

# Configuration Manual

MSc Research Project  
Data Analytics

Bryan O'Donohoe  
Student ID: x20212828

School of Computing  
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Bryan O'Donohoe
<b>Student ID:</b>	x20212828
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2022
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Jorge Basilio
<b>Submission Due Date:</b>	15/08/2022
<b>Project Title:</b>	Configuration Manual
<b>Word Count:</b>	XXX
<b>Page Count:</b>	2

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.

**ALL** 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.

<b>Signature:</b>	
<b>Date:</b>	13th August 2022

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

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

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Configuration Manual

Bryan O'Donohoe  
x20212828

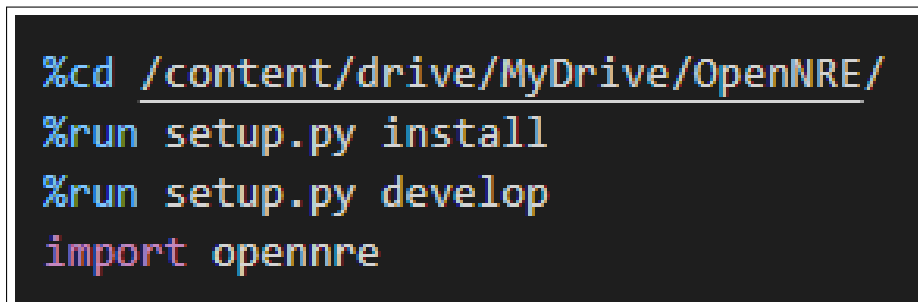
## 1 Introduction

This configuration manual is comprised of the specifications for the software and hardware deployed in this research. It will outline step by step the process that was used in cleaning, transforming, modelling, validating and deploying in this project. Any Python notebook can be utilised for this project, however Kaggle and Google Colab were chosen in this research for speed and convenience.

## 2 Data preparation

The dataset was downloaded using the OpenNRE (Han et al.; 2019) package which can be found at the following link: [OpenNREpackage](#)

The dataset is contained within this package. It is installed using the commands outlined below in figure 1

A terminal window with a black background and white text. The text shows a sequence of commands for installing the OpenNRE package. The first line is a directory change command, the next two are Python script execution commands, and the last is a Python import statement.

```
%cd /content/drive/MyDrive/OpenNRE/  
%run setup.py install  
%run setup.py develop  
import opennre
```

Figure 1: OpenNRE package install commands

## 3 Installation

The necessary packages need to be installed. For the exploratory analysis the packages need to be installed are shown below in figure 2

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
import json
import re
from nltk.tokenize.treebank import TreebankWordDetokenizer as Detok
import spacy
from spacy import displacy
from sklearn.feature_extraction.text import CountVectorizer
from tqdm.notebook import tqdm
import matplotlib.pyplot as plt
```

Figure 2: Package install commands

## 4 Exploratory Data Analysis

The exploratory data analysis notebook can be then fully run as it is self contained once the packages are correctly installed.

## 5 Model building

The model building notebook can be fully run with either the BERT or CNN section of code called. The output of the validation is then output into a csv file for analysis

## 6 Validation

The validation was carried out in Microsoft Excel. The 3 validation notebooks are named "Validation analysis.xlsx", "Validation analysiscnn.xlsx" and "Validation analysisbert.xlsx"

## 7 Application

The final application that was deployed is contained in "../OPENRE/OpenNRE/app.py"

This was then placed on the AWS EC2 engine. To start the application, in the root directory of the project, run the following command shown below in Table 1:

```
ssh -i "x20212828_16GB.pem" ubuntu@ec2-34-242-186-194.eu-west-1.compute.amazonaws.com
```

Table 1: Command

## References

- Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z. and Sun, M. (2019). OpenNRE: An open and extensible toolkit for neural relation extraction, *Proceedings of EMNLP-IJCNLP: System Demonstrations*, pp. 169–174.  
URL: <https://www.aclweb.org/anthology/D19-3029>