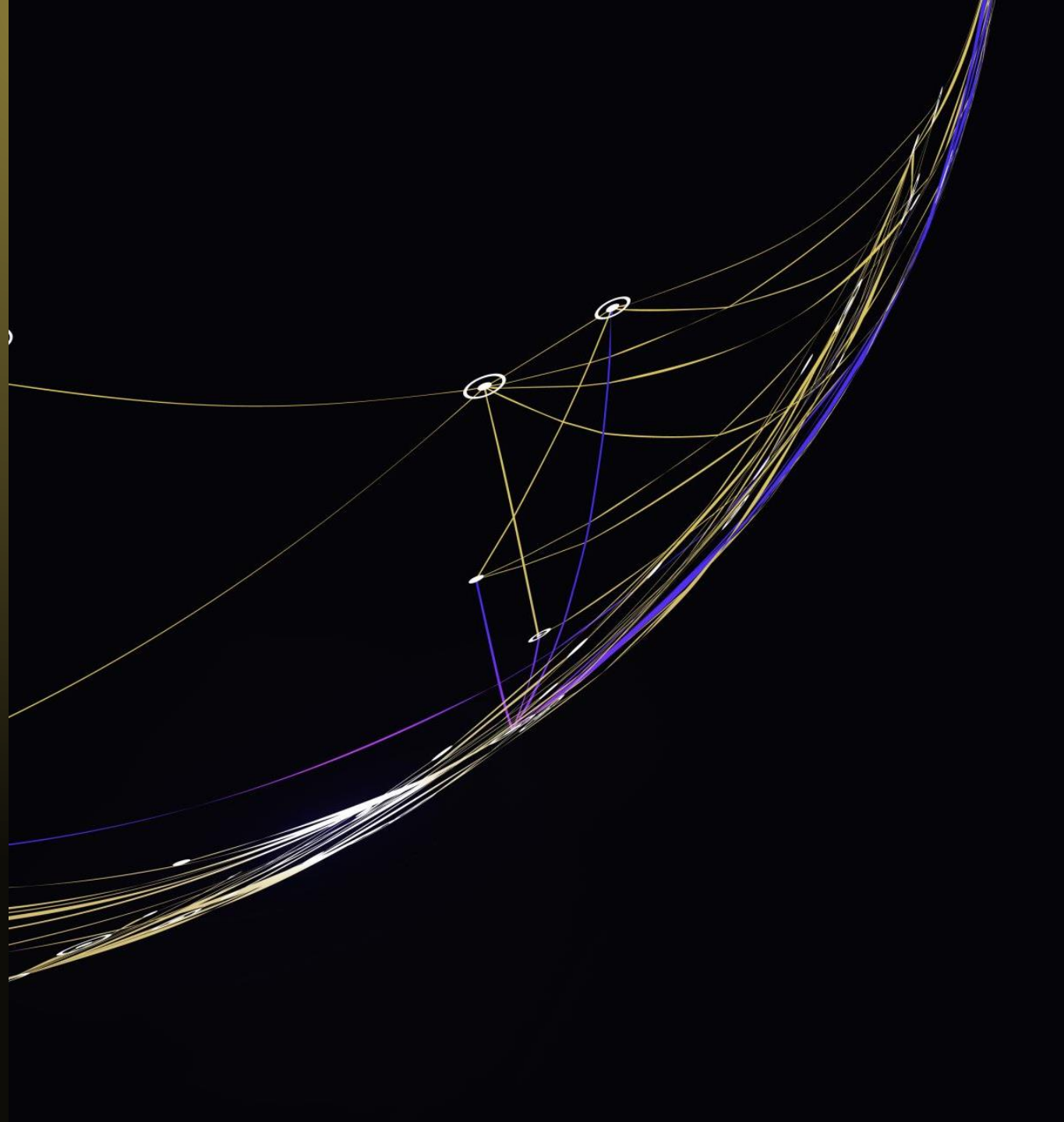
Optical Character Recognition

WITH NLP

LAKSHAY JAIN

1910110214



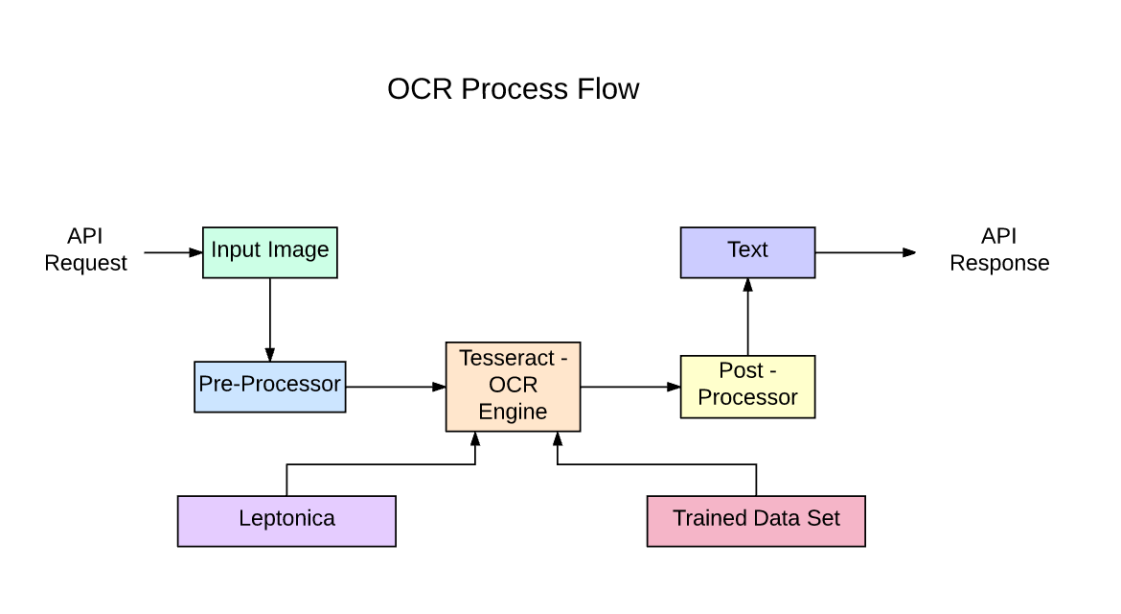
Project Overview

- Created an Optical Character Reader for extracting text from images and scanned handwritten text.
    - ❖ Text from Images Using Tesseract
    - ❖ Text from handwritten Images Using TensorFlow

- Used (NLP) Natural Language Processing to improve OCR accuracy.
    - ❖ Using BERT(Bidirectional Encoder Representations from Transformers)
    - ❖ Using NLTK
    - ❖ Using Python Spellchecker

OCR Using Tesseract

- Tesseract is used directly using an API to extract printed text from images.
- Tesseract includes a new neural network subsystem and uses LSTM.
- Doesn't work well while extracting handwritten text.



OCR Using TensorFlow

- OCR for extracting text from images containing handwritten text.
- Consists of a **Neural Network (NN)** which is trained using images containing handwritten text from the ***IAM dataset***.
- Image is split line-wise for text extraction, as the model is trained for extracting text from a line.

Model Overview

Model consists of :

Convolutional NN (CNN) layers

Recurrent NN (RNN) layers

Connectionist Temporal Classification (CTC).

Post-OCR Error Detection and Correction

I. Process scanned image using OCR

- ✓ Scanned text is cleaned by removing special and unwanted characters using NLTK library functions.

II. Process document and identify unreadable words

- ✓ Incorrect words are identified by Python enchant's SpellChecker function.
- ✓ NLTK's "Parts of Speech" tagging is used to exclude person names from incorrect words.
- ✓ Each incorrect word is replaced with a [MASK] token, and replacement word suggestions from SpellChecker are stored.

III. Load BERT model and predict replacement words

- ✓ BERT model looks for the [MASK] tokens and then predicts the original value of the masked words, based on the context provided by the other words in the sequence.

IV. Refine BERT predictions by using suggestions from Python SpellChecker

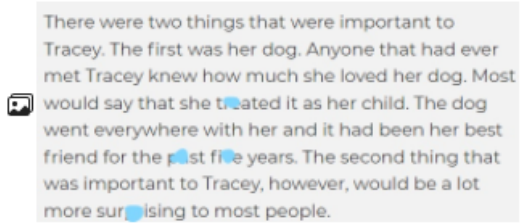
- ✓ The suggested word list from SpellChecker, which incorporates characters from the garbled OCR output, is combined with BERT's context-based suggestions to yield better predictions and the best prediction replaces the [MASK] token.

SAMPLE OUTPUT

Image to Text

Extract all the text from a selected image using tesseract OCR engine.

Input



Choose File p3.jpg

Output

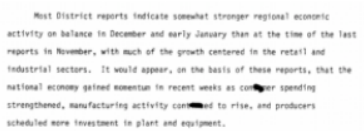
loading tesseract core
initializing tesseract
found in cache eng.traineddata
loading eng.traineddata
initializing api
recognizing text
success

```
There were two things that were important to
Tracey. The first was her dog. Anyone that had ever
met Tracey knew how much she ioved her dog. Most
would say that she ti slated it as her Chiid. The dog
went everywhere with her and it had been her best
friend for the p .st fi e years. The second thing that
was important to Tracey, however, would be a iot
more sur. ising to most people.
```

Image to Text

Extract all the text from a selected image using tesseract OCR engine.

Input


Most District reports indicate somewhat stronger regional economic activity on balance in December and early January than at the time of the last reports in November, with much of the growth centered in the retail and industrial sectors. It would appear, on the basis of these reports, that the national economy gained momentum in recent weeks as conmer spending strengthened, manufacturing activity conthd to rise, and producers scheduled more investment in plant and equipment.

Choose File t.png

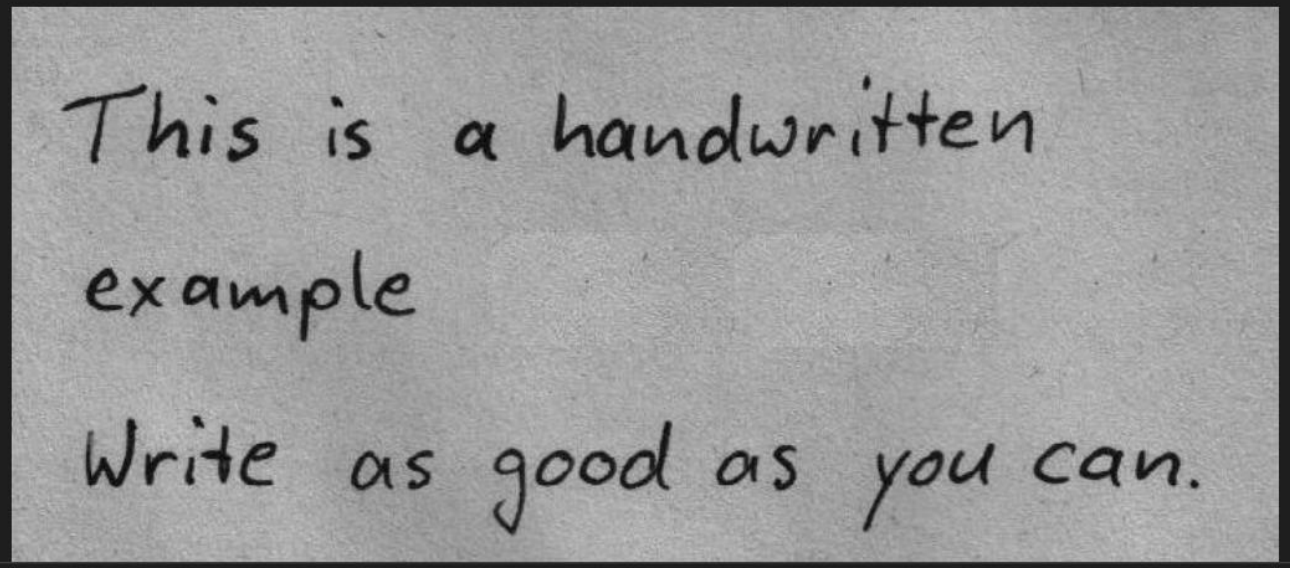
Output

initializing api
recognizing text
success

```
Most District reports indicate somewhat stronger regional
economic
activity on balance in December and early January than at the
time of the last
reports in November, with much of the growth centered in the
retail and
industrial sectors. It would appear, on the basis of these
reports, that the
national economy gained momentum in recent weeks as conmer
spending
strengthened, manufacturing activity conthd to rise, and
producers
scheduled more investment in plant and equipment.
```

data > t12.png

Search (Ctrl+Shift+F)



PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
his is a hand risen
 example
 write as good as you can
```