PROJECT REPORT

Document similarity

Comparison of two images containing English text.

Name: Abhinav kumar gaur

University Roll no: 2013211.

INTRODUCTION

It is easy for humans to understand the contents of an image by just looking at it. You can recognize the text on the image and can understand it without much difficulty. However, computers do not function similarly. They only understand information that is organized. And this is exactly where Optical Character Recognition comes in the picture.so this project we are going to extract text from two Images, and do some comparison on that text.

OPTICAL CHARACTER RECOGNITION

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, Handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example: from a television broadcast).

TESSERACT OCR

Tesseract is an open-source text recognition engine that is available under the Apache 2.0 license and its development has been sponsored by Google since 2006. In the year 2006, Tesseract was considered as one of the most accurate open-source OCR engines. You can use it directly or can use the API to extract the printed text from images. The best part is that it supports an extensive

variety of languages. It is through wrappers that Tesseract can be made compatible with different programming languages and frameworks. In this project we are going to use tesseract libraries. And use them to extract text from images.

Procedure followed during project.

Step 1: Save image 1.

Step 2: save image 2.

Step 3: install tesseract libraries.

Step 4: Test for working of tesseract using Command line interface.

Step 5: open graphical user interface to extract text.

Step 6: Select both the images in browse section.

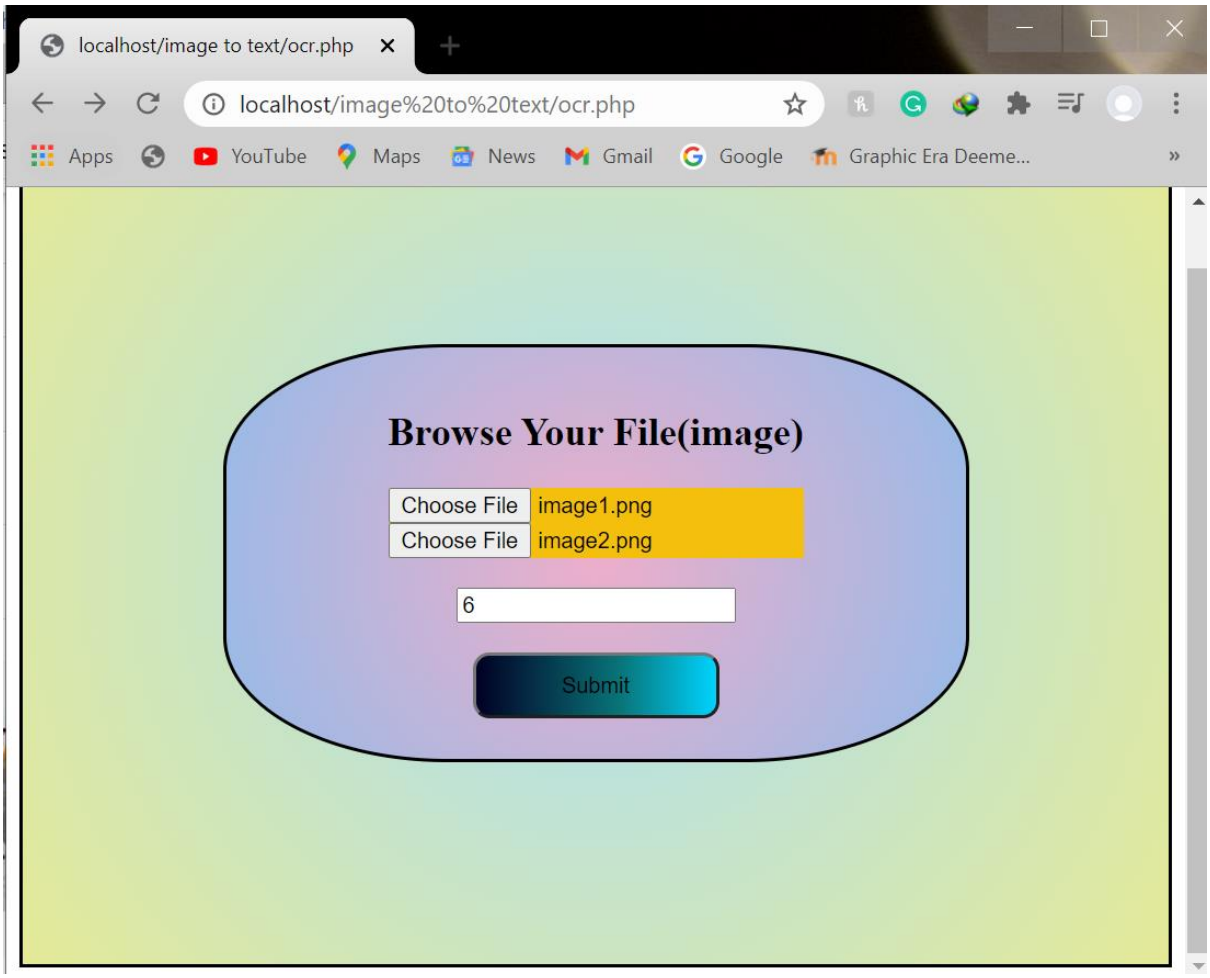
Step 7: provide Factor to check for plagiarism.

Step 8: extract text from images and save text of those images to two different text files.

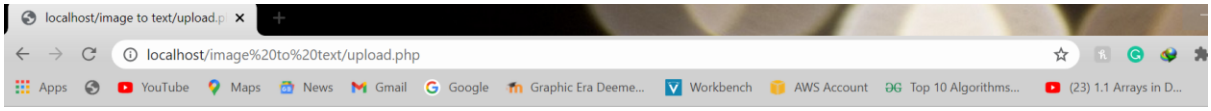
Step 9: Extract text from those files and check for similarities (plagiarism).

Step 10: print output for processed comparison, after applying Algorithm (number of words, no of unique words, percentage match between two images).

Graphical user interface:



Output:



AFTER EXTRACTING IMAGE ONE : !!

---

* Machine Learning Examples

To see end-to-end examples of the interactive machine learning analyses that Colaboratory makes possible, check out these

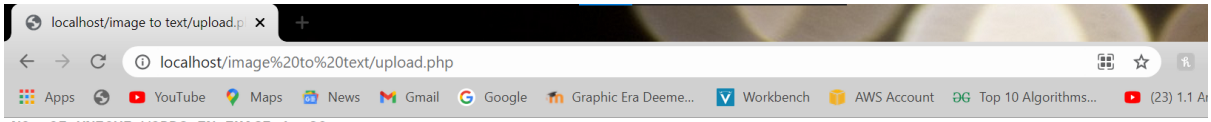
tutorials using models from TensorFlow Hub.

A few featured examples:

e Retraining an Image Classifier: Build a Keras model on top of a pre-trained image classifier to distinguish flowers.
Text Classification: Classify IMDB movie reviews as either positive or negative.

Style Transfer: Use deep learning to transfer style between images.
Multilingual Universal Sentence Encoder Q&A: Use a machine learning model to answer questions from the SQUAD

dataset.
Video Interpolation: Predict what happened in a video between the first and the last frame.
NO. OF CHARACTER IN IMAGE 1: 634
NO. OF UNIQUE WORDS IN IMAGE 1: 80



NO. OF UNIQUE WORDS IN IMAGE 1: 80

AFTER EXTRACTING IMAGE TWO: !!

Jachine Learning Examples

o see end-to-end examples of the interactive machine learning analyses that Colaboratory makes possible, check out these
Jtorials using models from TensorFlow Hub.

. few featured examples:

e Retraining an Image Classifier: Build a Keras model on top of a pre-trained image classifier to distinguish flowers.
¢ Text Classification: Classify IMDB movie reviews as either positive or negative.

NO. OF CHARACTER IN IMAGE 2: 420
NO. OF UNIQUE WORDS IN IMAGE 2: 55

OUTPUT: 27
percentage % match respect to image 1: 36.486486486486%
percentage % match respect to image 2: 55.102040816327%