Automatic Recognition of Historical Newspaper Content

Description of Project

ProQuest has recently acquired 80 million page images of historical newspapers that cover local and global news coverage back through the eighteenth century. We now wish to automatically decompose those page images into their constitutive headlines and article blocks to make them more useful for information retrieval purposes. This team will create, train and deploy a machine learning pipeline capable of breaking those newspaper page images into their constitutive headlines and article blocks.

The Project

The team will build upon progress made in 2016 to develop a tool that will automatically process our archive of historical newspaper articles. The tool will need to differentiate between title and article blocks, recognize the text, record it in a digitally readable form, and execute at a production scale. The tool should have a high quality and intuitive user interface as well as a complete set of documentation.

The 2016 team should develop a method for evaluating the accuracy of the tool output. The final report should include a list of lessons learned and documented ideas for further system development/improvement. The 2017 team will build upon this.
An extremely successful team will deliver one or more of the following:

- Classification accuracy that exceeds contemporary research in newspaper article decomposition.
- Improvement of underlying text OCR through the use of natural language processing techniques.
- A publishable paper documenting the research method and opportunities for future improvements.

Location
Work will take place on campus with regular visits to the ProQuest Ann Arbor Office to hold meetings and give presentations.

Project Sponsor Mentor

John York,
Director of Engineering
ProQuest Dialog

Over 15 years of experience as a software engineer, architect and technology leader. Industry experience in publishing, information, and software services. Worked for CareerSite.com, eePulse, JSTOR, and ProQuest. John has been the sponsor mentor for two previous MDP teams.

Executive Sponsor Mentor

Roger Valade
VP of Engineering, ProQuest

Senior technology leader with extensive experience in enterprise and application architecture, software development and methodology (with an emphasis on agile), strategic planning, project and program management, offshoring in China and India, and change management. Former positions include VP, Technology for a $200M publishing company; VP, Technical Solutions for a J2EE consultancy; and Architect at General Motors. Roger has been the Executive Sponsor Mentor for 3 previous MDP teams.
Project Faculty Mentor

Brent Griffin
Research Fellow Computer Vision

I am a research fellow at the University of Michigan currently conducting research in computer vision. I am passionate about pushing the boundaries of robotics, and I enjoy learning and developing new methods. My academic research interests include control of underactuated robots and applications of wireless power transfer. In addition to this academic work, I have electro-mechanical system design and manufacturing experience in the aerospace and scientific instrument industries.

Key Skills & Project Roles

MDP Sponsored Projects are both a professional and academic learning experience for students. By participating in this program, students are actively preparing for graduate school and a professional career. As part of the experience, MDP expects professional behavior. To best prepare you for future professional opportunities, your experiences on this MDP team will be very broad. In addition to key technical skills that you will bring to the team, you will engage deeply
in the self-directed learning of new and important concepts, demonstrate flexibility, collaboration, and cooperation, and develop strong professional communication skills. This also means that you will need to be able to work outside of your traditional area of study in the true multidisciplinary nature of our projects. You won’t always be able to anticipate how your skills and expertise will be used, so the MDP Sponsored Project will challenge you to grow and develop as a professional.

<table>
<thead>
<tr>
<th>Project Roles</th>
<th>Key Skills and/or Knowledge</th>
<th>Likely Majors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning / Machine Vision (2-3 students)</td>
<td>Machine vision / optical character recognition (OCR) skills, Experience with image processing, Experience with machine learning, Familiarity with text analysis methods for document clustering</td>
<td>EE, Robotics, Data Science, MIDAS, MICDE</td>
</tr>
<tr>
<td>Programming (2-3 students)</td>
<td>Collaborative programming experience. Software development: Solutions in any language, though Python is preferred for the machine learning components of the task · Experience using version control and GIT</td>
<td>CSE/CS-LSA</td>
</tr>
</tbody>
</table>

**Company Overview**

**Company Overview**

ProQuest connects people with vetted, reliable information. Key to serious research, the company has forged a 70-year reputation as a gateway to the world’s knowledge – from dissertations to governmental and cultural archives to news, in all its forms. Its role is essential to libraries and other organizations whose missions depend on the delivery of complete, trustworthy information.

ProQuest’s massive information pool, built through partnerships with content creators, is navigated through technological innovations that enable users to quickly find just the right information. The company is currently rolling out the all-new ProQuest platform, which moves beyond navigation to empower researchers to use, create, and share content—accelerating research productivity.
Legal Requirements

Citizenship

This project is open to all students regardless of citizenship.

Intellectual Property Agreements / Non-Disclosure Agreements (please select)

Students will sign the standard MDP IP/NDA agreement

Internship Information

Interviews Guaranteed to interested team members