MSc Project Report

Instance segmentation based objects detection in digital documents

Bravin Balasubramaniam
2019

Student ID - 2017385(IIT) 1715892(RGU)

A report submitted as part of the requirements for the degree of MSc Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

Digital documents have increased in numbers exponentially within the last twenty years. Because of this information captured in digital documents also lost vastly. There are multiple researches done on using Natural Language Processing to mechanically extracting, understanding and, eventually, summarizing key data from digital documents. However, while text is without argument, a basic way to convey data, there are contexts where graphical components are far more powerful. For example, in scientific research papers, several experiments, variables and numbers must be reported in a concise manner that fits better with tables/figures than text. Graphical components possess in conveying information that may be otherwise cumbersome to explain in words, each for the author to express and also the reader to grasp.

We developed an application which can identify/detect any graphical components in given digital document and extract them separately. Application not only have the capability to extract these graphical components but it also can classify them into three different categories/classes.

1. Tables
2. Charts
3. Other (Any other graphical components other than tables and charts)

The results shows that graphical components are extracted from digital documents and classified correctly with an 81% of accuray.