Midterm project (Data Mining)

Vn233@njit.edu

Apriori Algorithm can used to analyze customer based on their transactions of various items

It is a type of frequent itemset mining ,

The important property:

If an itemset is frequent , all its subsets must be frequent and if itemset not frequent , all its subsets must be not frequent

→ find the frequency of all individual items in all transaction

→ Pruning stage is implemented by following the important property that If an itemset is frequent , all its subsets must be frequent and if itemset not frequent , all its subsets must be not frequent

→We use generate_itemsets function to create k+1 candidates

K =2 ; C[k]= (a,c) , (d,f) , (f,i)

By using the k-1 itemset we can join and generate K+1 itemset

Usage:

Python Aprioro.py “support_ in_ percentage” “input_Text_file.txt”

Example:

Python .Apriori.py 30 example2.txt .

Source code:

```
import sys


def generate_itemsets(itemset):
    candidates_lst=dict()
    for i in range(len(itemset)):
        item1=str(itemset[i])
        l1=len(item1)-1
        for j in range(i+1,len(itemset)):
            item2=str(itemset[j])
            l2=len(item2)-1
            if item1[0:l1]==item2[0:l2]:
                supset=item1[0:l1]+item2[:l2]
                sortsupset=sorted(supset)
                sortsupset=','.join(sortsupset)
                candidates_lst.insert(len(candidates_lst),sortsupset)
    return candidates_lst
def pruning_stage(c,support):
    lst=dict()
    for item in c:
        if support < c[item]:
            lst.insert(len(lst),item)
    return sorted(lst)



def find_L_K(c):
    l_k= dict()
    f= open(str(sys.argv[2]),'r')
    for line in f:
        l=str(line.split())
        for i in range(len(c)):
            item =str(c[i])
            if not (item in l_k):
                l_k[item]=0
            flag = True

            for i in item:
                if not (item in l):
                    flag = False
                if flag:
                    l_k[item]+=1
    f.close()
    return l_k


support=sys.argv[1]/5
c1=dict()
f=open(str(sys.argv[2]),'r')
for line in f:
```

```
    for item in line.split(" "):
        if item in c1:
            c1[item]=c1[item]+1
        else:
            c1[item]=1
f.close()
l1=pruning_stage(c1,support)
L=dict()
L=generate_itemsets(l1)
print("frequent 1-pair itemset"+ l1)
k=2
while not L:
    ck ={}
    ck=find_L_K(L)
    freq_items=dict()
    freq_items=pruning_stage(ck,support)
    print 'Frequent'k'pair itemset\t',fruquent_items
    L = generate_itemsets(freq_items)
    k = k + 1
```

Screen shots of outputs for 5 different files of 20 transactions each

1)taking support as 30 % for example.txt

```
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$ python Apriori.py 30 example.txt
a -> apple , b->banana, c->chicken d -> dessert, e ->icecream
 f -> coke, g ->pretz  h -> chips i -> bars, j -> juice
Frequent 1-pair itemset is ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
Frequent 2 -pair itemset is ['ab', 'ad', 'af', 'ai', 'bd', 'bi', 'cd', 'cf', 'ci', 'de', 'df', 'dh', 'di', 'ef', 'eh', 'ei', 'fh', 'fi', 'hi', 'ij']
Frequent 3 -pair itemset is ['adf', 'cdf', 'def', 'dfh', 'dfi']
```

2) taking support as 20 % for example2.txt

```
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$ python Apriori.py 20 example2.txt
a -> apple , b->banana, c->chicken d -> dessert, e ->icecream
 f -> coke, g ->pretz  h -> chips i -> bars, j -> juice
Frequent 1-pair itemset is ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
Frequent 2 -pair itemset is ['ab', 'ac', 'ad', 'ae', 'ai', 'bc', 'bd', 'be', 'bi', 'cd', 'ce', 'cf', 'cg', 'ci', 'cj', 'de', 'df', 'dh', 'di', 'ef', 'eh', 'ei', 'ej', 'fh', 'fi', 'gi
', 'hi', 'hj']
Frequent 3 -pair itemset is ['abi', 'ade', 'adi', 'aei', 'bci', 'bde', 'cde', 'cdi', 'cei', 'cfi', 'def', 'deh', 'dei', 'dfi', 'efi', 'fhi']
Frequent 4 -pair itemset is ['adei', 'cdei', 'defi']
```

3) taking support as 25 % for example3.txt

```
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$ python Apriori.py 25 example3.txt
a -> apple , b->banana, c->chicken d -> dessert, e ->icecream
 f -> coke, g ->pretz  h -> chips i -> bars, j -> juice
Frequent 1-pair itemset is ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
Frequent 2 -pair itemset is ['ab', 'ad', 'af', 'ai', 'bd', 'bi', 'cd', 'cf', 'ci', 'de', 'df', 'dh', 'di', 'ef', 'eh', 'ei', 'fh', 'fi', 'hi', 'ij']
Frequent 3 -pair itemset is ['adf', 'cdf', 'def', 'dfh', 'dfi']
```

4) taking support as 40 % for example4.txt

```
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$ python Apriori.py 40 example4.txt
a -> apple , b->banana, c->chicken d -> dessert, e ->icecream
 f -> coke, g ->pretz  h -> chips i -> bars, j -> juice
Frequent 1-pair itemset is ['a', 'b', 'c', 'd', 'e', 'f', 'h', 'i', 'j']
Frequent 2 -pair itemset is ['bh', 'de']
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$
```

5) taking support as 10 % for example.5txt

```
bharat@DESKTOP-HUMECVL:/mnt/c/linux/test/Apriori-Python$ python Apriori.py 10 example5.txt
a -> apple , b->banana, c->chicken d -> dessert, e ->icecream
 f -> coke, g ->pretz  h -> chips i -> bars, j -> juice
Frequent 1-pair itemset is ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']



Frequent 2 -pair itemset is ['ab', 'ac', 'ad', 'ae', 'af', 'ag', 'ah', 'ai', 'aj', 'bc', 'bd', 'be', 'bf', 'bg', 'bh', 'bi', 'bj', 'cd', 'ce', 'cf', 'cg', 'ch', 'ci', 'cj', 'de', 'df
', 'dg', 'dh', 'di', 'dj', 'ef', 'eg', 'eh', 'ei', 'ej', 'fg', 'fh', 'fi', 'fj', 'gh', 'gi', 'gj', 'hi', 'hj', 'ij']
Frequent 3 -pair itemset is ['abc', 'abd', 'abf', 'abg', 'abh', 'ace', 'acf', 'acg', 'ach', 'acj', 'adf', 'adg', 'adh', 'adi', 'aej', 'afg', 'afh', 'afi', 'agh', 'agi', 'agj', 'ahi',
 'bcf', 'bcg', 'bch', 'bci', 'bdh', 'bdi', 'bfg', 'bfh', 'bgh', 'bhi', 'cde', 'cdf', 'cdg', 'cdh', 'cdi', 'ceg', 'ceh', 'cej', 'cfg', 'cfh', 'cgh', 'chi', 'chj', 'deh', 'dfg', 'dfh',
 'dfi', 'dgh', 'dgi', 'dhi', 'dij', 'efg', 'efh', 'egh', 'ehi', 'fgh', 'fgi', 'fgj', 'fhi', 'fij', 'ghi', 'gij']
Frequent 4 -pair itemset is ['abcf', 'abcg', 'abch', 'abdh', 'abfg', 'abfh', 'abgh', 'acej', 'acfg', 'acfh', 'acgh', 'adfg', 'adfh', 'adgh', 'adhi', 'afgh', 'afgi', 'bcfg', 'bcfh', '
bcgh', 'bdhi', 'bfgh', 'cdeh', 'cdfg', 'cdfh', 'cdgh', 'cdhi', 'cfgh', 'dfgh', 'dfgi', 'efgh', 'fghi', 'fgij']
Frequent 5 -pair itemset is ['abcfg', 'abcfh', 'abcgh', 'abfgh', 'acfgh', 'adfgh', 'bcfgh', 'cdfgh']
Frequent 6 -pair itemset is ['abcfgh']
```