A PROJECT REPORT ON
AUTOMATIC ADDRESS TRACER

UNDER THE GUIDENCE OF MR. CHANDAN KUMAR

Submitted In Partial Fulfillment of the Requirements for the Award Of the degree of

BACHELOR OF TECHNOLOGY

(INFORMATION TECHNOLOGY)

SUBMITTED TO:	SUBMITTED BY:
Mr. Nishant Singh	Divya Prakash Yadav (DT20173826226)
Mr. Chandan Kumar	Akhil Jain (CT20172177061)
	Nandan Goel (CT20172178653)
	(INSTITUTE OF ENGG AND TECHNOLOGY SITAPUR ROAD LUCKNOW)

TCS REMOTE INTERNSHIP 2018

OUTLINE

CONTENT **PAGE NO.**

1. **INTRODUCTION………………………………………………………………….3**
2. **TECHNOLOGIES USED……………………………………………………….4**
   - **IDE……………………………………………………………………………4**
   - **DATABASE…………………………………………………………………..5**
   - **FRONTEND…………………………………………………………………..5**
   - **BACKEND…………………………………………………………………. ...5**
3. **MODULE'S USED………………………………………………………...…..6**
   - **CSV MODULE…………………………………………………………….....6**
   - **NUMPY MODULE……………………………………………………………6**
   - **PYTESSERACT………………………………………………………………7**
   - **PIL MODULE…………………………………………………………………7**
   - **TKINTER MODULE………………………………………………………….8**
   - **OS MODULE………………………………………………………………….8**
   - **OPENPYXL MODULE……………………………………………………….9**
   - **RE MODULE…………………………………………………………………9**
4. **DATA FLOW DIAGRAM…………………………………………………….11**
5. **TEST CASES & SCREENSHOT…………………………………………….12**
6. **FUTURE ENHANCEMENTS………………………………………………..15**
7. **SOURCES……………………………………………………………………...16**

Introduction

Automatic Address Tracer is an application to automate the task to find the address manually using pincode. The application basically takes an image of post card as input containing postal address and provide details like **City name, State name, District name** and **list of post offices** by detecting **pincode** in the image.

This application is built in **Python** programming language and uses techniques like **Image processing** and **Machine learning** to train the alphabetical and numerical characters and builds a model to recognize text and filters out pincode from a bunch of text.

The front end is designed using **tkinter** module. The design consists of a simple and user friendly GUI. The home page contains instructions about the application, upload button and a submit button. The address details are displayed whenever user uploads and submits the image of postcard.

In the back end, using pytesseract module the application recognizes text image. The text is extracted and sent as a multiline string to "**extract.py**" module which uses regex module to separate out pincode. This pincode is then used to look for details in "**address_directory.csv**" using csv module. The details are returned as a python list which is finally displayed by application.

Technologies Used

➢ IDE

Pycharm

PyCharm is one of the most widely used IDEs for Python programming language. At present, the Python IDE is being used by large enterprises like Twitter, Pinterest, HP, Symantec and Groupon.JetBrains has developed PyCharm as a cross-platform IDE for Python. In addition to supporting versions 2.x and 3.x of Python, PyCharm is also compatible with Windows, Linux, and macOS. At the same time, the tools and features provided by PyCharm help programmers to write a variety of software applications in Python quickly and efficiently.

Sublime Text Editor

Sublime has highly customizable build systems that can add to your productivity if you learn how to use them to your advantage. You can define one for your project and whenever you are editing any file, you can run certain commands on the source file and see the output in the sublime console, without leaving the editor.

➢ Database

Database is present as a **CSV file** which contains **All India pincode** directory. The Pincode extracted from the Post card image is matched with pincode present in the CSV file and complete information about the address is returned. This CSV file is imported using CSV module of Python.

➢ Front End

Tkinter

Python offers multiple options for developing GUI (Graphical User Interface). Out of all the GUI methods, tkinter is most commonly used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the fastest and easiest way to create the GUI applications.

➢ Backend

The backend makes uses of image processing to reduce noise in incoming input image, convert it into a binary image and other optimizations. Later, machine learning is used to train model for recognizing text from image. Apart from these technologies the backend comprises of various python modules such as csv, re, pytesseract, os, PIL, opencv , etc in general to perform various tasks

MODULES USED

➢ CSV Module

The so-called **CSV** (Comma Separated Values) format is the most common import and export format for spreadsheets and databases. There is no “CSV standard”, so the format is operationally defined by the many applications which read and write it. The lack of a standard means that subtle differences often exist in the data produced and consumed by different applications. These differences can make it annoying to process CSV files from multiple sources. Still, while the delimiters and quoting characters vary, the overall format is similar enough that it is possible to write a single module which can efficiently manipulate such data, hiding the details of reading and writing the data from the programmer.

➢ NumPy Module

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

➢ Python-tesseract Module

Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images.

Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Python Imaging Library, including jpeg, png, gif, bmp, tiff, and others, whereas tesseract-ocr by default only supports tiff and bmp. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.

➢ PIL Module

Python Imaging Library (abbreviated as **PIL**) (in newer versions as Pillow) is a free library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats. It is available for

Windows, Mac OS X and Linux. The latest version of PIL is 1.1.7, was released in September 2009 and supports Python 1.5.2–2.7, with Python 3 support to be released "later". Some of the file formats supported are PPM, PNG, JPEG, GIF, TIFF, and BMP. It is also possible to create new file decoders to expand the library of file formats accessible.

➢ Tkinter Module

The **Tkinter** module (“Tk interface”) is the standard Python interface to the Tk GUI toolkit. Both Tk and Tkinter are available on most Unix platforms, as well as on Windows systems. (Tk itself is not part of Python; it is maintained at ActiveState.)

Running python -m Tkinter from the command line should open a window demonstrating a simple Tk interface, letting you know that Tkinter is properly installed on your system, and also showing what version of Tcl/Tk is installed, so you can read the Tcl/Tk documentation specific to that version.

➢ OS Module

The OS module in Python provides a way of using operating system dependent functionality.

The functions that the OS module provides allows you to interface with the underlying operating system that Python is running on – be that Windows, Mac or Linux.

The OS module in python provides functions for interacting with the operating system. OS, comes under Python’s standard utility modules. This module provides a portable way of using operating system dependent functionality. The *os* and *os.path* modules include many functions to interact with the file system.

➢ Openpyxl Module

Openpyxl is a Python library for reading and writing Excel 2010 xlsx/xlsm/xltx/xltm files.

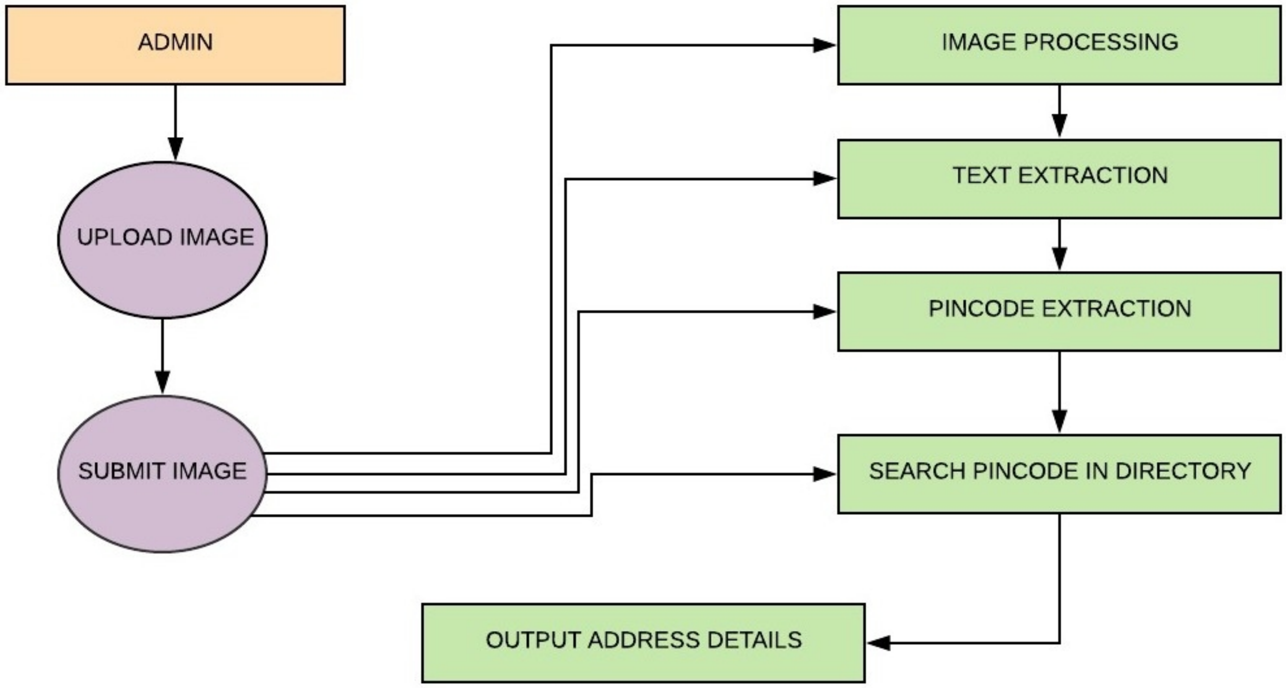
It was born from lack of existing library to read/write natively from Python the Office Open XML format.

➢ RE Module

A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing)

Regular expressions use the backslash character (`'\'`) to indicate special forms or to allow special characters to be used without invoking their special meaning. This collides with Python’s usage of the same character for the same purpose in string literals; for example, to match a literal backslash, one might have to write `'\\\\'` as the pattern string, because the regular expression must be `\\`, and each backslash must be expressed as `\\` inside a regular Python string literal.

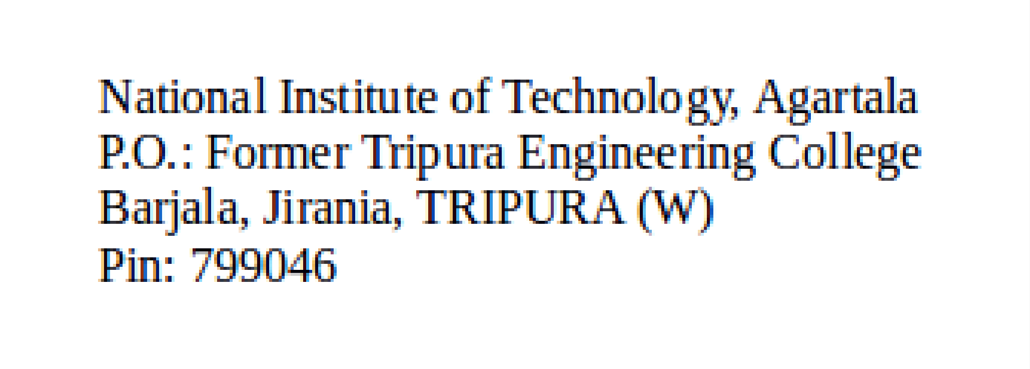
***Data Flow Diagram* (DFD)**



Test Cases & Screenshots

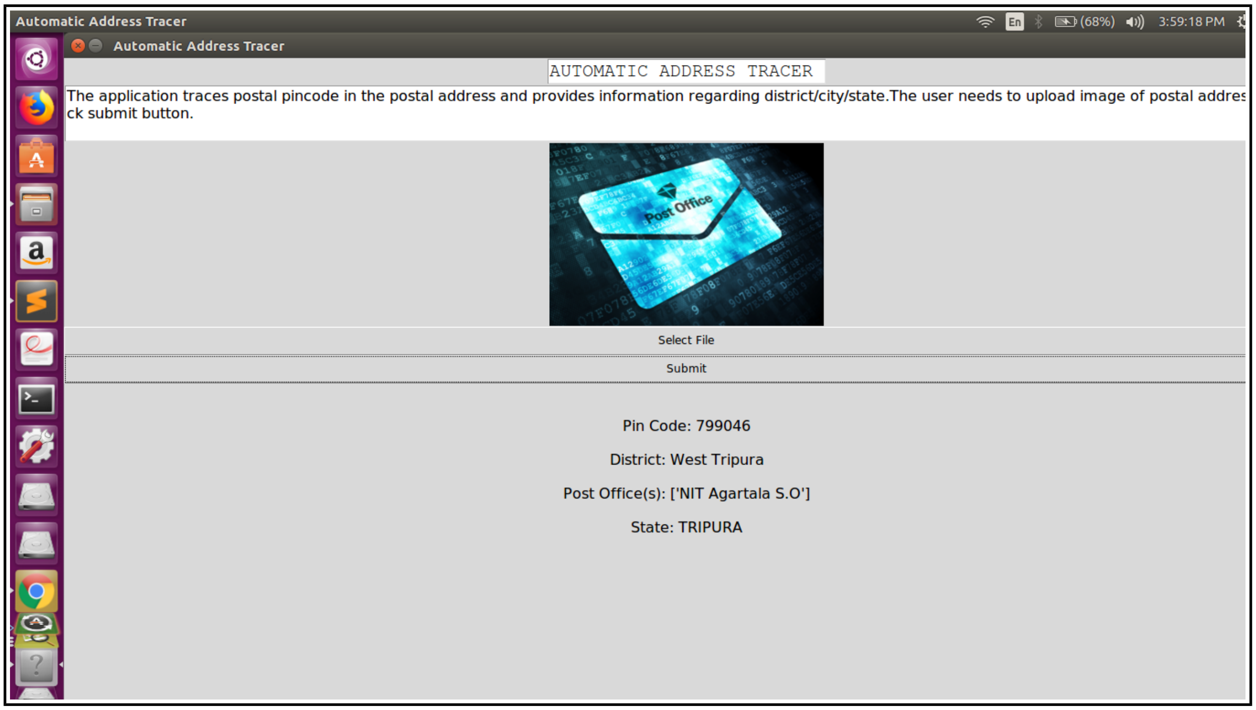
➢ ***EXAMPLE 1***

INPUT



National Institute of Technology, Agartala
P.O.: Former Tripura Engineering College
Barjala, Jirania, TRIPURA (W)
Pin: 799046

OUTPUT



AUTOMATIC ADDRESS TRACER

The application traces postal pincode in the postal address and provides information regarding district/city/state.The user needs to upload image of postal addres
ck submit button.

Pin Code: 799046

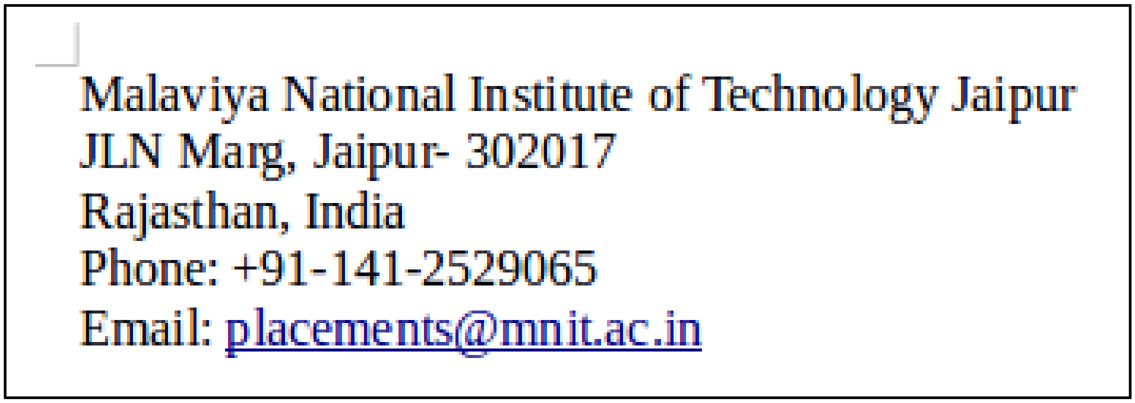
District: West Tripura

Post Office(s): ['NIT Agartala S.O']

State: TRIPURA

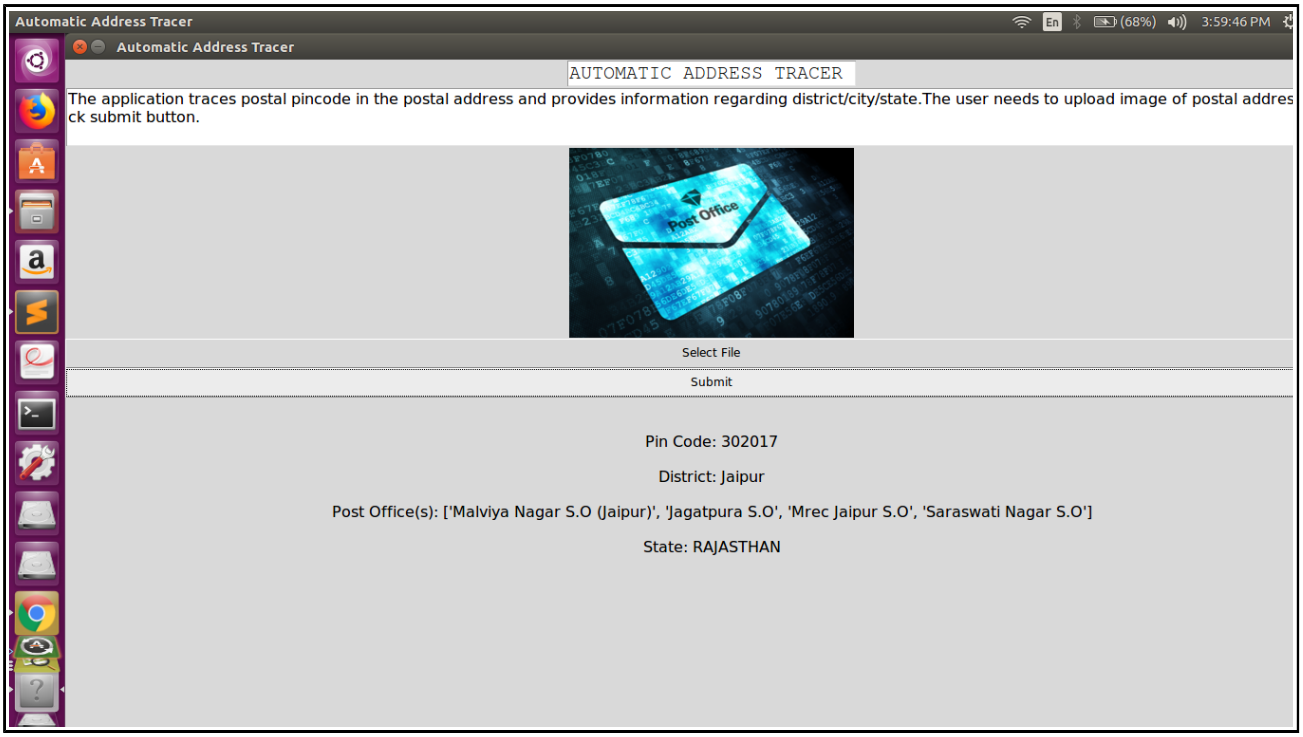
- ## *EXAMPLE 2*

INPUT



Malaviya National Institute of Technology Jaipur
JLN Marg, Jaipur- 302017
Rajasthan, India
Phone: +91-141-2529065
Email: placements@mnit.ac.in

OUTPUT



AUTOMATIC ADDRESS TRACER

The application traces postal pincode in the postal address and provides information regarding district/city/state.The user needs to upload image of postal addres
ck submit button.

Pin Code: 302017

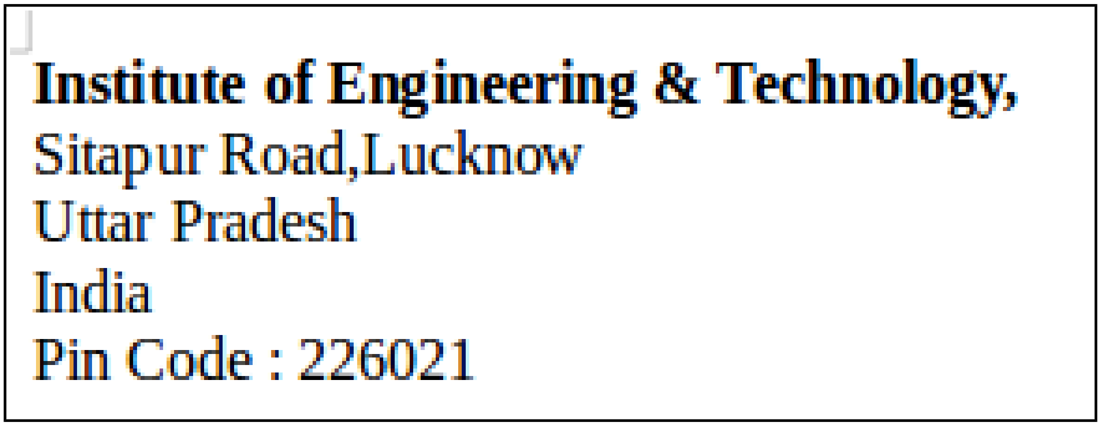
District: Jaipur

Post Office(s): ['Malviya Nagar S.O (Jaipur)', 'Jagatpura S.O', 'Mrec Jaipur S.O', 'Saraswati Nagar S.O']

State: RAJASTHAN

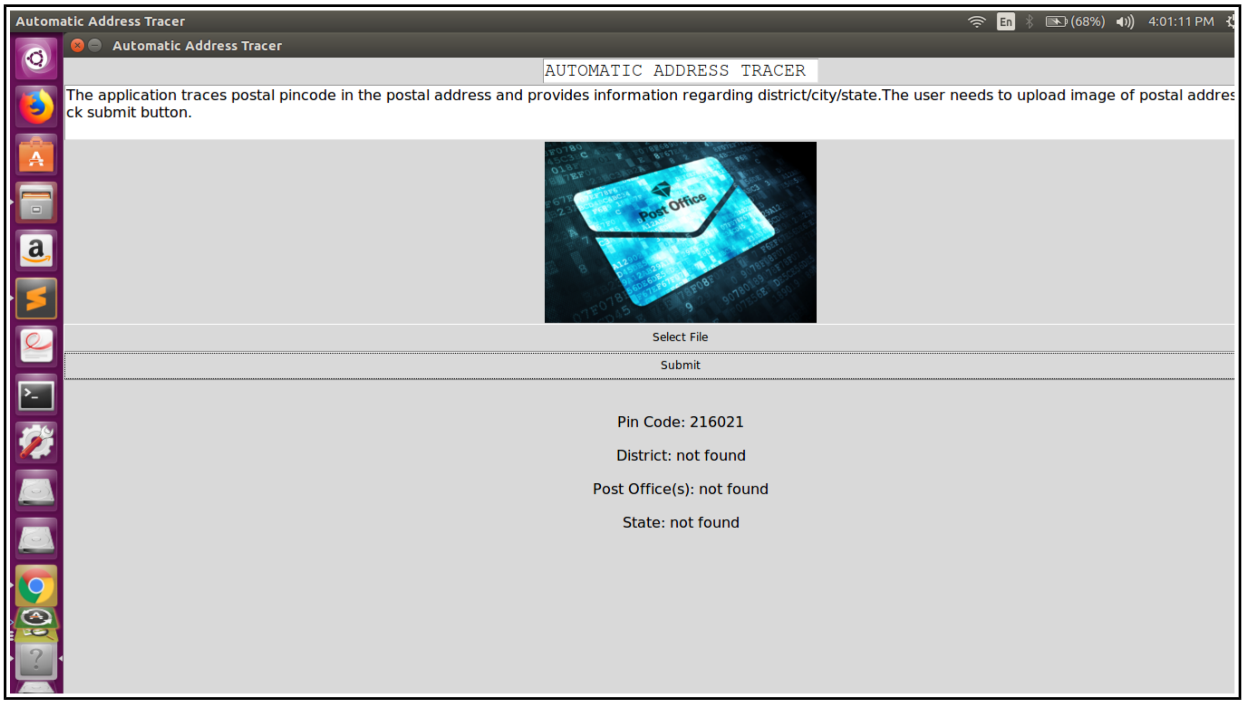
- ## *EXAMPLE 3 (Errored)*

INPUT



Institute of Engineering & Technology,
Sitapur Road,Lucknow
Uttar Pradesh
India
Pin Code : 226021

OUTPUT



The application traces postal pincode in the postal address and provides information regarding district/city/state.The user needs to upload image of postal addres
ck submit button.

```
Pin Code: 216021

District: not found

Post Office(s): not found

State: not found
```

Future Enhancements

- ✔ Efficiency of the project can be increased by training on the images of handwritten characters dataset (the more the merrier).
- ✔ Integrating the project to a RaspberryPi will be more practical and handy to use.
- ✔ The faster the processor the quicker will one get result so a faster processing speed can be used in practical cases.
- ✔ Currently the Project is only giving District , State and Post Office from CSV file but other things like Post office’s phone number , division , circle , taluk etc can also be printed by modifying the code accordingly.
- ✔ Project currently have a simple GUI based on tkinter but it can be made much more user friendly and easily navigable by using many other modules.
- ✔ After increasing the Efficiency the idea of this project can be implemented in Post Offices to automatically categorize the mails according to their destination State.

Resources

- ✗ **Google :** For searching different queries
- ✗ **Tesseract :** https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/
- ✗ **Opencv :**http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_tutorials.html
- ✗ **Pincodes :** https://data.gov.in/catalog/all-india-pincode-directory
- ✗ **Tkinter :** http://effbot.org/tkinterbook/
- ✗ **Github :** Different Projects for learning more about Character Recognition