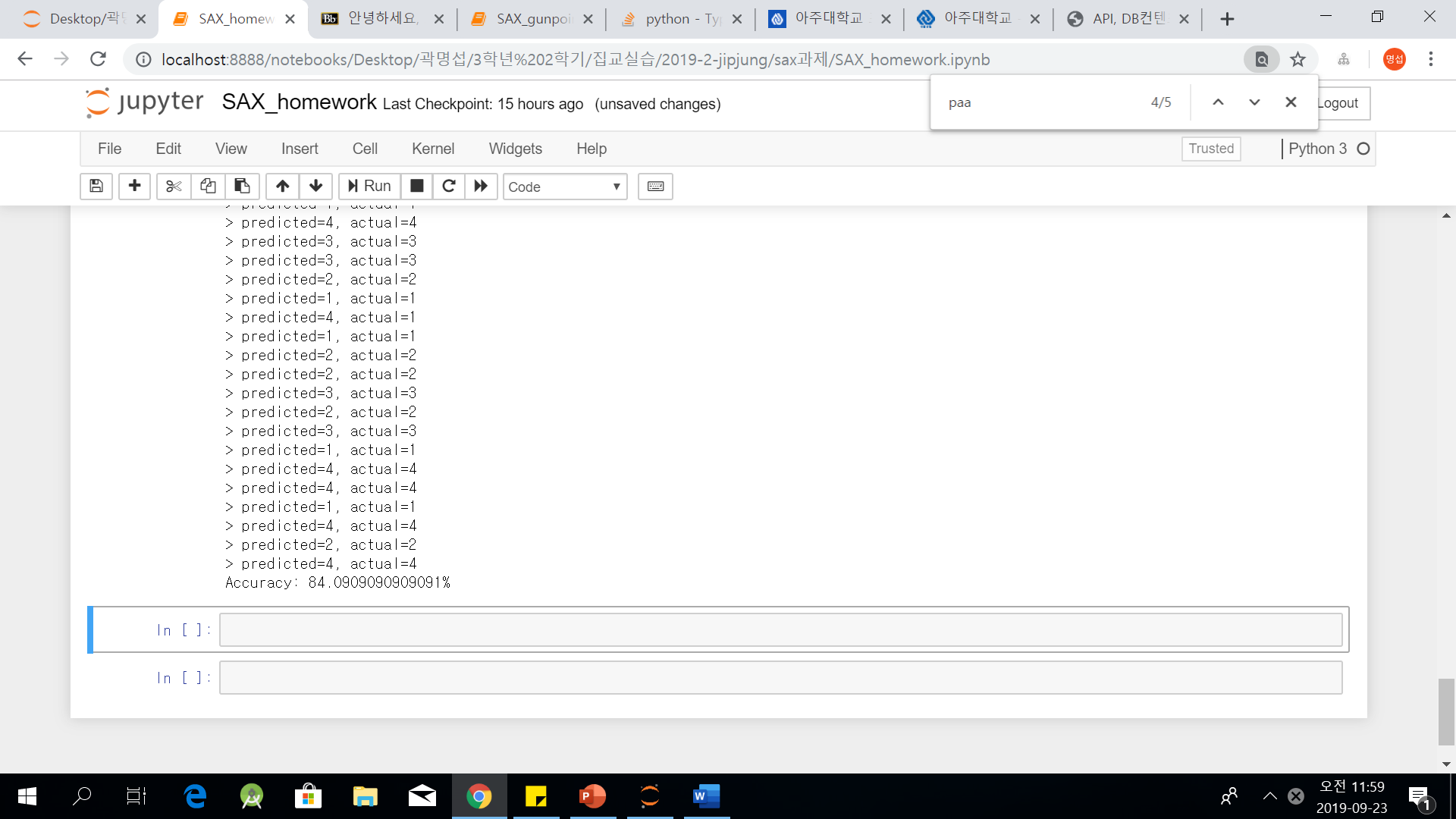


이상하게도 facefour은 accuracy에 변화가 없이 거의 accuracy가 그대로였다.

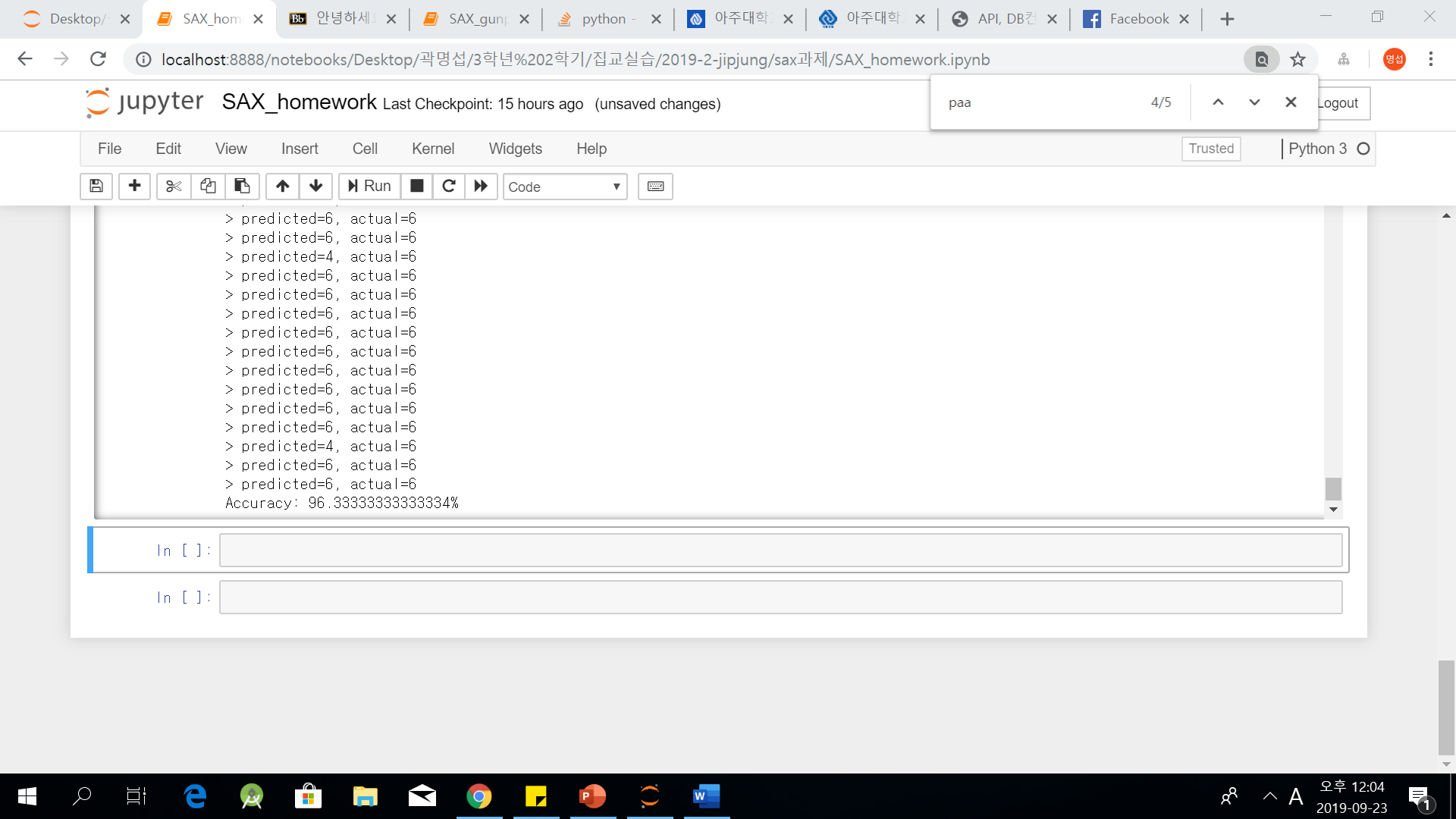


하지만 syntheticcontrol에서는 확연한 변화가 있었는데 62.3%->84%로 accuracy가 크게 오른 것을 볼 수 있었다.

이번엔 여기서 더 accuracy를 올려보고 싶다는 생각에 PAA단위를 5->3으로 줄여보았다.

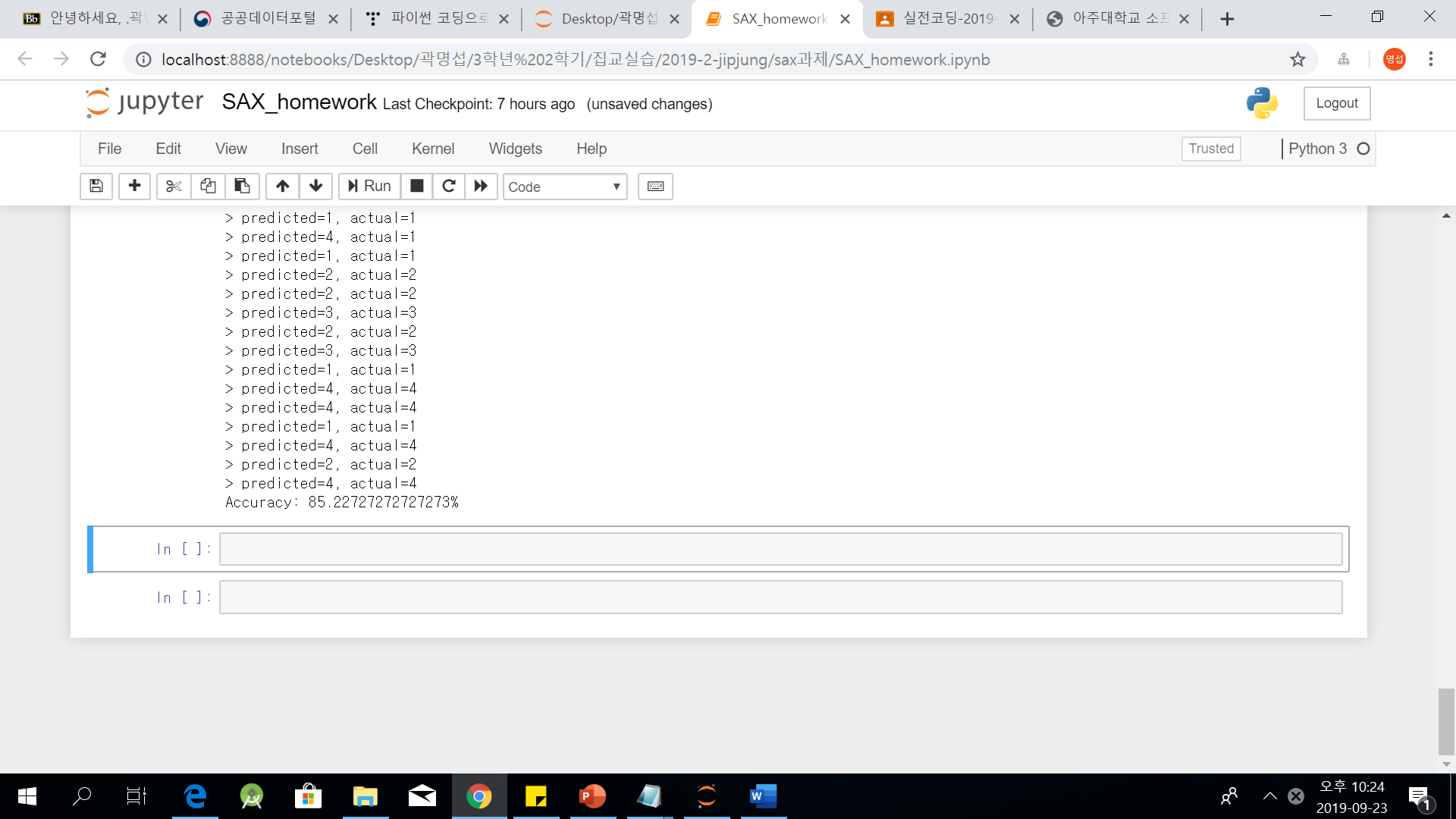


facefour같은경우 살짝 오른 수치인 84%정도가 나오는 것을 볼수있었다

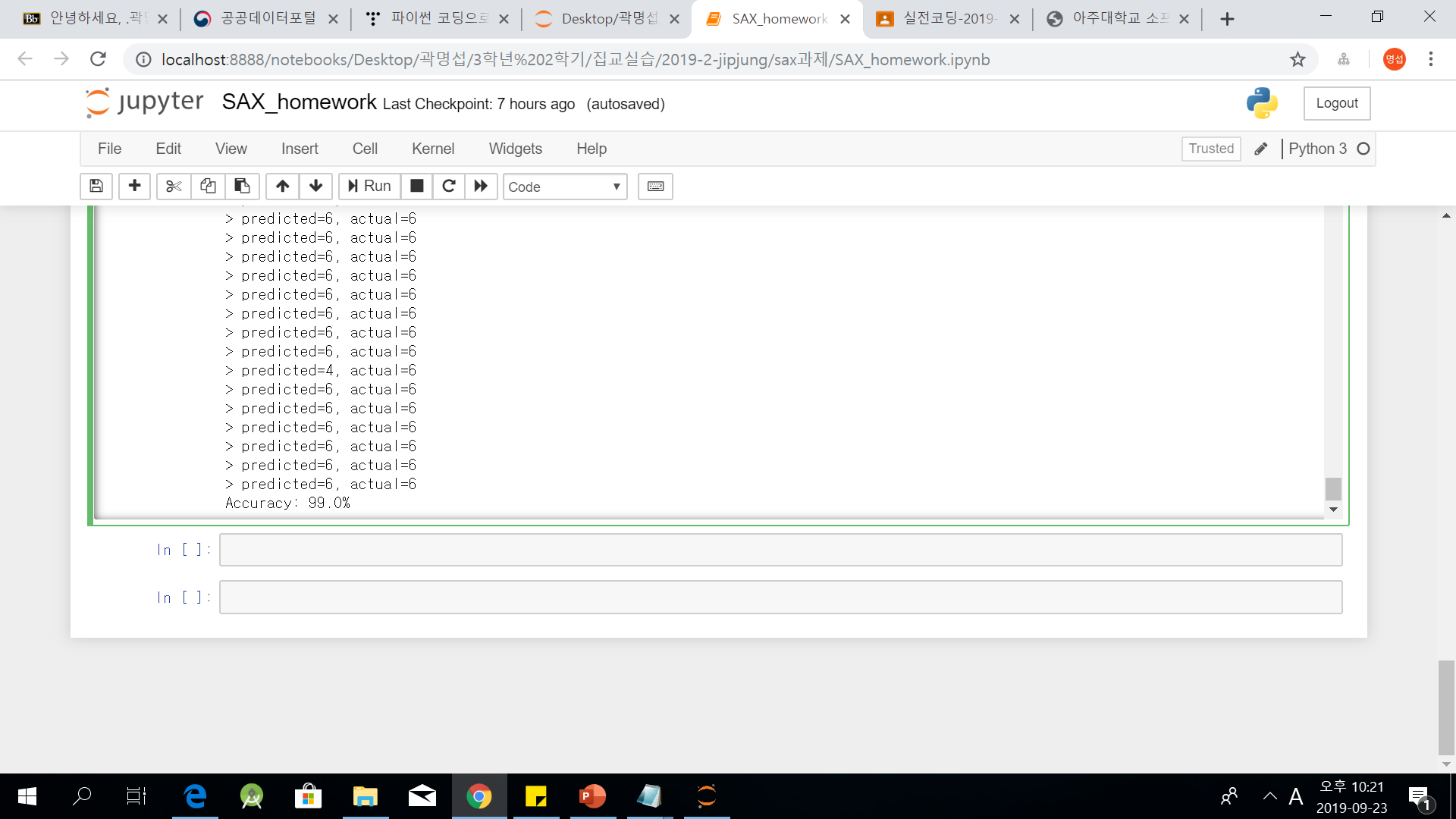


Synthetic control같은경우에 확연히 오른 수치인 96%까지 오른 것을 확인할 수 있었다.

그리고 한번 더 정말 최대한으로 끌어올려보고자 하는 마음에 SAX의 알파벳의 개수를 10개까지 한번 늘려보았는데



FaceFour같은경우 accuracy가 85%까지 오른 것을 확인 해볼수 있었으며



Synthetic control은 accuracy가 99%까지 오른 것을 확인해 볼 수가 있었다.

이 이상으로 sax의 알파벳개수를 늘리거나 하는 경우엔 오히려 accuracy는 더 줄어들 뿐이었다.