Big Data Analytics for Historical Document Processing

James Patrick Philips
Mentor: M. H. N. Tabrizi
Department of Computer Science

INTRODUCTION

Historical Document Processing is the process of digitizing written material from the past for future use by historians and other scholars. It incorporates algorithms and software tools from various subfields of computer science, including computer vision, document analysis and recognition, natural language processing, and machine learning, to convert images of ancient manuscripts, letters, diaries, and early printed texts automatically into a digital format usable in information retrieval systems. Within the past twenty years, as libraries, museums, and other cultural heritage institutions have scanned an increasing volume of their historical document archives, the need to transcribe the full text from these collections has become acute. Big Data Analytics and infrastructure will be an essential tool in this field. This study compares performance analysis of two OCR systems, discusses HDP workflow, and highlights the role of OCR software in a RESTful API for HDPaaS.

LITERATURE REVIEW

HISTORIC DOCUMENT PROCESSING WORKFLOW

CASE STUDY

- Used images from the sermons of C.H. Spurgeon
- Compared OCR performance of OCRopy and Ocular OCR software
- OCRopy uses BLSTM neural network & Ocular uses unsupervised machine learning with a multiple models.

ACKNOWLEDGEMENTS

James Patrick Philips wishes to thank his mentor, Dr. M.H.N. Tabrizi for his inspiration and guidance on this project. He also thanks Rebecca Stamilio-Ehret, Trey Cherry, and Timothy Boyd of Edgecombe Community College for their encouragement and support.