Project Proposal

Author: Michael Dickens mdickens@stanford.edu

Author: Mihail Eric meric@stanford.edu

TODO: 1) Take a limited collection of classes from explorecourses.stanford.edu (scrape using BeautifulSoup or JS?) 2) Manually determine the relevance of the results of a search query. 3) Base-line system: Keyword search (list of professor's names, list of course names) 4) How to input queries: comma-delimited phrases?

Task Definition

This project provides intelligent Stanford course search. It allows Stanford students to quickly find courses that meet a wide range of criteria not offered by Stanford ExploreCourses.

Types of Queries to Support

- 'courses taught [by {NAME PROFESSOR}] [in {QUARTER OF YEAR}] [at {TIME OF DAY}]'
- 'courses taught by {PROFESSOR QUALIFIER}'
- 'courses that meet on {DAY OF WEEK}'
- 'courses in {NAME DEPARTMENT} department'
- 'courses on {TOPICS/SUBJECT MATTER OF COURSE}'

Literature Review

"Information Retrieval By Natural Language Querying." Douglas E. Appelt et al. All web page information characterized and subsequently stored in a database on which search queries could be run. We will be employing a similar model for parsing web pages off Explore Courses. The authors employed a syntactic and semantic parsing scheme for dissecting search queries. In particular, text was parsed into tokens including multi-words (such as 'because of'), noun and verb phrases, and other such word groups.

"Natural Language Processing for Online Applications: Text retrieval.." by Peter Jackson and Isabelle Moulinier Book on natural processing for text retrieval to be read for various other methods for query parsing.

"Wikipedia-based Semantic Interpretation for Natural Language Processing" Gabrilovitch et al. This paper sought to use the vast corpus of information available on Wikipedia to develop a high-dimensional space of concepts to be employed in more accurate semantic parsing. Their study investigated the degree of semantic relatedness among sentence fragments. This is something that is directly relevant to our project for if we want to support more complicated queries, it will be necessary to have a framework on which to base the relatedness between courses and the topics that they cover.

Baseline System

For our preliminary baseline, we implemented a system for parsing single web pages on Explore Courses and extracting course information for the courses available on those pages. Information was outputted as a dictionary with course names as keys and a dictionary of course attributes as a value, in a format very similar to JSON. So far we have only catalogued a limited collection of courses. In the future, we intend to catalogue all courses available on Explore Courses and ideally save those to an external database file. Our first attempt at performing course searches based on queries simply involved doing a key word search through the attributes of a course (i.e. 'instructor, jerry cain').

Data and Preliminary Results

Though we have not decided on a good measure for quantitatively determining the quality of a search result, the key word search is limited in that it cannot support more complicated queries. We can perhaps use machine learning algorithms to rate the relevance of a query. Perhaps we can calculate the dot product between a weight vector and the feature vector of a query result. The weight vector can be updated via the Perceptron algorithm to improve quality of results.