Configuration Manual

MSc Research Project
Data Analytics

Soumyadip Dipak Ghosh
Student ID: X18192181

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Soumyadip Dipak Ghosh		
Student ID:	X18192181		
Programme:	MSc in Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Lecturer:	Dr. Catherine Mulwa		
Submission Due Date:	17 August, 2020		
Project Title:	Text Classification using Graph Based Learning		
Word Count: 240		**Page Count:** 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.

Signature: Soumyadip Dipak Ghosh

Date: 17 August, 2020

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

Soumyadip Dipak Ghosh
X18192181

1 Prerequisite Configuration

1.1 Python Setup

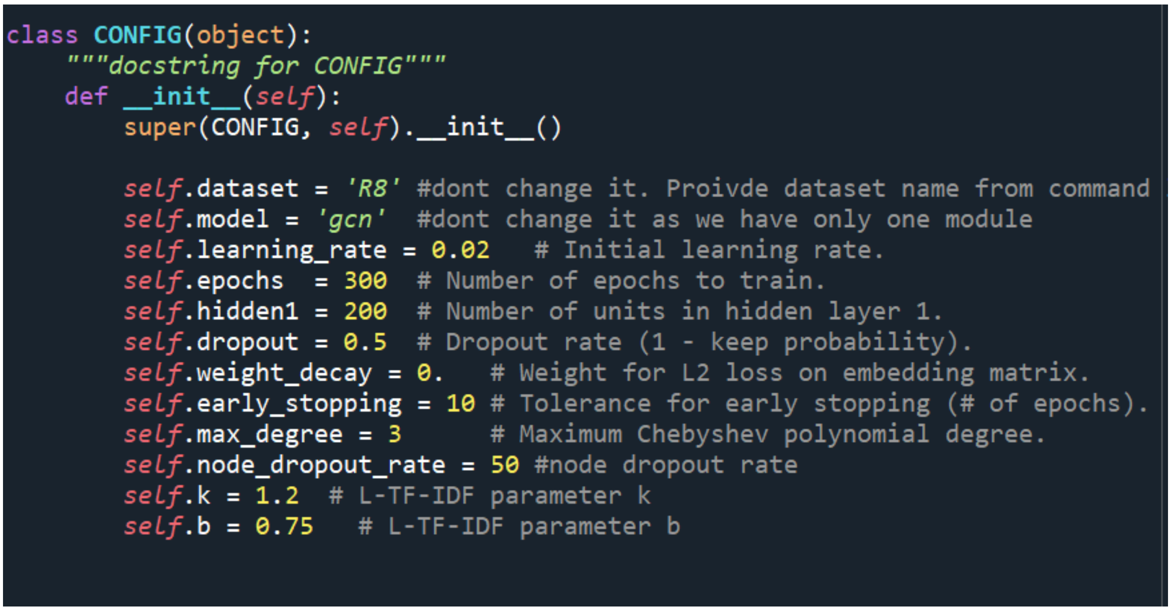
- Install python on your system. Minimum required version 3.7.x. Download it from: https://www.python.org/downloads/
- For Instructions on how to download and install it on system visit the link: https://docs.python.org/3/using/index.html

1.2 Installing required Packages

- NLTK - https://www.nltk.org/install.html
- Numpy - https://numpy.org/install/
- Scipy - https://www.scipy.org/install.html
- Pickle – installed by default in python 3.x
- Sklearn - https://scikit-learn.org/stable/install.html
- Pytorch - https://pytorch.org/
  - Required build above 1.5.x
  - If system has CUDA enabled GPU, select correct CUDA version. I
  - ncase of any CUDA related errors, install CPU only version.
- Networkx - https://pypi.org/project/networkx/

2 Setting up configuration file

- Configurations have been set up and values have been set at default values required for the project.
- Configuration file name: config.py
- Location: root directory of project
- Proceed without making changes to run project at default values.



```
class CONFIG(object):
    """docstring for CONFIG"""
    def __init__(self):
        super(CONFIG, self).__init__()

        self.dataset = 'R8' #dont change it. Proivde dataset name from command
        self.model = 'gcn'  #dont change it as we have only one module
        self.learning_rate = 0.02   # Initial learning rate.
        self.epochs  = 300  # Number of epochs to train.
        self.hidden1 = 200  # Number of units in hidden layer 1.
        self.dropout = 0.5  # Dropout rate (1 - keep probability).
        self.weight_decay = 0.   # Weight for L2 loss on embedding matrix.
        self.early_stopping = 10 # Tolerance for early stopping (# of epochs).
        self.max_degree = 3      # Maximum Chebyshev polynomial degree.
        self.node_dropout_rate = 50 #node dropout rate
        self.k = 1.2  # L-TF-IDF parameter k
        self.b = 0.75   # L-TF-IDF parameter b
```

3 Code Running Sequence

- Navigate to root directory of the project folder. There are 4 datasets to choose from.
- Replace <dataset> with ‘R8’, ‘R52’, ‘ohsumed’ or ‘mr’ depending upon the dataset you want to train on.
- No need to type ‘python’ in following commands, if running from Anaconda prompt.
- Then run the following commands:
  - cd preprocess
  - python remove_words.py <dataset>
  - python build_graph.py <dataset>
  - cd..
  - python train.py <dataset>

- After training is completed, the best model will save in the root folder with name ‘<dataset>_model.pt’.

Note: If dataset has to be changed, then follow the process from the 1st step.