Paris, Texas or Paris, France? Entity Linking for Geographic Data Identification

Description of Project

The Sponsor, ProQuest, is a content aggregator and research and learning hub for students, librarians, and instructors. New methods for exploring and analyzing large amounts of text data are changing the way our users access and analyze our content. Accurate place name identification would have an impact across many products within the ProQuest portfolio.

This project will focus on identifying place names in more than 80 million pages of ProQuest’s historic newspaper XML. There are several challenges to this task, which make high rates of precision and recall difficult. These challenges include:

- Ambiguity often exists between entities: Is the newspaper referring to Ann Arbor the place or Ann Arbor the person?
- Place name abbreviations, and pseudonyms, differ based on historical, political, and language-based, naming conventions
- Varying degree of OCR quality and spelling variations in historical content.

A high quality entity-linking engine would enhance a number of ProQuest’s services as well as enable new products. Location Browse: Location-based content indexing and search.

- GIS Data Visualizations: For products such as Indian Tribe Claims or Historic Newspapers, provide a more useful way to explore content.
- Topic-Product Development: We want to move beyond query-based, automatic product development. Rich and automated metadata production is critical for topic products.

The student team will deliver an end to end system/engine that labels geographical entities with increased precision and recall. The team will begin their investigation with Stanford’s CoreNLP NER for GIS data extraction. This open-access software is commonly used, but only somewhat reliable.
Phase I - After reviewing the state of current solution techniques, benchmark most promising techniques for benefits/limitations and develop a strategy for team’s tool development. Develop a comprehensive set of content-specific target case examples, create a Master Evaluation Data Set incorporating multiple examples of target cases, and verify that that Master data set is 100% correct. Create an automatic test harness/environment to efficiently evaluate team’s new developments on the Master Evaluation Data Set. Demonstrate the correctness of test environment.

Phase II - Implement most promising solution techniques identified in the team’s strategic plan.

Phase I (Base Level Goal end of first term)

- Implement end-to-end system that relies on Stanford CoreNLP. This will include text processing, validation, and some parameter optimization.

- Develop a comprehensive set of content-specific target case examples, create a Master Evaluation Data Set incorporating multiple examples of target cases, and verify that that Master data set is 100% correct.

Phase II (Target Goal end of project)

- Improve the test system with additional post-processing or pre-processing decisions. These might include: A) Entity linking and disambiguation, B) Building a historic knowledge base, C) Improving pre-processing text processing (e.g. Spelling normalization / Part-of-speech tagging).

- Validate precision and recall with out of sample datasets and prioritize future avenues of development.

Long Term Stretch Goals (one or more of the following)

- Beat Watson! Develop a pipeline which outperforms Watson’s location entity tagger for the same test set.

Location

Work will take place on campus with regular visits to the ProQuest Ann Arbor Office to hold meetings and give presentations.
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>General Programming (2 positions)</td>
<td>Solid programming experience-- EECS 281 (or equivalent)</td>
<td>Computer Science, Data Science</td>
</tr>
<tr>
<td>Front End Developer (1 position)</td>
<td>Front End Development. Web-App development in Java</td>
<td>Computer Science, SI HCI</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>Experience / Strong interest in Natural Language Programming</td>
<td>Computer Science, Electrical Engineering, Mathematics</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Experience / Strong interest in Machine Learning</td>
<td>Computer Science, Electrical Engineering, Mathematics</td>
</tr>
<tr>
<td>Data Visualization</td>
<td>Experience / Strong interest in data visualization especially D3 (Data Driven Documents) and mapping visualizations</td>
<td>Data Science, Computer Science, Mathematics</td>
</tr>
<tr>
<td>Academic Credit Substitutions completed by MDP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additional Desired Skills/ Knowledge/ Experience

- This project will be run in Python; students should have experience in Python or be prepared to quickly develop their skills.
  - Project experience with the following python libraries: scikit-learn, NLTK, Keras, Theano would be particularly beneficial.
  - Please include a description of your programming language experience in your personal statement.
- Familiarity with entity recognition systems and disambiguation processes and libraries.
- Experience with data visualization, especially d3.js and mapping visualizations.
**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 and Right to Work Options (please select)

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**

Internships may be available based on internal hiring needs and business conditions, to be determined in Winter 2018.