UNIVERSITY OF CALIFORNIA, RIVERSIDE

Department of Computer Science and Engineering

Final Report (Machine Learning Project)

Project Title: Image Text Extraction and Information Retrieval
Course: CS 205 Artificial Intelligence
Group Members: Cristina Lawson (claws001), Hamsa Veerendra (hveer003), Rucha Kolhatkar (rkolh001)

Capable of taking pictures of text, recognizing the text, converting the same to computer text, and searching either typed or handwritten text is a very applicable concept. With this, the question that arises is how can the python libraries be used to extract text from images and search through results with specific queries?

Accomplishments

- Successfully stacked machine learning libraries.
  - PIL allows us to process the image
  - Pytesseract allows us to convert the text in the image into a text string.
  - Pandas and Elasticsearch were used to manipulate data frames and initialize a search engine for the data.
- Worked with multiple datasets.

What We Learned

We learned more about the applications of image text recognition.
This project allowed us to get a better understanding of python libraries, combining multiple datasets into one, and know-how these libraries typically work in converting the document and dictionaries that store the images and results into Pandas data frames. This activity is an example of what can be done using image processing and information retrieval techniques. Also learned that the typed text had higher accuracy in comparison to the handwritten text.

Challenges

Initially, we were confused while selecting the title for the project as machine learning consists of a wide range of techniques. Eventually, image recognition and information retrieval intrigued us to implement the machine learning libraries. Also, we faced issues with colab, as the machine learning libraries we were trying to implement were not imported. As a reason use of jupyter notebook and pip install is needed for the code to work.

Most Interesting Activities

This project was very interesting since this was our first time delving into a succinct tutorial project. The project was a pre-defined project for which we were just required to implement the underlying machine learning techniques and libraries.

Most “Hated” Activities

The code was not working on colab which made us spend more time figuring out what exactly can be done. It took a lot of rewriting the code. As of which we had to do it in the repository and write the final report doc (which we wouldn't have been written if the code was working on colab).

Team Review

Our team worked together very well during the project. We were able to communicate very well in order to finish each of the parts of the project and were able to turn in the parts of the project by their respective due dates. We also were able to put in equal amounts of work among the team members. Everybody in the group did their part in the project and everyone contributed either together or individually to certain parts of the project overall resulting in equal contribution.

Final Thoughts

The project taught machine learning techniques and the libraries like information retrieval and text recognition which underlies real-world applications that deal with processing documents containing free text. Overall, we thoroughly enjoyed working on this project. Parts of the project presented their challenges, but with some discussion, we were able to work around those challenges.