**Data Science**

**Assignment1 : Apriori Algorithm**

2016025532 컴퓨터전공 심수정

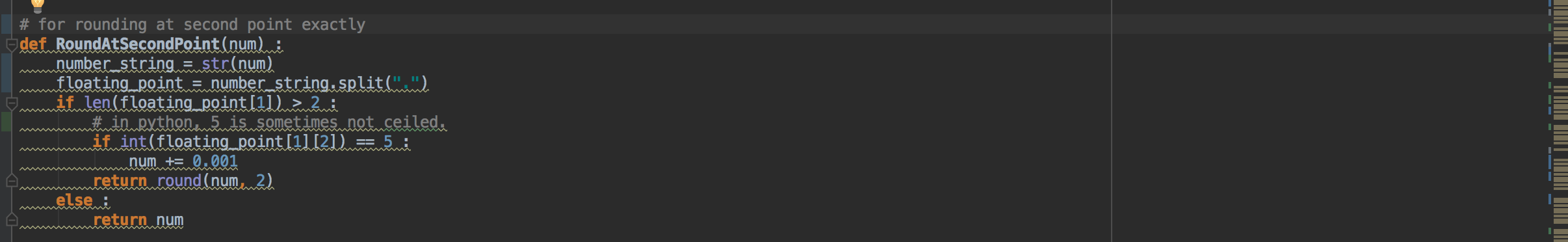
**1. Summary of Algorithm**

Apriori 알고리즘은 기본적으로는 frequent한 pattern을 생성하는 알고리즘입니다. 이 알고리즘은 downward closure를 이용합니다. 우선 downward closure은 frequent itemset의 subset은 반드시 frequent해야 한다는 것을 의미합니다. 이를 활용하여 apriori 알고리즘은 frequent 하지 않은 itemset의 supserset을 생성, test하지 않는 방식으로 진행됩니다. 이를 통해서 많은 candidate를 줄입니다.

**2. Detailed Description of Code**



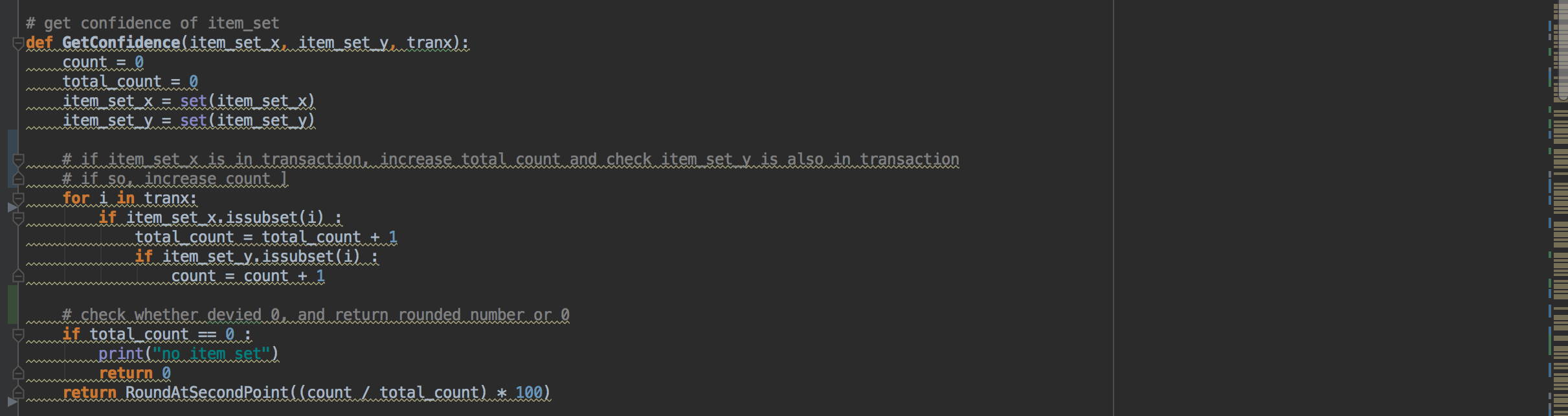
실행할 때 받은 argument를 각각 min\_support, input\_filename, output\_filename에 넣어주고, input file을 열어 데이터를 읽어들인 후 input file을 닫아줍니다.



python에서는 5를 올리지 않는 경우가 있기에 셋째자리가 5인 경우 셋째자리를 6으로 만들어 준 후 반올림해줍니다.



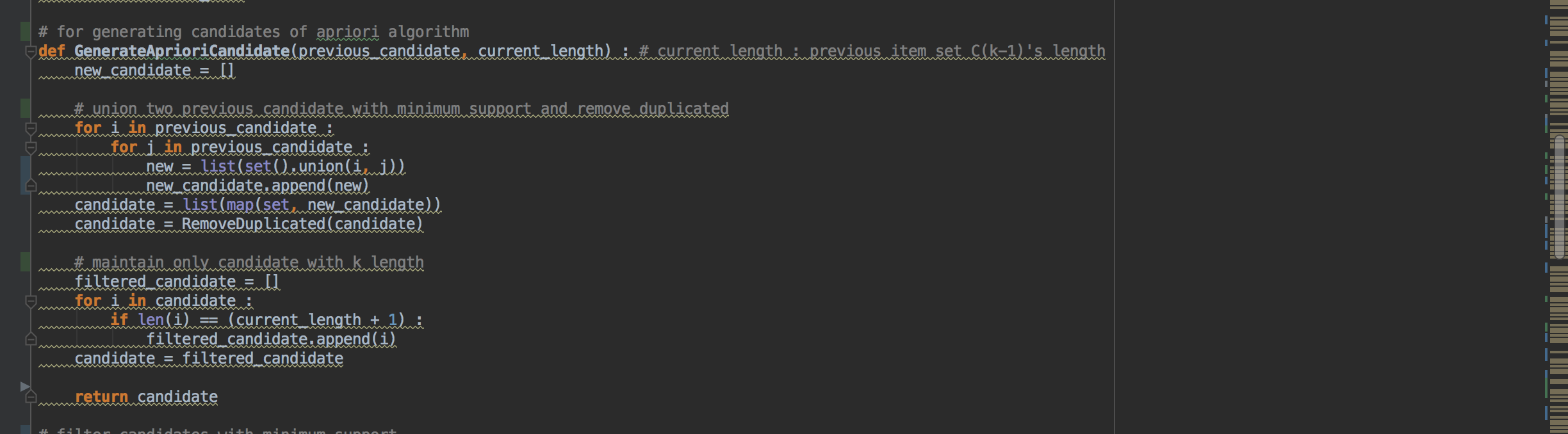
item set, transaction, transaction length를 받아 support를 구합니다.



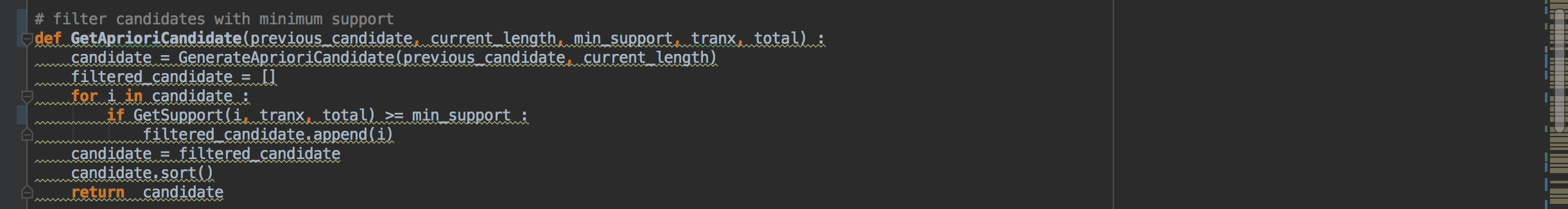
item\_set\_x, item\_set\_y, transaction list를 받아 confidence를 구합니다.



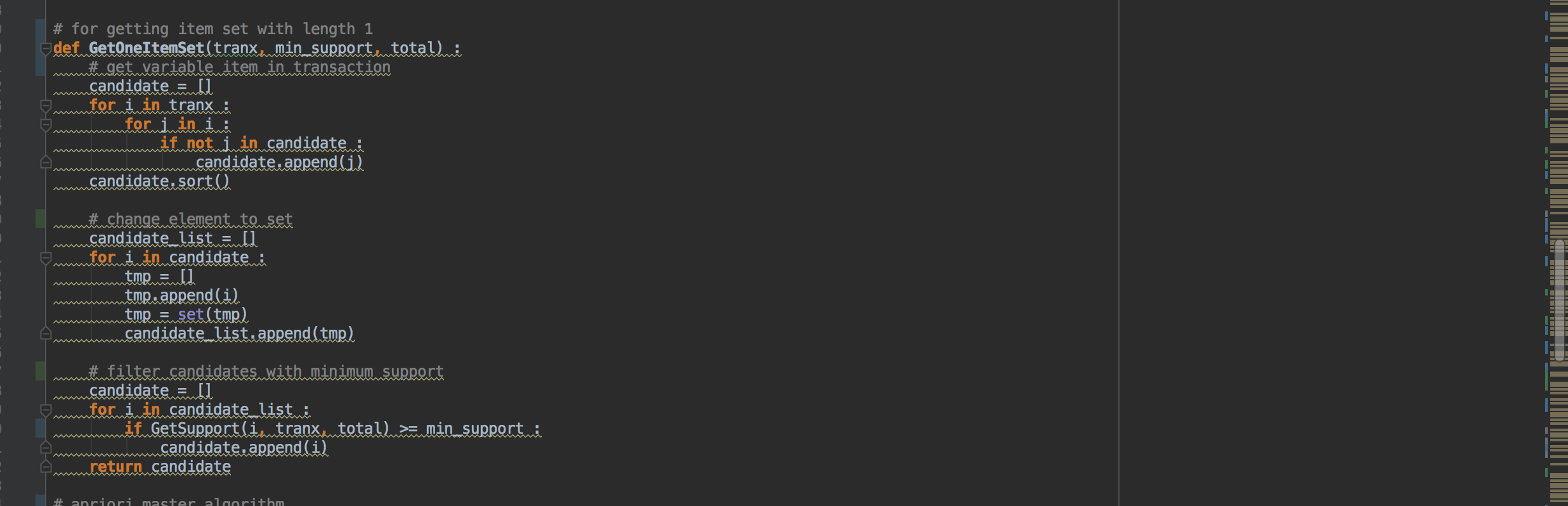
중복되지 않은 항목만 새로운 list에 추가하여 return 시킵니다.



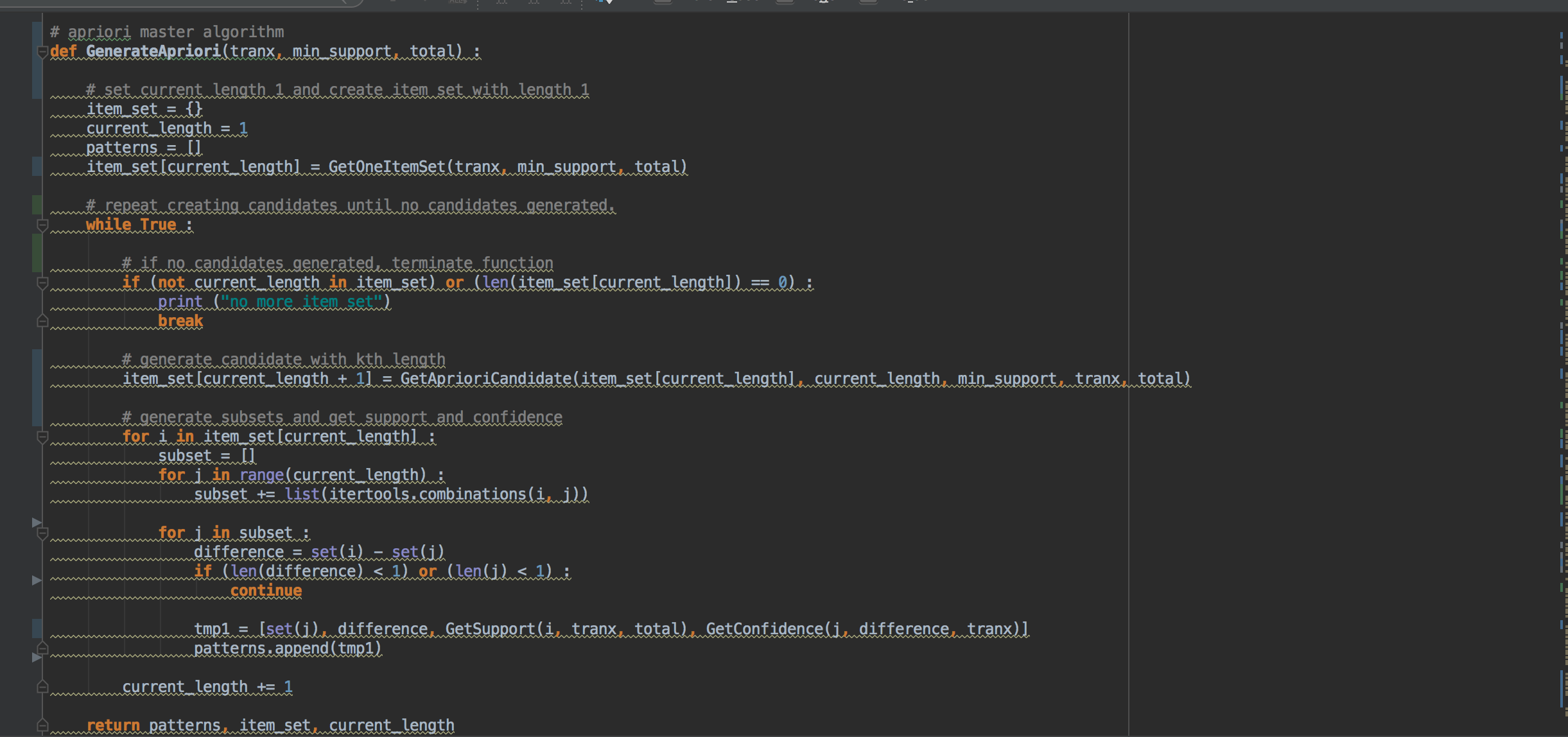
이전 candidate들을 합집합하고, 해당 개수에 해당하는 candidate들만 list로 return 시킵니다.



생성된 candidate들의 support를 확인하고 frequent pattern이 아닌 candidate들을 제거합니다.



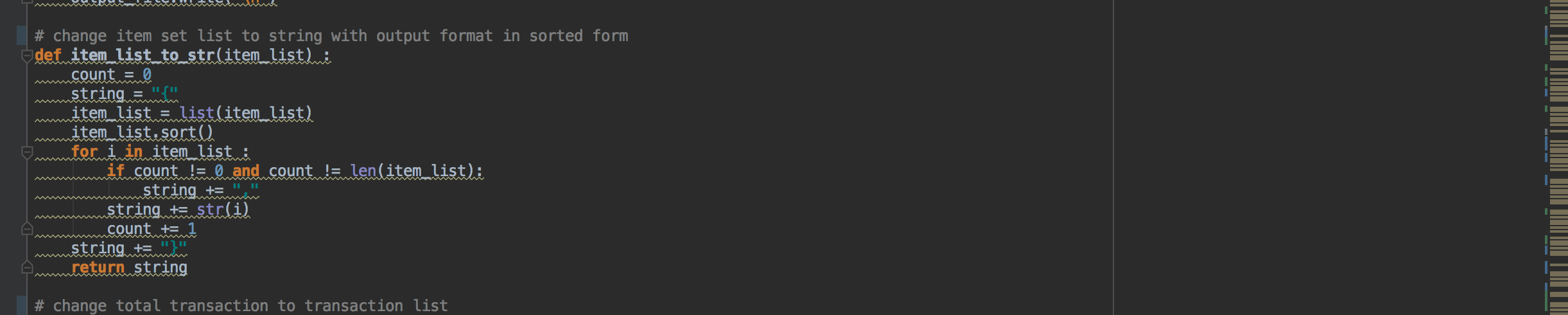
1개짜리 item을 set으로 저장하고, 빈발패턴만 return합니다.



length별로 frequent한 item set을 생성하고, item set의 subset을 구해서 support, confidence를 구해 list에 넣어주고, 마지막에 생성된 pattern들, item set, support, confidence를 담은 list, 큰 length를 return해줍니다.



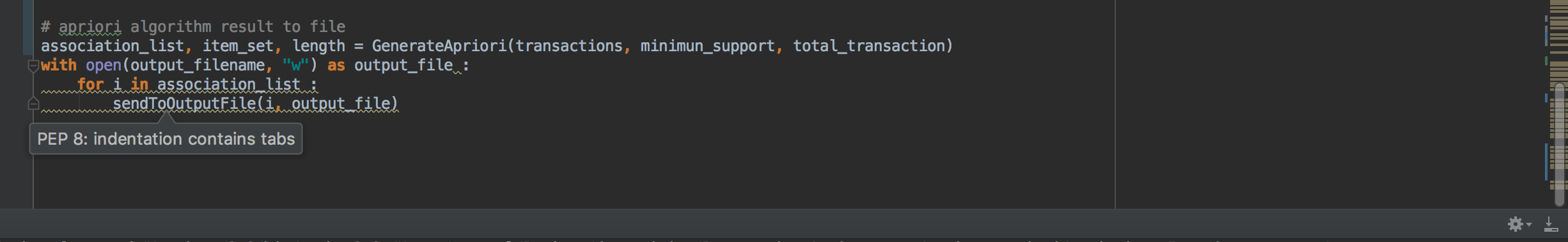
item set, support, confidence를 담은 list를 set \t set \t support \t confidence \n의 형태로 output file로 써줍니다.



list를 {element1, element2, …}의 string 형태로 변환하여 return해줍니다.



input file의 transaction을 list의 list 형태로 바꿔줍니다.



apriori algorithm을 통해 frequent pattern, support, confidence를 구하고 이를 output file에 보내줍니다.

**3. Instruction for Compiling Source Code**

이 코드는 python3을 기반으로 작성되었습니다. 따라서 python3가 설치되어 있어야 합니다.

“python3 apriori.py [minimum\_support] [input\_file] [output\_file]”의 형태로 complie, 실행하면 됩니다.

**4. Other Sepcification of Implementation and Testing**

프로그램 종료 시 no more item set을 출력을 한 후 apriori algorithm을 종료하게 되니 해당 message가 뜨기 전에 종료하면 apriori 알고리즘이 제대로 수행되지 않고, output file에 존재하지 않을 수 있습니다. 또한 “no more item set”이 출력된 후 파일에 저장하는 별도의 시간이 들 수도 있습니다. 그 외에 과제 명세서에 나온 기본적인 format 외의 다른 특별한 특이사항은 존재하지 않습니다.