National
College of
Ireland

Configuration Manual

MSc Research Project
Data Analytics

Sufal Addya
Student ID: X18180825

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sufal Addya

Student ID: X18180825

Programme: Data Analytics **Year:** 2020

Module: Research Project

Lecturer: Christian Horn

Submission Due Date: 28/09/2020

Project Title: Extractive text summarization from images

Word Count: 671 **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:

Date: 28/09/2020

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

Sufal Addya
Student ID: x18180825

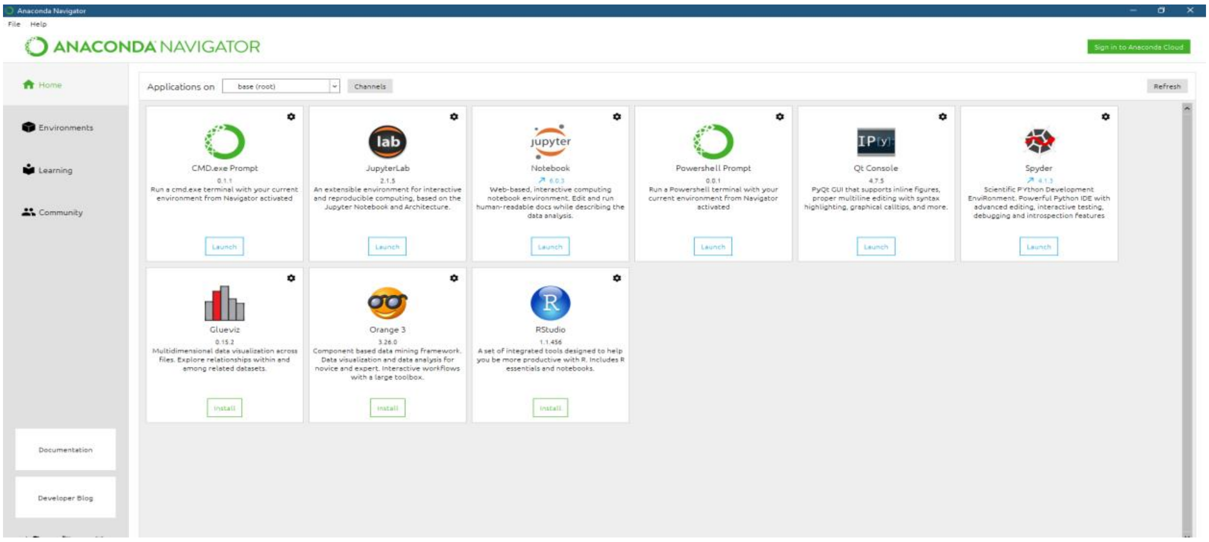
1 Introduction

The project is based on image processing and text analytics. The project is trying to develop a unique pipeline by using optical character recognition, text pre-processing and implementation of text summarization algorithms to find an extractive summary from given image. The project is totally based on python programming language.

2 Installation of IDE

The project is implemented on the Anaconda Spyder (3.7) IDE. Python programming language should be installed to run this project. Anaconda can be installed from given link below,

Link: https://www.anaconda.com/products/individual



CMD.exe Prompt
0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated

JupyterLab
2.1.5
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook
↗ 6.0.3
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Powershell Prompt
0.0.1
Run a Powershell terminal with your current environment from Navigator activated

Qt Console
4.7.5
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

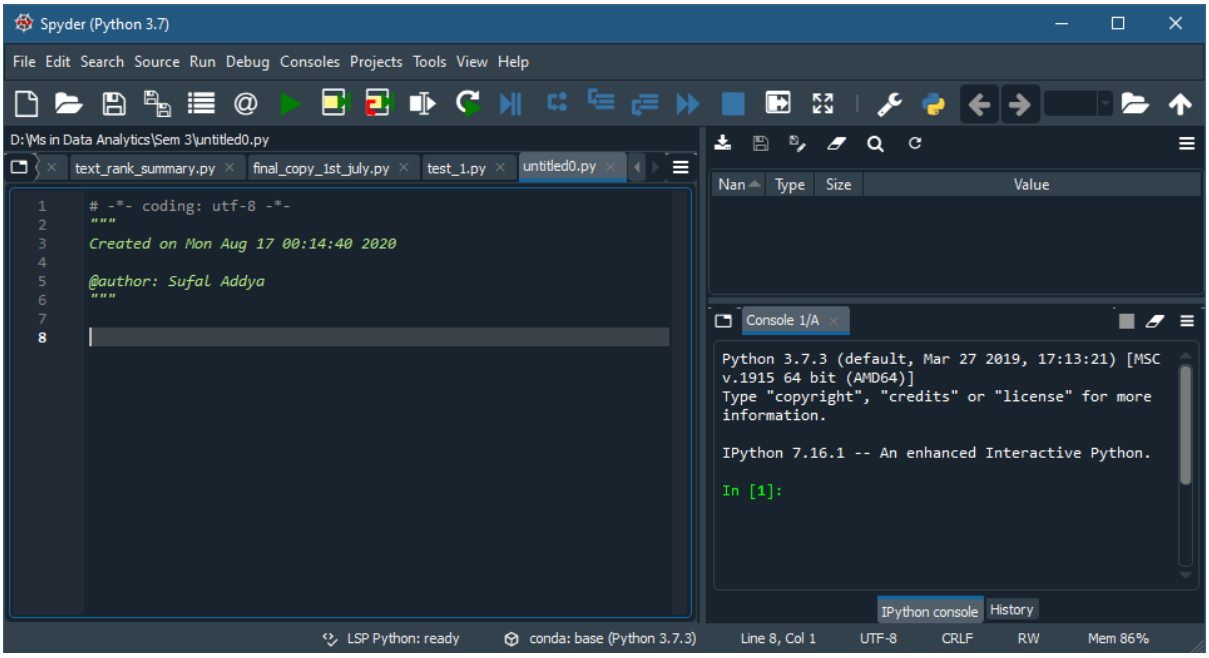
Spyder
↗ 4.1.3
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz
0.15.2
Multidimensional data visualization across files. Explore relationships within and among related datasets.

Orange 3
3.26.0
Component based data mining Framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

After installing Anaconda navigator, any of the IDE can be used to run the code. Here the research project is implemented on the Spyder (3.7).



```
# -*- coding: utf-8 -*-
"""
Created on Mon Aug 17 00:14:40 2020

@author: Sufal Addya
"""

```

```
Python 3.7.3 (default, Mar 27 2019, 17:13:21) [MSC
v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more
information.

IPython 7.16.1 -- An enhanced Interactive Python.

In [1]:
```

3 Configuring OCR

The optical character recognition is implemented by using PyTesseract and OpenCV. So, need to install Tesseract engine in the python, to install the Tesseract engine need to open Anaconda command prompt and can install with following command

```
$ pip install pillow
$ pip install pytesseract
$ pip install opencv-python
```

After installing this, need to set the path to run the PyTesseract. project can follow the uploaded code file for further processing.

4 Configuring Text pre-processing

After installing Tesseract and OpenCV, need to install Regular expression (Re) and Natural language toolkit (NLTK) libraries for text pre-processing. NLTK library can be installed by following commands

```
$ pip install nltk
```

5 Configuring Text summarization algorithms and others

Here two unsupervised extractive text summarization algorithms implemented to find summary of given images.

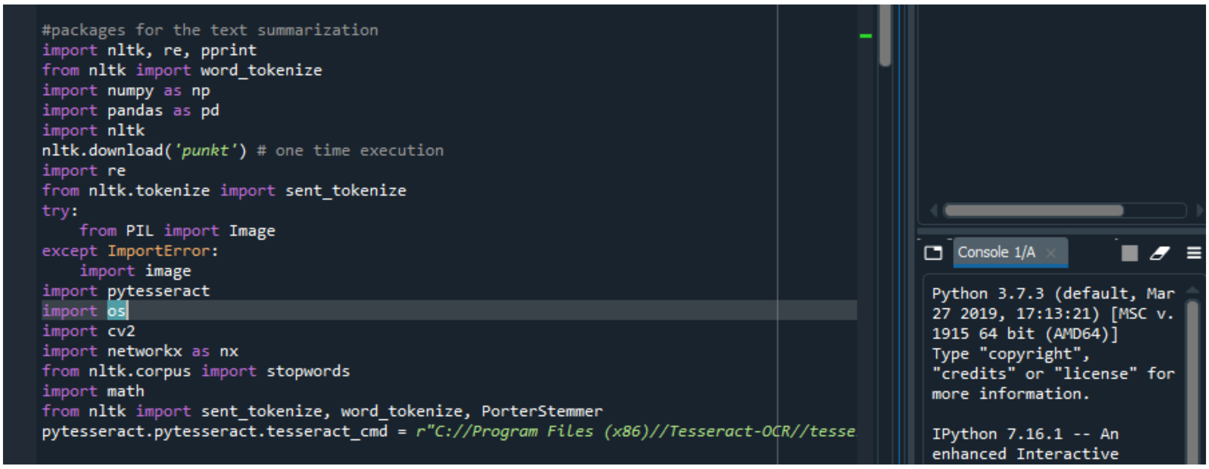
Text rank algorithm: In this algorithm, need to install glove word embedding file to implement Textrank algorithm. Glove can be download from given link

Link: https://nlp.stanford.edu/projects/glove/

Also, need to install networkx for making graph and implementing Textrank algorithm by following command

```
$ pip install networkx
```

Now, need to import all the packages in the IDE, providing a screenshot for that.



```
#packages for the text summarization
import nltk, re, pprint
from nltk import word_tokenize
import numpy as np
import pandas as pd
import nltk
nltk.download('punkt') # one time execution
import re
from nltk.tokenize import sent_tokenize
try:
    from PIL import Image
except ImportError:
    import image
import pytesseract
import os
import cv2
import networkx as nx
from nltk.corpus import stopwords
import math
from nltk import sent_tokenize, word_tokenize, PorterStemmer
pytesseract.pytesseract.tesseract_cmd = r"C://Program Files (x86)//Tesseract-OCR//tesse
```

```
Python 3.7.3 (default, Mar
27 2019, 17:13:21) [MSC v.
1915 64 bit (AMD64)]
Type "copyright",
"credits" or "license" for
more information.

IPython 7.16.1 -- An
enhanced Interactive
```

After installing all the packages into the python, need to import all the library and need to set the default path for getting the image as an input to execute the code. Then should run the given code to get summarization.

This project is based on pipeline architecture so that, here need to input an screenshot or captured image (jpg or png any format) of any news or any story to execute the total pipeline of the project and finally will get a extractive summary of that image.

References

https://www.google.com/search?client=firefox-b-d&q=pytesseract+download