In [39]:

```
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
import numpy as np
import matplotlib.pyplot as plt
from itertools import permutations
import seaborn as sns

from pandas.plotting import parallel_coordinates

```

In [2]:

```
# !pip install mlxtend
#https://goldinlocks.github.io/Market-Basket-Analysis-in-Python/
```

In [10]:

```
df = pd.read_csv("tel_samp_rec.csv",encoding="latin-1")
```

In [11]:

```
df.head()
```

Out[11]:

	Defence.date	Domains	Full.Text.Language	def.date	n.disc	these.id	disc1.lev1	
0	2010/09/23	Sciences du Vivant [q-bio] / Ecologie, Environ...	French	2010.0	1	tel-00662843v1	Sciences du Vivant [q-bio]	Env
1	2009/11/02	Sciences de l'Homme et Société	French	2009.0	1	tel-00491490v1	Sciences de l'Homme et Société	
2	1996/05/30	Sciences du Vivant [q-bio] / Alimentation et N...	French	1996.0	1	tel-01776364v1	Sciences du Vivant [q-bio]	Alin
3	2018/02/02	Informatique [cs] / Autre [cs.OH] \r\n\r\nInf...	French	2018.0	1	tel-02437294v1	Informatique [cs]	Au
4	2015/07/08	Informatique [cs] / Automatique \r\n\r\nInfor...	French	2015.0	1	tel-01245100v1	Informatique [cs]	A

5 rows × 25 columns

In [12]:

```
cols = ['disc1.rec.lev1','disc2.rec.lev1','disc3.rec.lev1']
#subset columns shown above and take columns where all the 3 columns are not null
df_sub = df[df[cols].notnull().all(axis=1)]
```

In [13]:

```
df_sub.head()
```

Out[13]:

	Defence.date	Domains	Full.Text.Language	def.date	n.disc	these.id	disc1.lev1	dis
53	1985/10/28	Planète et Univers [physics] / Sciences de la ...	French	1985.0	2	tel-00711880v1	Planète et Univers [physics]	S de
104	2018/12/17	Sciences de l'ingénieur [physics] / Génie civi...	English	2018.0	2	tel-02182014v1	Sciences de l'ingénieur [physics]	Gé
113	2003/06/17	Sciences de l'ingénieur [physics] / Traitement...	French	2003.0	3	tel-00130932v1	Sciences de l'ingénieur [physics]	Tra d [e
193	1997/10/24	Planète et Univers [physics] / Sciences de la ...	French	1997.0	2	tel-00675418v1	Planète et Univers [physics]	S de
212	2002/12/13	Sciences du Vivant [q-bio] / Autre [q-bio.OT] ...	French	2002.0	2	tel-00008546v1	Sciences du Vivant [q-bio]	A

5 rows × 25 columns

In [14]:

```
df_sub = df_sub[cols]
```

In [15]:

```
df_sub.head()
```

Out[15]:

	disc1.rec.lev1	disc2.rec.lev1	disc3.rec.lev1
53	VIII	VIII	VIII
104	IX	IX	V
113	IX	VI	V
193	VIII	VIII	VIII
212	X	IX	IX

In [16]:

```
# Getting the list of transactions from the dataset
transactions = []
for i in range(0, len(df_sub)):
    transactions.append([str(df_sub.values[i,j]) for j in range(0, len(df_sub.columns))
```

In [17]:

```
#check transactions
transactions[:1]

```

Out[17]:

```
[['VIII', 'VIII', 'VIII']]
```

In [19]:

```

# Extract unique items.
flattened = [item for transaction in transactions for item in transaction]
items = list(set(flattened))
```

In [20]:

```
print('# of items:',len(items))
print(list(items))
```

```
# of items: 13
['IV', 'pharmacie', 'VI', 'I', 'III', 'XII', 'II', 'VIII', 'X', 'V', 'I - Dr
oit', 'VII', 'IX']
```

In [21]:

```
#remove nan if present in list
if 'nan' in items: items.remove('nan')
print(list(items))
```

```
['IV', 'pharmacie', 'VI', 'I', 'III', 'XII', 'II', 'VIII', 'X', 'V', 'I - Dr
oit', 'VII', 'IX']
```

In [22]:

```
# Compute and print rules.
rules = list(permutations(items, 2))
print('# of rules:',len(rules))
print(rules[:5])
```

```
# of rules: 156
[('IV', 'pharmacie'), ('IV', 'VI'), ('IV', 'I'), ('IV', 'III'), ('IV', 'XI
I')]
```

In [23]:

```
# Import the transaction encoder function from mlxtend
from mlxtend.preprocessing import TransactionEncoder

# Instantiate transaction encoder and identify unique items
encoder = TransactionEncoder().fit(transactions)

# One-hot encode transactions
onehot = encoder.transform(transactions)

# Convert one-hot encoded data to DataFrame
onehot = pd.DataFrame(onehot, columns = encoder.columns_)

# Print the one-hot encoded transaction dataset
onehot.head()
```

Out[23]:

	I	I - Droit	II	III	IV	IX	V	VI	VII	VIII	X	XII	pharmacie
0	False	False	False	False	False	False	False	False	False	True	False	False	False
1	False	False	False	False	False	True	True	False	False	False	False	False	False
2	False	False	False	False	False	True	True	True	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	True	False	False	False
4	False	False	False	False	False	True	False	False	False	False	True	False	False

In [24]:

```
def leverage(antecedent, consequent):
    # Compute support for antecedent AND consequent
    supportAB = np.logical_and(antecedent, consequent).mean()

    # Compute support for antecedent
    supportA = antecedent.mean()

    # Compute support for consequent
    supportB = consequent.mean()

    # Return leverage
    return supportAB - supportB * supportA

# Define a function to compute Zhang's metric
def zhang(antecedent, consequent):
    # Compute the support of each book
    supportA = antecedent.mean()
    supportC = consequent.mean()

    # Compute the support of both books
    supportAC = np.logical_and(antecedent, consequent).mean()

    # Complete the expressions for the numerator and denominator
    numerator = supportAC - supportA*supportC
    denominator = max(supportAC*(1-supportA), supportA*(supportC-supportAC))

    # Return Zhang's metric
    return numerator / denominator

def conviction(antecedent, consequent):
    # Compute support for antecedent AND consequent
    supportAC = np.logical_and(antecedent, consequent).mean()

    # Compute support for antecedent
    supportA = antecedent.mean()

    # Compute support for NOT consequent
    supportnC = 1.0 - consequent.mean()

    # Compute support for antecedent and NOT consequent
    supportAnC = supportA - supportAC

    # Return conviction
    return supportA * supportnC / supportAnC

```

In [25]:

```
# Create rules DataFrame
rules_ = pd.DataFrame(rules, columns=['antecedents','consequents'])

# Define an empty list for metrics
zhangs, conv, lev, antec_supp, cons_supp, suppt, conf, lft = [], [], [], [], [], [], [

# Loop over lists in itemsets
for itemset in rules:
    # Extract the antecedent and consequent columns
    antecedent = onehot[itemset[0]]
    consequent = onehot[itemset[1]]

    antecedent_support = onehot[itemset[0]].mean()
    consequent_support = onehot[itemset[1]].mean()
    support = np.logical_and(onehot[itemset[0]], onehot[itemset[1]]).mean()
    confidence = support / antecedent_support
    lift = support / (antecedent_support * consequent_support)

    # Complete metrics and append it to the list
    antec_supp.append(antecedent_support)
    cons_supp.append(consequent_support)
    suppt.append(support)
    conf.append(confidence)
    lft.append(lift)
    lev.append(leverage(antecedent, consequent))
    conv.append(conviction(antecedent, consequent))
    zhangs.append(zhang(antecedent, consequent))

# Store results
rules_['antecedent support'] = antec_supp
rules_['consequent support'] = cons_supp
rules_['support'] = suppt
rules_['confidence'] = conf
rules_['lift'] = lft
rules_['leverage'] = lev
rules_['conviction'] = conv
rules_['zhang'] = zhangs

# Print results
rules_.sort_values('zhang',ascending=False).head()
```

Out[25]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leve
2	IV	I	0.143959	0.021994	0.020566	0.142857	6.495362	0.01
9	IV	I - Droit	0.143959	0.004856	0.003999	0.027778	5.720588	0.00
97	X	pharmacie	0.199372	0.019709	0.016281	0.081662	4.143453	0.01
45	I	I - Droit	0.021994	0.004856	0.001428	0.064935	13.372804	0.00
123	I - Droit	I	0.004856	0.021994	0.001428	0.294118	13.372804	0.00

In [26]:

```
# Function to convert rules to coordinates.
def rules_to_coordinates(rules):
    rules['antecedent'] = rules['antecedents'].apply(lambda antecedent: list(antecedent
    rules['consequent'] = rules['consequents'].apply(lambda consequent: list(consequent
    rules['rule'] = rules.index
    return rules[['antecedent','consequent','rule']]
```

In [ ]:

```

```

In [32]:

```
rules_.head()
```

Out[32]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverag
0	IV	pharmacie	0.143959	0.019709	0.000857	0.005952	0.302019	-0.0019ε
1	IV	VI	0.143959	0.162239	0.000857	0.005952	0.036689	-0.0224ς
2	IV	I	0.143959	0.021994	0.020566	0.142857	6.495362	0.0173ς
3	IV	III	0.143959	0.032562	0.022279	0.154762	4.752820	0.0175ς
4	IV	XII	0.143959	0.049414	0.034276	0.238095	4.818332	0.0271ϵ

In [33]:

```
#remove rows where antecedent = consequent
rules_ = rules_[rules_['antecedents'] != rules_['consequents']]
```

In [35]:

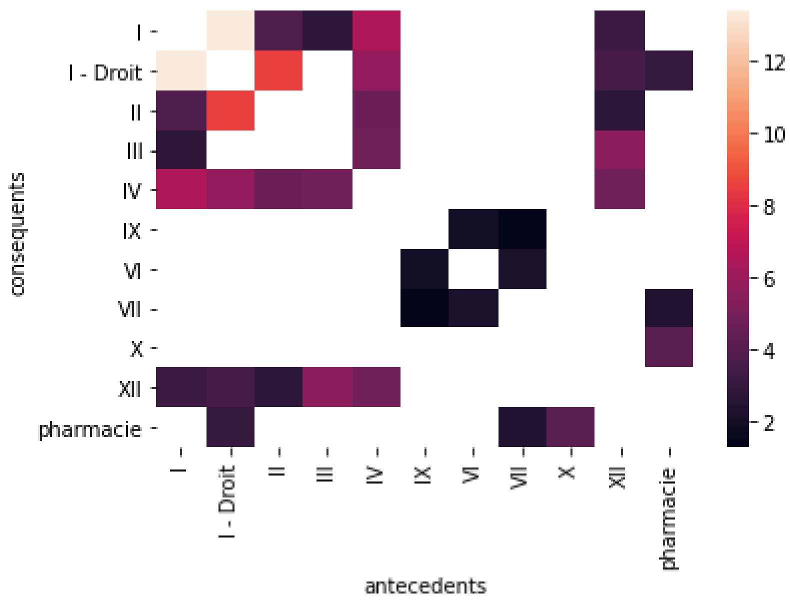
```
#filter rules with lift > 1
rules_ = rules_.query("lift>1")
#create support table based on lift values > 1
support_table = rules_.pivot(index='consequents', columns='antecedents',
values='lift')




sns.heatmap(support_table)
```

Out[35]:

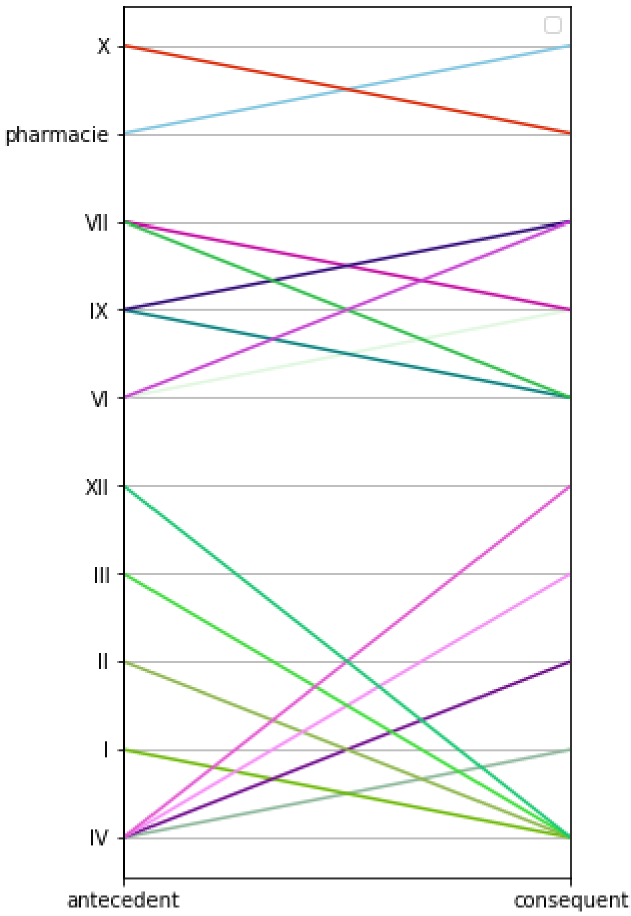
<AxesSubplot:xlabel='antecedents', ylabel='consequents'>



In [41]:

```
# Generate frequent itemsets
frequent_itemsets = apriori(onehot, min_support = 0.01, use_colnames = True, max_len =
# Generate association rules
rules = association_rules(frequent_itemsets, metric = 'lift', min_threshold = 1.00)
# Generate coordinates and print example
coords = rules_to_coordinates(rules)
# Generate parallel coordinates plot

plt.figure(figsize=(4,8))
parallel_coordinates(coords, 'rule')
plt.legend([])
plt.grid(True)
plt.show()
```



https://www.galaxie.enseignementsup-recherche.gouv.fr/ensup/pdf/qualification/sections.pdf (https://www.galaxie.enseignementsup-recherche.gouv.fr/ensup/pdf/qualification/sections.pdf)

> LINK : Reference what I , II etc means