IMPLEMENTATION OF APRIORI ALGORITHM

Report submitted to the SASTRA Deemed to be University

as the requirement for the course

CSE405: NATURAL LANGUAGE PROCESSING

Submitted by

JAISAIARUN P SRINIVASAN

Reg. No.:122003096, B.Tech COMPUTER SCIENCE

JUNE 2021

SCHOOL OF COMPUTING

THANJAVUR, TAMIL NADU, INDIA – 613 401

1. INTRODUCTION

Association rule mining is a technique to identify underlying relations between different items. Different statistical algorithms have been developed to implement association rule mining, and Apriori is one such algorithm.

In this project apriori algorithm is applied on data on abstracts of various covid research paper abstracts, the data was taken from kaggle website. The abstracts were extracted, preprocessed and tokenized. The cleaned data is run in apriori algorithm which produces association rules from which can infer useful information.

There are three major components of Apriori algorithm:

- Support
- Confidence
- Lift

1.1 Support

Support refers to the default popularity of a word and can be calculated by finding number of abstracts containing a particular item divided by total number of abstracts.

Support(A) = (Number of abstracts containing (A)) / (Total number of Abstracts)

1.2 Confidence

Confidence refers to the likelihood that a word B is also present if word A is also present. It can be calculated by finding the number of abstracts where A and B are both present, divided by total number of abstracts where only A is present.

Confidence(A→B) = (Number of abstracts containing both (A and B)) / (Number of abstracts containing (A))

1.1 Lift

Lift(A -> B) refers to the increase in the ratio of presence of B when A is sold. Lift (A –> B) can be calculated by dividing Confidence (A -> B) divided by Support(B). Mathematically it can be represented as:

Lift(A→B) = (Confidence (A→B)) / (Support (B))

2. Implementing Apriori Algorithm with Python

STEP 1: import the necessary libraries:

1. apriori
2. pandas – csv file importer and table manipulation
3. NLTK – Natural Language ToolKit in python
4. Re – word manipulation

```
import re
import pandas as pd
from nltk.tokenize import WordPunctTokenizer
from nltk.corpus import stopwords
from apyori import apriori
```

STEP 2: **Import dataset**

CSV file is imported using the pandas.read_csv function

```
df = pd.read_csv(r'D:\PROJECT\NLP\metadata.csv',encoding='UTF-8')
```

Extracting column in which contains data on which apriori is applied on

```
title=list(df["abstract"])
```

The abstract column contains abstracts of papers which are used as data in this project.

STEP 3: **Pre-processing**

Perform preprocessing on text, where the text was tokenized and stop words were removed using **nltk.tokenize WordPunctTokenizer** and **nltk.corpus stopwords** respectively.

```
word_punct_token = WordPunctTokenizer().tokenize(text)
stop_words = stopwords.words('english')
tokens = [x for x in clean_token if x not in stop_words]
```

Characters which are not alphabets are removed

```
re.sub(r'[^a-zA-Z]+', '', token)
```

Empty tokens or tokens token which are or smaller than length two are removed

```
new_token != "" and len(token) > 2
```

STEP 4: **Apply Apriori**

Apply the apriori function with appropriate parameters.

The apriori class requires some parameter values to work.

- The first parameter is the list of list that you want to extract rules from.
- The second parameter is the **min_support** parameter. This parameter is used to select the items with support values greater than the value specified by the parameter.
- Next, the **min_confidence** parameter filters those rules that have confidence greater than the confidence threshold specified by the parameter.
- Similarly, the **min_lift** parameter specifies the minimum lift value for the short listed rules.
- Finally, the **min_length** parameter specifies the minimum number of items that you want in your rules.

association_rules = apriori(clean, min_support=0.05, min_confidence=0.05, min_lift=3, min_length=2)

STEP 5: **Display the results**

The obtained result is in such

RelationRecord(items=frozenset({‘virus’, ‘covid’ }), support= 0.054379797957218194, ordered_statistics[OrderedStatistic(items_base=frozenset({'virus'}), items_add=frozenset({'covid'}), confidence= 0.3341192629048873, lift= 3.596496969276992)])

Hence a result is more readable format for inference

=====================================

Rule: virus -> covid

Support: 0.054379797957218194

Confidence: 0.3341192629048873

Lift: 3.596496969276992

=====================================

2. Snapshot of Dataset

	cord_uid	sha	source_x	title	doi	pmcid	pubmed_i	license	abstract	publish_ti	authors	journal	mag_id	who_covic	arxiv_id	pdf_json_	pmc_json_	url	s2_id
2	ug7v899j	d1aafb70c	PMC	Clinical fea	10.1186/1	PMC35282	11472636	no-cc	OBJECTIVE	7/4/2001	Madani, T	BMC Infect Dis				document	document	https://www.ncbi.nlm.	
3	02tnwd4m	6b0567729	PMC	Nitric oxid	10.1186/r	PMC59543	11667967	no-cc	Inflammat	########	Vliet, Albe	Respir Res				document	document	https://www.ncbi.nlm.	
4	ejv2xln0	06ced00a5	PMC	Surfactant	10.1186/r	PMC59549	11667972	no-cc	Surfactant	########	Crouch, Er	Respir Res				document	document	https://www.ncbi.nlm.	
5	2b73a28n	348055649	PMC	Role of en	10.1186/r	PMC59574	11686871	no-cc	Endothelir	########	Fagan, Kar	Respir Res				document	document	https://www.ncbi.nlm.	
6	9785vg6d	5f48792a5	PMC	Gene expr	10.1186/r	PMC59580	11686888	no-cc	Respirator	########	Domachov	Respir Res				document	document	https://www.ncbi.nlm.	
7	zjufx4fo	b2897e127	PMC	Sequence	10.1093/e	PMC12534	11742998	green-oa	Nidovirus	########	Pasternak	The EMBO Journal				document	document	http://europepmc.org/	
8	5yhe786e	3bb07ea1	PMC	Debate: Tı	10.1186/c	PMC13726	11299062	no-cc	Recent evi	3/8/2001	Alvarez, G	Crit Care				document	document	https://www.ncbi.nlm.	
9	8zchiykl	5806726a	PMC	The 21st Ir	10.1186/c	PMC13727	11353930	no-cc	The 21st Ir	5/2/2001	Ball, Jonat	Crit Care				document	document	https://www.ncbi.nlm.	
10	8qnrcgnk	faaf1022c	PMC	Heme oxy	10.1186/1	PMC19368	12964953	no-cc	Heme oxy	8/7/2003	Slebos, Dir	Respir Res				document	document	https://www.ncbi.nlm.	
11	jg13scgo	5b44feca5	PMC	Technical	10.1197/ja	PMC21277	12807803	bronze-oa	This repor	9/1/2003	Tsui, Fu-Cl	Journal of the American Medical Informat				document	document	https://academic.oup.c	
12	5tkvsudh	9d4e3e8el	PMC	Conservat	10.1093/e	PMC30201	10775274	no-cc	Regulatior	########	Ivanov, Iva	EMBO J				document	document	https://www.ncbi.nlm.	
13	6lvn10f4	14e0cac6e	PMC	Heteroger	10.1093/e	PMC30207	10970862	green-oa	Heteroger	9/1/2000	Shi, Steph	The EMBO Journal				document	document	http://europepmc.org/	
14	tvxpckxo	d09b7902	PMC	A Method	10.1251/b	PMC30219	14702098	no-cc	The UBA d	########	Pridgeon, .	Biol Proced Online				document	document	https://www.ncbi.nlm.	
15	mcuixluu	44102e3e	PMC	Vaccinia vi	10.1093/e	PMC30661	10921875	green-oa	We exami	8/1/2000	Ploubidou	The EMBO Journal				document	document	http://europepmc.org/	
16	6iu1dtyl	6e8517cb	PMC	The site of	10.1186/1	PMC34038	14733617	no-cc		########	Barry, Joh	J Transl Med				document	document	https://www.ncbi.nlm.	
17	t35n7bk9	30a4842a	PMC	Multi-face	10.1186/1	PMC42049	15169556	no-cc	There are	########	Shieh, Biel	Retrovirology				document	document	https://www.ncbi.nlm.	
18	eiqypt0m	6a8ac55ea	PMC	Herpes sir	10.1186/c	PMC46889	15153242	no-cc	INTRODUC	########	Verheij, Jo	Crit Care				document	document	https://www.ncbi.nlm.	
19	sgmk96vr	367af6bb9	PMC	Logistics o	10.1186/1	PMC52075	15298713	cc-by	BACKGRO	8/6/2004	Porco, Tra	BMC Public Health				document	document	https://www.ncbi.nlm.	
20	di0fcy0j	4df2c6eec	PMC	Protection	10.1186/1	PMC52169	15450125	cc-by	BACKGRO	########	Kremer, Te	Respir Res				document	document	https://www.ncbi.nlm.	
21	4k8f7ou1	83b05e8a	PMC	Bioinform	10.1186/1	PMC54488	15627404	cc-by	BACKGRO	1/3/2005	Bratlie, Ma	BMC Genomics				document	document	https://www.ncbi.nlm.	
22	wnnsmx6	025339bfc	PMC	Managing	10.1186/1	PMC54496	15679928	cc-by	In the 198	########	Howse, Ge	Aust New Zealand Health Policy				document	document	https://www.ncbi.nlm.	
23	gdsfkw1b	8d7400a2	PMC	Protein se	10.1186/1	PMC54505	15631634	cc-by	Lactococci	1/4/2005	Le Loir, Yv	Microb Cell Fact				document	document	https://www.ncbi.nlm.	

3. CODE

```
import re
import pandas as pd
from nltk.tokenize import WordPunctTokenizer
from nltk.corpus import stopwords
from apyori import apriori

stop_words = stopwords.words('english')
stop_words = stop_words + ['conclusions', 'conclusion']

def tokenizing(text):
    word_punct_token = WordPunctTokenizer().tokenize(text)
    clean_token=[]
    for token in word_punct_token:
        token = token.lower()
        if token in ["coronavirus","cov"]:
            token="covid"
        # remove any value that are not alphabetical
        new_token = re.sub(r'[^a-zA-Z]+', '', token)
        # remove empty value and double character value
        if new_token != "" and len(new_token) > 2:
```

```
            vowels=len([v for v in new_token if v in "aeiou"])
            if vowels != 0: # remove line that only contains consonants
                clean_token.append(new_token)
    tokens = [x for x in clean_token if x not in stop_words]
    tokens = list(set(tokens))
df = pd.read_csv(r'D:\PROJECT\NLP\metadata.csv',encoding='UTF-8')
#title=list(df["title"])
title=list(df["abstract"])
clean=[]
for i in range(len(title)):
    clean.append(tokenizing(str(title[i])))
print(len(clean))
association_rules = apriori(clean, min_support=0.05, min_confidence=0.05, min_lift=3,
min_length=2)
for item in association_rules:
    # first index of the inner list
    # Contains base item and add item
    pair = item[0]
    items = [x for x in pair]
    print("Rule: " + items[0] + " -> " + items[1])

    #second index of the inner list
    print("Support: " + str(item[1]))

    #third index of the list located at 0th
    #of the third index of the inner list

    print("Confidence: " + str(item[2][0][2]))
    print("Lift: " + str(item[2][0][3]))
    print("=====================================")
  # print("Length : "+str(len(association_rules)))
```

4. RESULTS

Rule: public -> health
Support: 0.06505382653679001
Confidence: 0.33980403031983736
Lift: 4.137184009076776
=====================================
Rule: hospital -> patients
Support: 0.05888412624786473
Confidence: 0.7497272425406765
Lift: 3.0136906321339336
=====================================
Rule: respiratory -> sars
Support: 0.07879035127427038
Confidence: 0.5669173133888241
Lift: 3.483241975751956
=====================================
Rule: respiratory -> severe
Support: 0.0879983309023373
Confidence: 0.6331711501601727
Lift: 4.586385606740432
=====================================
Rule: respiratory -> syndrome
Support: 0.08517055160324655
Confidence: 0.6128245338909217
Lift: 6.2947194483510565
=====================================
Rule: severe -> sars
Support: 0.07778069621491122
Confidence: 0.477898592194117
Lift: 3.4616662874898037
=====================================
Rule: sars -> syndrome
Support: 0.06593122050903753
Confidence: 0.4050932814467209
Lift: 4.160976619080486
=====================================
Rule: severe -> syndrome
Support: 0.07741185543676896
Confidence: 0.5607340439886656
Lift: 5.759664140137451
=====================================
Rule: respiratory -> covid
Support: 0.08039052414509973
Confidence: 0.6647105032114967
Lift: 5.89954199792484
=====================================
Rule: covid -> sars
Support: 0.0670917649776367
Confidence: 0.5547494724519817
Lift: 3.5080568683385027
=====================================
Rule: covid -> severe
Support: 0.07507027534522938
Confidence: 0.62071992976295
Lift: 5.255311261502366
=====================================
Rule: covid -> syndrome
Support: 0.07475173103683379
Confidence: 0.6180860403863039
Lift: 7.360065081565504
=====================================
Rule: disease -> acute
Support: 0.05900334750948647
Confidence: 0.487870246291761
Lift: 6.017993106542068
=====================================
Rule: disease -> severe
Support: 0.0551435591644825
Confidence: 0.45595551653497224
Lift: 5.287407597851772
=====================================
Rule: disease -> syndrome
Support: 0.05416370942052878
Confidence: 0.4478536112009611
Lift: 7.295485586091714
=====================================
Rule: respiratory -> sars
Support: 0.06544874696591203
Confidence: 0.5411641482987539
Lift: 6.868406340961135
=====================================
Rule: respiratory -> severe
Support: 0.07699271818888001
Confidence: 0.6366156831939375
Lift: 7.234406342325618
=====================================
Rule: respiratory -> syndrome
Support: 0.07858916539528368
Confidence: 0.6498159358008718
Lift: 7.629584681192813
=====================================
Rule: severe -> sars
Support: 0.06326737044467667
Confidence: 0.5231273970703757
Lift: 6.725671310847533
=====================================
Rule: sars -> syndrome
Support: 0.06290225533096008
Confidence: 0.5201084361474362
Lift: 7.888651721169673
=====================================
Rule: severe -> syndrome
Support: 0.07340117768252495
Confidence: 0.606918965543798
Lift: 7.840129423580831
=====================================
Rule: analysis -> results
Support: 0.05338504555556176
Confidence: 0.2522222808963053
Lift: 3.1044437144933257
=====================================
Rule: covid -> background
Support: 0.06488430880542159
Confidence: 0.503265424071666
Lift: 3.8643070596446893
=====================================
Rule: patients -> background
Support: 0.05987701581730832
Confidence: 0.4644271059095506
Lift: 4.154513676271408
=====================================
Rule: results -> background
Support: 0.09711316891976222
Confidence: 0.7532437509030486
Lift: 4.188114208772029
=====================================
Rule: study -> background
Support: 0.06482656100682356
Confidence: 0.5028175119202428
Lift: 4.186833800686981
=====================================
Rule: patients -> background
Support: 0.06201182153322268
Confidence: 0.48098540673313106
Lift: 3.57525226167988
=====================================
Rule: study -> background

Support: 0.06842555284202996
Confidence: 0.530732553099263
Lift: 3.21423139878707
======================================
Rule: clinical -> results
Support: 0.05954729451563568
Confidence: 0.28133636091601977
Lift: 3.3824443730762597
======================================
Rule: results -> compared
Support: 0.05227293472449643
Confidence: 0.24696801675731814
Lift: 3.3573903426765916
======================================
Rule: public -> health
Support: 0.0518873284564386
Confidence: 0.2710297652061379
Lift: 4.18987431155258
======================================
Rule: covid -> sars
Support: 0.07263741647526066
Confidence: 0.44629735607187826
Lift: 3.4834320745952496
======================================
Rule: respiratory -> covid
Support: 0.07585266487462208
Confidence: 0.5457798866058144
Lift: 3.451336098339893
======================================
Rule: respiratory -> covid
Support: 0.08016512144734611
Confidence: 0.5768091465948234
Lift: 4.883541607879399
======================================
Rule: respiratory -> covid
Support: 0.07771363425524899
Confidence: 0.5591700510675941
Lift: 6.658503345177628
======================================
Rule: severe -> covid
Support: 0.07492311160041504
Confidence: 0.46034107817328607
Lift: 3.8974673379346885
======================================
Rule: covid -> sars
Support: 0.06312952086092653
Confidence: 0.38787913471443286
Lift: 4.618799792817322
======================================
Rule: virus -> covid
Support: 0.054379797957218194
Confidence: 0.3341192629048873
Lift: 3.5964969692769992
======================================
Rule: covid -> severe
Support: 0.07195003138872279
Confidence: 0.5211712319525031
Lift: 6.2060197693717285
======================================
Rule: results -> data
Support: 0.06866772102969913
Confidence: 0.32442660752319097
Lift: 3.16196223916151
======================================
Rule: results -> disease
Support: 0.06810514570142152
Confidence: 0.3217686715600852
Lift: 3.093041327976905
======================================
Rule: patients -> disease
Support: 0.052541182563145356
Confidence: 0.38058291728511673
Lift: 3.191392597407634
======================================
Rule: disease -> sars
Support: 0.05092796986682613
Confidence: 0.36644014636696287
Lift: 4.510795937817283
======================================
Rule: disease -> severe
Support: 0.059126294435533897
Confidence: 0.4254292492661546
Lift: 4.933415132275757
======================================
Rule: disease -> syndrome
Support: 0.05561671854654379
Confidence: 0.40017692709799346
Lift: 6.5188376972131925
======================================
Rule: severe -> disease
Support: 0.05206429751665838
Confidence: 0.31989241158292325
Lift: 3.709575838345937
======================================
Rule: disease -> severe
Support: 0.052462943610206085
Confidence: 0.3800161921468088
Lift: 6.190421563988392
======================================
Rule: results -> health
Support: 0.053066501247166165
Confidence: 0.2507172906655401
Lift: 3.0862735633287453
======================================
Rule: results -> may
Support: 0.05590545753953396
Confidence: 0.2641301860555174
Lift: 3.0398896745013118
======================================
Rule: patients -> results
Support: 0.102893537276204
Confidence: 0.48612944676207076
Lift: 3.6134888495060102
======================================
Rule: patients -> study
Support: 0.06771395093672518
Confidence: 0.31992043794335606
Lift: 3.013647493909814
======================================
Rule: results -> study
Support: 0.10882665787409862
Confidence: 0.5141609899491296
Lift: 3.113870419809812
======================================
Rule: results -> used
Support: 0.055395041513215865
Confidence: 0.2617186812413089
Lift: 3.0866497639984125
======================================
Rule: results -> using
Support: 0.0646644946043065
Confidence: 0.3055130168453293
Lift: 3.0424179342533137
======================================
Rule: severe -> respiratory
Support: 0.06596102582444296
Confidence: 0.47460694036752576
Lift: 6.101860274638935
======================================
Rule: respiratory -> sars

Support: 0.06358591475307228
Confidence: 0.45751739112951867
Lift: 6.939313235780375
======================================
Rule: respiratory -> severe
Support: 0.07435308494328607
Confidence: 0.5349900143418178
Lift: 6.910957130833853
======================================
Rule: severe -> sars
Support: 0.06183857813742858
Confidence: 0.3799473503490901
Lift: 4.908128712396465
======================================
Rule: disease -> covid
Support: 0.05494609894992148
Confidence: 0.4543228131786886
Lift: 6.318511091016454
======================================
Rule: disease -> covid
Support: 0.05221332409368556
Confidence: 0.4317268148421977
Lift: 5.474907126293829
======================================
Rule: disease -> covid
Support: 0.0515967266312356
Confidence: 0.4266284675692744
Lift: 7.563954490888935
======================================
Rule: respiratory -> covid
Support: 0.06264145882116252
Confidence: 0.5179520354881938
Lift: 6.828396027276351
======================================
Rule: respiratory -> covid
Support: 0.07149736316100272
Confidence: 0.5911772407313278
Lift: 7.374494419242209
======================================
Rule: respiratory -> covid
Support: 0.07303047407216985
Confidence: 0.6038537960353034
Lift: 7.770242658475562
======================================
Rule: severe -> covid
Support: 0.06049733894418396
Confidence: 0.5002233414968501
Lift: 6.676489147496633
======================================
Rule: covid -> sars
Support: 0.06013967515931873
Confidence: 0.4972659920213176
Lift: 7.876916935846666
======================================
Rule: covid -> severe
Support: 0.06894155736498658
Confidence: 0.5700445142707515
Lift: 7.922783399370392
======================================
Rule: disease -> severe
Support: 0.0521872444427058
Confidence: 0.43151117477627343
Lift: 7.298126474791266
======================================
Rule: disease -> syndrome
Support: 0.05264922683149006
Confidence: 0.43533108451550295
Lift: 7.827342135462177
======================================
Rule: severe -> respiratory
Support: 0.062039764016415276
Confidence: 0.5129769111100843
Lift: 7.776969880295464
======================================
Rule: respiratory -> sars
Support: 0.06225957821753037
Confidence: 0.5147944488085887
Lift: 8.096045339722275
======================================
Rule: respiratory -> severe
Support: 0.07262065098534509
Confidence: 0.6004651664279222
Lift: 8.075860831947134
======================================
Rule: severe -> sars
Support: 0.06061283454138002
Confidence: 0.5011783189316575
Lift: 8.10462229286467
======================================
Rule: results -> covid
Support: 0.05983603350862585
Confidence: 0.46410923276983096
Lift: 4.138979400069814
======================================
Rule: patients -> results
Support: 0.056322731955210065
Confidence: 0.4368588354284063
Lift: 4.245736389212833
======================================
Rule: results -> study
Support: 0.06060352038031582
Confidence: 0.4700621297500361
Lift: 4.319365667682731
======================================
Rule: results -> disease
Support: 0.05544161231853686
Confidence: 0.2619387090528243
Lift: 3.1517012658883776
======================================
Rule: disease -> covid
Support: 0.05567632917735467
Confidence: 0.40060584127494736
Lift: 5.080249129377842
======================================
Rule: disease -> covid
Support: 0.05277403658975032
Confidence: 0.3797230822845041
Lift: 6.732340506728338
======================================
Rule: severe -> disease
Support: 0.05089816455142069
Confidence: 0.31272748082865975
Lift: 3.96582712612503
======================================
Rule: disease -> covid
Support: 0.05018842547832874
Confidence: 0.36354068276885715
Lift: 6.445432944776061
======================================
Rule: results -> covid
Support: 0.053923404065072456
Confidence: 0.41404909028492964
Lift: 4.666077264605544
======================================
Rule: patients -> covid
Support: 0.06569277798579404
Confidence: 0.31037123092358876
Lift: 3.5146620202659666
======================================
Rule: results -> covid

Support: 0.06758914117846491
Confidence: 0.31933076340849487
Lift: 3.157120695814833
=====================================
Rule: severe -> respiratory
Support: 0.06317981733067321
Confidence: 0.45459541329902015
Lift: 6.067492441097467
=====================================
Rule: respiratory -> covid
Support: 0.06081215758815388
Confidence: 0.43755947833313225
Lift: 6.931138908801
=====================================
Rule: respiratory -> covid
Support: 0.06982640266608546
Confidence: 0.5024193440293806
Lift: 6.982892631623343
=====================================
Rule: severe -> covid
Support: 0.05909276345570278
Confidence: 0.3630765709053451
Lift: 5.0462322795074215
=====================================
Rule: disease -> severe
Support: 0.05058520873966361
Confidence: 0.36397388984947787
Lift: 6.937732898743996
=====================================
Rule: patients -> results
Support: 0.06382435727631576
Confidence: 0.30154371512559186
Lift: 3.7459512767587264
=====================================
Rule: severe -> respiratory
Support: 0.06098726381616081
Confidence: 0.4388194137279343
Lift: 7.096208013591653
=====================================
Rule: covid -> acute
Support: 0.05052187244442706
Confidence: 0.417741016280825
Lift: 7.915654046481597
=====================================
Rule: covid -> severe
Support: 0.05927345818034824
Confidence: 0.49010366126026217
Lift: 7.757282024022589
=====================================
Rule: covid -> acute
Support: 0.05950631220695321
Confidence: 0.4920290189917287
Lift: 8.090964677227227
=====================================
Rule: covid -> severe
Support: 0.06838270770113465
Confidence: 0.5654236557152319
Lift: 8.097562442377646
=====================================
Rule: covid -> severe
Support: 0.05787260835629274
Confidence: 0.4785207091477596
Lift: 8.097788648968315
=====================================
Rule: severe -> acute
Support: 0.06036880352149802
Confidence: 0.4991605440290806
Lift: 8.184668614315006
=====================================
Rule: covid -> severe
Support: 0.05824331196664785
Confidence: 0.4190759580200249
Lift: 7.091832121443658
=====================================
Rule: covid -> severe
Support: 0.05762857733641073
Confidence: 0.47650293424518275
Lift: 8.181247222308459
=====================================

Let us consider one of the above rules :

Rule: respiratory -> severe
Support: 0.07435308494328607
Confidence: 0.5349900143418178
Lift: 6.910957130833853

The support value is 0.0743.This number is calculated by dividing number of abstracts containing the word “respiratory” by total number of abstracts. The confidence value is 0.535 which shows that out of all abstracts that contain “respiratory”, 53.5% also contain “severe”. . Finally, the lift of 6.91 tells us that “severe” is 6.91 times more likely to be present in the abstract which contain “respiratory” compared to the default likelihood of the presence of “severe”.

Reference

1. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge?select=metadata.csv
2. **https://www.geeksforgeeks.org/implementing-apriori-algorithm-in-python/**
3. https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/