

Configuration Manual

MSc Research Project Data Analytics

Saurabh Shantkumar Yeramwar Student ID: x20131283

School of Computing National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland Project Submission Sheet School of Computing



Student Name:	Saurabh Shantkumar Yeramwar
Student ID:	x20131283
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	31/01/2022
Project Title:	Configuration Manual
Word Count:	746
Page Count:	5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

<u>ALL</u> internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Saurabh
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).Attach a Moodle submission receipt of the online project submission, to
each project (including multiple copies).You must ensure that you retain a HARD COPY of the project, both for

your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only					
Signature:					
Date:					
Penalty Applied (if applicable):					

Configuration Manual

Saurabh Shantkumar Yeramwar x20131283

1 Introduction

This document gives all the information related to the environment needed to execute the Pre-processing Techniques for the optimizing Association Rule Mining Algorithm research project. It gives detail understanding about the system configuration needed for the execution with the information about the dataset.

2 System Configuration

The research project needs the special environment setup to execute, this section gives basic hardware and operating system requirements.

2.1 Operating System

Operating system is the software which enables the user to interact with the hardware. Various operating system available each with its own advantages and disadvantages. For this research project stable version of windows 10. Figure 1 shows all details about the operating system on which research project got developed and executed.

Windows specifications

Edition	Windows 10 Home Single Language
Version	21H1
Installed on	20-01-2021
OS build	19043.1348
Experience	Windows Feature Experience Pack 120.2212.3920.0

Figure 1: Windows 10

2.2 Hardware Configuration

Hardware is the physical device on which software embedded. Computer or laptop can be used to execute this project. This research is completed on the laptop with 24 Gb of RAM with high performing Intel i5 11th Gen processor. Figure 2 gives all the important details related hardware, this research project needs basic hardware requirements to execute.

Device name	DESKTOP-IEOQJHN	
Processor	11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz GHz	1.38
Installed RAM	24.0 GB (23.7 GB usable)	
Device ID	5CB0A4E3-C60F-48EC-88F3-5C555EE45DA6	
Product ID	00327-35909-76090-AAOEM	
System type	64-bit operating system, x64-based processor	

Figure 2: Hardware Configuration

3 Development Environment

Python scripting language is getting use to develop this research project. Python is user friendly language which is easy to learn and get used to develop variety of applications. Python documentation ¹ is available for public access with wide community support.

3.1 Anaconda Distribution

There are various ways to install python on the machine, use of anaconda distribution is more preferred. Anaconda distribution ² comes with suite of applications which helps develop the variety of applications. Figure 3 shows the snip of anaconda distribution showing various application which can be used for development. Pycharm which is integrated development environment for python which can be integrated with cloud or GitHub. Pycharm is heavy wright application comparatively which mostly get used for data engginering. Spyder, RStudio, Datalore and other applications are present in the anaconda distribution with their own features.

	IDA.NAVIGATOR						- 0
ne	Applications on base (root)	* Channels					Re
ironments	¢	0	•	\$	\$	0	
ning	0	E	٠ŏ	lab	Jupyter	\circ	
runity	CMD, exe Prompt 0.1.1 Run a cmd.exe terminal with your current	Datalore Online Data Analysis Tool with smart	IBM Watson Studio Cloud	JupyterLab 2.2.6 An extensible environment for interactive	Notebook 6.1.4 Web-based, interactive computing	Powershell Prompt 0.0.1 Run a Powershell terminal with your	
	environment from Navigator activated	coding assistance by JetBreins. Edit and run your Python notebooks in the cloud and share them with your team.	tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data	and reproducible computing, based on the Jupyter Notebook and Architecture.	notebook environment. Edit and run humen-readable docs while describing the date analysis.	current environment from Navigator activated	
	Launch	Launch	science tools or visual modeling.	Launch	Launch	Launch	
	0	¢	•	•	*	¢.	
	PyCharm Community	Ot Console	Souder	Glueviz	Orange 3	PyCharm Professional	
	2020.3.5 An IDE by JetBrains for pure Python development. Supports code completion, listing, and debugging.	4.7.7 PyQt GUI that supports inline Figures, proper multiline editing with syntax highlighting, graphical califips, and more.	4.1.5 Scientific PYthon Development EnviRonment: Powerful Python IDE with advanced editing, interactive testing,	1.0.0 Multidimensional data visualization across files. Explore relationships within and among related datasets.	3.26.0 Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows	A full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL	
your ments in s for free	[units]	Turns	debugging and introspection features		with a large toolbox.	-	
Now	•	Lincitot	Lineitori				
up, port, and environment	R						
ientation	RStudio						
nda Blog	A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.						

Figure 3: Python Distribution

¹https://docs.python.org/3/

²https://www.anaconda.com/

3.2 Python Libraries

When we install Python preloaded libraries are present inside it, which can be used in the development of the script. Libraries are the collection of functions which can be used for similar goals. Libraries helps to reuse the already written code and avoid repetition of the code. If development needs new libraries which are not present in the anaconda we can use the pip command ³ to install them. Mlxtend ⁴ library is used for association rule mining algorithms.

Table 1 gives details about the important libraries which are required for smooth execution the project python code with version.

Python Libraries	Version
apyori	1.1.2
bitmap	0.0.7
conda	4.10.1
ipython	7.19.0
jupyter-client	6.1.7
mlxtend	0.19.0
pandas	1.1.3
numpy	1.19.2
scikit-learn	0.24.2

Table 1: Python Libraries Version

3.3 Jupyter notebook

Jupyter notebook ⁵ is light weight application to develop the python script. Each cell can contain small part of code and can run individually. This can help in developing and debugging the python code. Figure 4 shows the snapshot of the Jupyter notebook post installation.

C ARM_PROJECT/ ×	+					
$\leftarrow \ \rightarrow \ \mathbf{C}$	○ □ localhost:8888,	tree/ARM_PROJECT				
		💭 jupyter		Juit	Logout	
		Files Running Clusters				
		Select items to perform actions on them.	Upl	oad N	lew - O	
		0 - ARM_PROJECT	Name 🔶 🛛 Last Modif	ed	File size	
		D	seconds	ago		
		Capriori	8 days	ago		
		Съвк	6 days	ago		
		C clus_arm_final	a day	ago		
		Chuster	4 days	ago		
		C mbxtend	5 days	ago		
		Comultihash	2 days	ago		
			2 days	ago		
		C stopwords	8 days	ago		
		C x20131283_ARM	a day	ago		

Figure 4: Jupyter Notebook UI

³https://pip.pypa.io/en/stable/

⁴http://rasbt.github.io/mlxtend/

⁵https://jupyter.org/

4 Dataset

Dataset for this research project is available publicly over UCI machine learning archive ⁶ and can be downloaded. Once the file is downloaded, update the path in the code to execute properly. Location of the file needs to be updated as shown in Figure 5. This script internally uses pandas function to consume the input file and convert it into datframe. This dataframe will get further analysed and used in the execution.

In [3]: 🕅	# da da	<pre>f Loading the Data data = pd.read_excel('D:\\SEM_3\\Project\\Datasets\\Online_Retail.xlsx') data.head()</pre>									
Out[3]:		InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country		
	0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom		
	1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom		
	2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom		
	3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom		
	4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom		

Figure 5: Input Data Script

5 Association Rules Mining

Association rules mining is implemented with the help of mlxtend library. Mlxtend provide implementation of apriori and fp growth algorithm which generates association rules. These functions need multiple parameters as inputs including datafarme name, support and other. Figure 6 shows the python implementation of apriori algorithm.

```
In [9]:  # Building the model
start_time = datetime.datetime.now()
frq_items = apriori(basket_France, min_support = 0.05, use_colnames = True)
# Collecting the inferred rules in a dataframe
rules = association_rules(frq_items, metric ="lift", min_threshold = 1)
rules = rules.sort_values(['confidence', 'lift'], ascending =[False, False])
print(rules.head())
end_time = datetime.datetime.now()
```

Figure 6: Operating System

.

5.1 Rules Genrated

The rules generated from apriori algorithm are shown in Figure 7. Each rule have antecedent and consequent items. With various other parameters which can give more details about the rule including support, lift and confidence.

⁶http://archive.ics.uci.edu/ml/machine-learning-databases/00352/

				ant	ecedents \							
45			(JUMBO BAG	WOODLAND	ANIMALS)							
259	(RED TOADSTOOL LED NIGHT LIGHT, PLASTERS IN TI											
270	(PLASTERS IN TIN WOODLAND ANIMALS, RED TOADSTO											
300	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED											
301	(SET/20 R	ED RETROSPOT	PAPER NAP	KINS, SET/	6 RED							
		c	onsequents	antecede	ent support	consequent support	\					
45		-	(POSTAGE)		0.076531	0.765306						
259			(POSTAGE)		0.051020	0.765306						
270			(POSTAGE)		0.053571	0.765306						
300	(SET/6 RE	D SPOTTY PAP	ER PLATES)		0.102041	0.127551						
301	(SET/6 RED SPOTTY PAPER CUPS) 0.102041 0.137755											
	support	confidence	lift	leverage	conviction							
45	0.076531	1.000	1.306667	0.017961	inf							
259	0.051020	1.000	1.306667	0.011974	inf							
270	0.053571	1.000	1.306667	0.012573	inf							
300	0.099490	0.975	7.644000	0.086474	34.897959							
301	0.099490	0.975	7.077778	0.085433	34.489796							

Figure 7: Association Rules

6 Documentation

Technical documentation is important part of any project which helps to understand the project. Documentation is completed with the help of overleaf tool, an online light weight Latex editor. Figure 8 shows the interactive user interface of overleaf.



Figure 8: Overleaf