National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Viktor Avgustin
Student ID: x20141432

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Viktor Avgustin
Student ID:	x20141432
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15/08/2022
Project Title:	Configuration Manual
Word Count:	2456
Page Count:	3

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.

Signature:	Viktor Avgustin
Date:	12th August 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

Viktor Avgustin
x20141432

1 Configuration description

This section describes the datasets and the code used in this research paper.

1.1 Datasets

There are four datasets which are used in this research project.

- UpdatedResumeDataSet.csv - this the labelled Kaggle dataset
- dataset.csv - Un-labelled scraped dataset
- NewResumes.csv - Labelled scraped dataset
- Job-Data.csv - list of job descriptions

The Output.csv file contains the scraped labelled dataset and the predictions from the three classifiers.

1.2 Software Guide

The research was primarily conducted in Python with the use of the PyCharm software. Docker was used to run containers for the linkedin resume crawling. Chromedriver is also used in the crawling/scraping of resumes.

1.3 Code Guide

- /chromedriver/chromedriver.exe - this extension is used in the scraping of LinkedIn profiles
- /linkedin folder contains LinkedIn crawler [1]
- conf.py - contains configuration file for the crawler
- /features/tfidf.py - contains the code for the TF-IDF vectorizer
- /features/word2vec.wordvectors - contains the word2vec vectors
- /models/models.py - contains the KNN and Linear SVM models

[1]Source: https://github.com/eracle/linkedin

- /models/cnn.py - contains the CNN model
- /models/job-title-norm-models - the folder contains the node2vec models for the job title normalization exercise
- gender.py - code for gender extraction from CVs
- google-urls.py - code to search google for linkedin profile which can be subsequently fed into the LinkedIn crawler for scraping.
- job-title-norm.py - used for job title normalization of the Un-labelled scrapped dataset [2]
- Scraper v1.py - a manual scraper used to scrap data from postjobsfree.com. Search words are manually input. There is no restriction to ensure that the returned data contains only the search string (e.g. there may be job titles which do not match the search string).
- Scraper v2.py - an automated scraper which was used to create the Labelled scraped dataset. This scraper ensures that the search only brings back the correctly labelled results and it automatically searches the list of labels.
- similarity.py - used for matching resumes and job descriptions on the unlabelled dataset
- /preprocessing/Preprocessing.py - used for cleaning/tokenization of Un-labelled dataset
- /preprocessing/Preprocessing-labeled.py - used for cleaning/tokenization of Labelled dataset
- DA-project.ipynb - notebook used in the Code Demo video.

2 Required Libraries

A list of the required python libraries is saved in /requirements/production.txt The following libraries are used in the research paper:

Scrapy ==v2.6.1- used in the scraping of data and data collection
Selenium ==3.14.0- used in the scraping of data
beatifulsoup ==4.9.2 - used for data collection

NLTK - used for pre-processing of the text data, including tokenization
Numpy - data manipulation
Pandas - data maniupulation
Spacy - used in the pre-processing
Gensim - used in the vectorization
Re - used for pre-processing in cleaning of the collected data

[2]Source: https://www.kaggle.com/code/estasney/exploring-job-titles-node2vec/notebook

matplotlib - for some of the plots
seaborn - for some of the plots
Sklearn - used for train/testing split, TF-IDF and Word2vec, in addition to some of the classifiers

Keras - used for Machine Learning (neural networks)
TensorFlow - used for the CNN

Some transformations, graphs, figures and statistical analysis were performed in R. The log in R is saved in the r-script.txt file.

3 Parameters

Parameters used for configuration of the embeddings and model building.
The following parameters are used in the CNN classifier
MAX-SEQUENCE-LENGTH= 100
MAX-VOCAB-SIZE =20000
EMBEDDING-DIM=300
VALIDATION-SPLIT =0.2
BATCH-SIZE =128
EPOCHS =10