

Configuration Manual

MSc Research Project Data Analytics

Sachin Muttappanavar Student ID: 20144253

School of Computing National College of Ireland

Supervisor: Dr. Rejwanul Haque

National College of Ireland Project Submission Sheet School of Computing



Student Name:	Sachin Muttappanavar
Student ID:	20144253
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Rejwanul Haque
Submission Due Date:	31/01/2022
Project Title:	Configuration Manual
Word Count:	481
Page Count:	6

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:	Sachin Muttappanavar
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

Sachin Muttappanavar 20144253

1 Introduction

The goal of this documentation is to outline the configuration set up to be followed in the implementation process of research project. We have described the software and hardware prerequisites for re-creating project. It also outlines the coding process and procedure to be followed to execute the code.

2 Scraper Setup

We have used Eclipse integrated development environment to code the automation script using Java Selenium.Following tools, programing language and libraries are used for implementation of scraper:

- JavaSE 1.8v is used for the implementation of the code
- Selenium library v3.141.59 is used for automating the web page
- Jsoup library v1.14.3 is used for parsing the html files
- Eclipse IDE Photon version tool is used for development activities.

Snapshot of the scraper implementation is shown in the Figure 1 To scrape the match commentary, we must execute the project's Scraper.java class, and for reports (news articles), we must execute the Reports.java method.

3 Anaconda Setup

For data processing and model development following tools, programming language and libraries are used.

- Anaconda navigator of version 2.0.3 for web based interactive computing Jupyter Notebook v6.4.0(Figure 2) to implement.
- Python version 3.8.5 was installed on Anaconda Navigator.
- Libraries required for data processing code are pandas, numpy, nltk, os, re
- Libraries required for executing the BART model are pandas, numpy, simpletransformers, rouge, matplotlib.

🔘 eclipse-workspace - test/src/main/java/te	st/Scraper.java - Eclipse IDE	
File Edit Source Refactor Navigate Se	arch Project Run Window Help	
📬 🕶 🔛 🚳 🗢 💷 🔌 🗠 📖 😣		Quick Access 🔡 😫 😻
🛤 Package Explorer 😂 🛛 🖻 😒 👻 🌣	Carperjava 🛙 D ParseHtmljava	- 0 - 0
> 😂 Interview	1 package test;	^ = 🗈
> 😂 Practice	2	- e .
✓	% 3*import java.io.BufferedWriter;□	B.
✓	23	
✓ J test	24 public class scraper {	
> 🔎 Commentary.java	25 public static (bromeDriver driver)	*
> 🔊 ParseCommentaryHtml.java	27 public static int i:	
> III ParseHtmLiava	28	
> ParsereportHTMLiava	299 public static void driverInitialization() {	
> 🔊 Reportiava	30 System.setProperty("webdriver.chrome.driver",	
> D Scraper, java	31 "C:\\Users\\sachi\Downloads\\NCI\\CricketNews\\New folder\\chromedriver.exe");	
src/main/resources	32 ChromeOptions options = new ChromeOptions();	
(# src/test/iava	33 options.addAnguments(start-maximized); 34 options.addAnguments("	
estimate and the second sec	35 driver - new (homologications);	
IRE System Library (JavaSE-1.8)	<pre>36 driver.manage().timeouts().imlicitlyWait(10.TimeUnit.SECONDS);</pre>	
Mayen Dependencies	37	
Beferenced Libraries	38 }	
htmlFiles	39	
> Commentany	40° public static void scrape(int i, int count) throws InterruptedException, IOException {	
> D mater_commentary	41 // First Innings	
> sic	42 driven findElement/By ymath("//div[contains(@class 'ci_dd_menu')]/decondent::u]/li["+i+"]/conn")) click():	
Diarget	AA	
Commentary_data_z.txt	45 JavascriptExecutor js = (JavascriptExecutor) driver;	
Commentary_data.txt	<pre>46 js.executeScript("window.scrollBy(0,2000)");</pre>	
Commentary_data1.ntm	<pre>47 js.executeScript("window.scrollBy(0,2000)");</pre>	
Commentary_data1.0t	48 Thread.sLeep(1000);	
Commentary_data10.txt	49 js.executeScript("window.scrollBy(0,2000)");	
commentary_data100.ntml	50 Inread.steep(1000);	
commentary_data11.txt	52 Thread science (1980): 51 - 51 - 51 - 51 - 51 - 51 - 51 - 51	
Commentary_data 12.txt	<pre>53 is.executeScipt("window.scroll8v(0.2000)"):</pre>	
commentary_data13.txt	54 Thread.sLeep(1000);	
commentary_data14.txt	55 js.executeScript("window.scrollBy(0,2000)");	
commentary_data15.txt	56 Thread.sLeep(1000);	
commentary_data16.txt	57 js.executeScript("window.scrollBy(0,2000)");	
iii commentary_data17.txt	58 Ihread.steep(1000);	
commentary_data18.txt	59 JS.EXECUTESCIPT(MINON.SCIOLES(0,2000)); Dread cf.exe(1990).	
commentary_data19.txt	60 In course cop(1000), 61 is executed scrollBv(0.2000)).	
commentary_data2.html	4 62 Thread.sleep(1990):	~
<		>
	Writable Smart Insert 58:5	1 0

Figure 1: Scraper Implementation

- Libraries required for executing the T5base model are pandas, numpy,rouge, pytorch, sentencepiece, transformers, CUDA, tqdm, torch.
- For Named Entity Recognizeer, satudford NER jars 4.2.0v is used. To run NamedEntity_MaskingPlayers, NER jars needs to installed path of the jars needs to be updated.

4 System Configuration

For implementing the pre-trained models, we leveraged 'Google Colab'(Figure 9) of Google cloud platform. GPU runtime environment was set to execute and train the models as GPU can process numerous computations parallelly. We have used google colab GPU computational power for running executing computational intensive activities. Another advantage of utilising Google Colab is that it gives a simple method to connect to a drive where data may be saved. Local system configuration is shown in the Figure 4.

Steps to follow to connect to GPU: Runtime - Change runtime type - select GPU under hardware accelerator

5 Libraries Used

Liraries and Modules used of the data Processing shown in the Figure 6. Commentary data processing code can be found with name commentary_processing.ipynb and Reports cleaning code is in Report_Processing.ipynb. We can find the Named recognization code in NamedEntity_MaskingPlayers.ipynb. Also, libraries used in the building BART and T5base are shown in Figure 7 and Figure 5 respectively.

	Applications on Project	 Channels 				
nents	*	*	¢	, °	i i i	¢
ity	DataTore Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebools in the cloud and share them with your team.	IBM Watson Studio Cloud IBM Watson Studio Cloud provides you the 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	Notebook A 62.0 Web-based, interactive computing notebook environment. Edit and run human-readble docs while describing the deta analysis.	CMD.exe Prompt 0.1.1 Run a cmd.exe terminal with your current environment from Nevigator activated	Glueviz 1.0.0 Multidimensional data visualization across Mes. Explore relationships within and among related datasets.	JupyterLab 3.2.1 An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
	Launch	science tools or visual modeling. Launch	Launch	Install	install 🖉	install 🔷
		Powershall Promot	Pu/Charm Professional		R	Sourier
DA.	3.26.0 Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.	0.0.1 Run a Powershell terminal with your current environment from Navigator activated	A full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL	5.1.1 PyQt CUI that supports inline figures, proper multiline editing with syntax highlighting, graphical colltips, and more.	1.1.436 A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.	5.1.5 Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
ar free	Install	Install	Install	Install	Install	Install
port, and vironment						
ation						

Figure 2: Anaconda Navigator



Figure 3: Google Colab

Device specifications

Device name	Sach
Processor	Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.60 GHz
Installed RAM	16.0 GB (15.8 GB usable)
Device ID	4A66B777-A5B0-4536-9DEE-6D2793126972
Product ID	00327-35900-64630-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 4: Local Machine Configuration

import pandas as pd import numpy as np import nltk import re nltk.download('stopwords') nltk.download('punkt') import os.path from nltk.tag import StanfordNERTagger from nltk.tokenize import word_tokenize executed in 473ms, finished 19:06:06 2021-12-14





Figure 6: Libraries and modules used for T5Base



Figure 7: Libraries and modules used for BART

6 Models Parameters

This section provides information about the parameters defined and values assigned for them in building model. Figure 7 illustrates the BART model parameters and Figure 7 illustrates T5 base model parameters. BART model implementation is given in BART.ipynb and T5 model implementation is in T5.ipynb.



Figure 8: BART model

7 Dataset

Dataset used for this research study are provided under the Dataset folder. Dataset without masking is in file data2_withoutMask.xlsx and masked data in Data_2.xlsx.



Figure 9: T5 model