

Student Projects for Fall 2009 / Spring 2010

Steven Skiena

Dept. of Computer Science
SUNY Stony Brook
skiena@cs.sunysb.edu

This is a list of (primarily) master's student projects which I interested in having students work on. Some are available as undergraduate research projects for strong students. Please feel free to talk to me if you are interested.

I *very strongly* recommend that M.S. students wait until the beginning of their second semester at Stony Brook before selecting their masters project. It is much better to spend a semester taking classes and getting to know what area you are really interested in than to grab the first project that will take you on. Rest assured that projects *will* be available when you are ready to start working seriously on them. Typically my M.S. students arrive at Stony Brook in August, select their project in January, work on them some in the spring, a lot in the summer, and some the next fall.

I encourage M.S. students potentially interested in working with me to take my Computational Biology course (CSE 549), which is typically offered in the fall. Most of the projects below involve techniques for string/text processing and algorithmic data analysis which are discussed in the course, even if the projects themselves are not biologically-oriented.

1 Text Mining and Document Analysis

The increasing volume of informative content on the World-Wide Web coupled with decreasing costs of computation and communication have created exciting new opportunities in text mining. Towards this end, we have started the *Lydia* project, a natural language system for rapidly assimilating the primary vocabulary associated with high-quality curated text, and extracting relations between them.

Lydia-style text analysis has natural applications in the financial, legal, medical, and homeland security sectors, and we look forward to interacting with associated companies. The ultimate goal of *Lydia* is to build a *relational encyclopedia* of much of the world's knowledge through the analysis of news sources, reference texts, and primary sources such as government documents. *Lydia* (<http://www.textmap.org>) is still at a relatively early stage of development, but it is already producing interesting analysis of significant volumes of text.

Projects in text mining include:

- *Text Extraction from Newspapers and Web Sources* – The *Lydia* system currently spiders text from roughly 500 online news sources on a daily basis. We would like to increase this to several thousand news sources, and perhaps millions of web blogs. Our current bottleneck is the amount of human effort needed to develop customized spidering programs to retrieve, analyze, and format the resulting text for downstream analysis.

This particular project involves building a set of general text extraction scripts and quickly customizing them to particular text sources. We are also interested in performing meta-analysis of the contents of different online news sources – how informative are they are on specific topics (news, sports, business, etc.), how politically biased are they, etc.

- *High-Throughput Parallel Processing* – We have acquired a 28-node cluster computer to dedicate to this project, which should increase our ability to analyze text by one or two orders of magnitude. Properly exploiting this machine will require parallelizing our text processing pipeline, and identifying/resolving subsequent performance bottlenecks associated with the database and analysis components of our system.
- *Six-Degrees of Separation Analysis* – *Lydia*'s juxtaposition analysis and relation extraction implicitly defines a graph of the set of actors. Social network analysis concerns itself with the task of constructing models of how people are connected to each other, and the possible implications of these models. Analysis of such relation graphs have implications in business, law enforcement, and many other human endeavors.

The project revolves around building a 6-degrees of separation server for the relationship data derived from *Lydia* analysis . Check out <http://www.cs.virginia.edu/oracle/> or <http://www.baseball-reference.com/oracle/> for inspiration. The challenge in this project concerns the large number and variety of relevant actors, and the dynamic nature of the their relationships. In addition, we may consider other measures of distance, such as the number of short paths between two actors instead of the length of the absolute shortest path.

- *Separating Good News from Bad News* – Can we write a program which reads a news article and identifies whether it is good or bad news? More specifically, given a news article and a selected keyword or topic, measure how positive or negative the story is for the topic. Techniques for doing this would be to build a lexicon of positive words (grow, increased, succeed) and negative words (decreased, crashed, failed) and analyze how close the words appear next to the keyword, and whether they are modifiers.
- *Foreign Language Analysis* – We have started to process non-English documents through automatic translation, with encouraging preliminary results. That said, we look forward to improving our results through more rigorous analysis, particularly with respect for sentiment analysis.

- *Literary Analysis* – The same techniques *Lydia* applies to identify characters in the news can be used to identify characters in a novel. This project involves doing *Lydia* analysis of the English-language novels and plays made available by Project Gutenberg. Over 10,000 such books are currently available for analysis, with dozens of new ones appearing each month.

Lydia analysis can make the following contributions:

1. *Character lists and concordance* – what characters and places are in the novel and where do they appear? Which are historical people versus fictional characters?
2. *Character interaction graphs* – which characters interact with which other characters, and in what ways? Character interaction graphs can be built up by juxtaposition analysis or relation extraction. Knuth pioneered character interaction graph construction for GraphBase – how do our extracted graphs compare to his (perhaps constructed by hand?) We are interested both in embedded graphs for visualization and combinatorial representations for analysis.
3. *Marked up editions* – can we annotate the html versions of these books with hyperlinks around every actor to help people navigate through it?
4. *Recommendations* – can we cluster books by similar keywords and themes, or similar actor structures/densities to make recommendations to other Gutenberg books?

I envision the result of this project will be a webpage for each Gutenberg book with our contributed analysis and links to relevant resources.

Useful background for these projects include Internet programming experience, natural language processing, and/or Hadoop Map/Reduce programming. There is room for several students on this project.

2 International Dining: Restaurant Menu Translation Service

Travelers to a foreign land often have difficulty making sense of restaurant menus due to both language and cultural differences. This project involves building the workflow / infrastructure for a service to ease the complexities of international dining.

The input to this system will be a restaurant menu, submitted either scanned or in some meaningful text representation (e.g. .doc or .txt) or perhaps scraped from a webpage. The output will be a meaningful menu in all popular languages, both as a webpage and in a printable document.

The primary technical challenges are:

- Set up a database of dishes with associated translations.

- Segment and parse arbitrary restaurant menus.
- Match up menu entries to database entries.
- Farm out translation (and possibly segmentation) tasks to Amazon Turk.
- Identify photos of dishes in Flickr or other photo services.
- Assemble a printable translated menu.
- Generate a webpage for each restaurant.
- Create an international dining website with proper search capabilities.

3 Who Flies There?

A great aggravation in making flight arrangements is identifying which are the airlines which fly into a particular city. Which airlines can you take to get to Champaign, Illinois? This project involves building a service which collected and displays such information automatically. The primary technical challenges are:

- Set up a database of airports with connecting airlines.
- Identify a comprehensive source of this connection information, and figure out how to scrape/update this data as it changes.
- Generate an appropriate webpage for each city and airline.
- Create an international flying website with proper search capabilities.