January 29, 2019

Yarowsky@jhu.edu
Johns Hopkins University
Department of Computer Science
Professor

Prof. David Yarowsky

(601) 466/666

INFORMATION RETRIEVAL AND WEB AGENTS
Office Hours: Instructor - Wed 3-4, Thursday after class and by appointment.

Web Page: \texttt{http://www.cs.jhu.edu/~yarowsky/cs466.html}

Classroom: Hackerman B-17

Meeting Time: Tu, Th: 3:00-4:15 PM

Winston Wu

Malone Ugrad CS Lab

410-316-5372

Hackerman 324G

Prof. David Yarowsky

601.466/666 - Information Retrieval and Web Agents
Selected papers distributed in class.

- Primary Readings
- L. Wall, T. Christiansen and R. Schwartz, Programming Perl, O'Reilly

Selected papers distributed in class.

- Another Perl Reference of your choice
Recommended Readings

- A python reference guide of your choice.
J. Williams, Bots and Other Internet Beasties, Indianapolis, IN: Sams Net, 1996.
Prerequisites

Students should have a solid programming background and have taken 601.226 (Data Structures) or its equivalent. Knowledge of Perl (or willingness to learn the language on your own relatively quickly) is also very important. No background in NLP is assumed, and although 601.465 is helpful, it is not necessary as a pre-requisite.

The material covered will be complementary to that in 601.465 (Natural Language Processing) and 520.666 (Information Extraction). Similarities and differences will be discussed in the first class.
Final grades will be based on the following (subject to change):

- Class Participation: 6%
- Final Project: 31%
- Comprehensive Exam: 32%
- Assignments (4): 31%

Final grades will be based on the following (subject to change):
Assignments

1. Machine Learning for preliminary text analysis and corpus processing
2. A vector-based information retrieval system
3. Vector-based and Bayesian text classification and information extraction
   (a) Named entity classification and word sense disambiguation
   (b) Email/News routing and filtering - supervised IR
   (c) Relational text classification problems (gender detection, authorship)
4. Build and unleash a web agent

Considerable infrastructure will be provided in support of each assignment. Partial code, supporting routines and training data will also be provided so students may receive feedback during assignment development and debugging. The first 3 assignments will be empirically evaluated on held-out (previously unseen) test sets. A portion of the grade will be based on this objective. These will include partial code, supporting routines and training data.

The first 3 assignments will be empirically evaluated on held-out (previously unseen) test sets. A portion of the grade will be based on this objective. These will include partial code, supporting routines and training data.

Assignments
Course Requirements

Final Project:
The final project for the course will be on a topic of your own choosing.

Late Homework Policy:
One homework assignment may be handed in up to 5 days late without penalty, and without the need for permission or excuse. No other late homeworks will be accepted.

Several options will be suggested.
Sample Final Project Areas (Previous Years)

- Image retrieval (caption-based and graphically-based)
- Speech-based retrieval: broadcast news and RealAudio
- Music lyrics IR
  - White-water rafting: river report extraction and synthesis
  - Specialized web bots
  - English
  - Foreign language classification (Dutch vs. German vs. Danish vs. En-
  - Resume Finder bot (locates, classifies, extracts info)
- FriendFinder - builds profiles of likes/dislikes and classifications homophilies
- Email Filter and classifer
- Comparative Shopping Bot (online book sellers, music stores, etc.)
Sample Final Project Areas (Previous Years)

- Image retrieval (caption-based and graphically-based)
- Speech-based retrieval: broadcast news and RealAudio
- Music lyrics IR
- White-water rafting: river report extraction and synthesis
- Specialized web bots
- Foreign language classification (Dutch vs. German vs. Danish vs. English)
- Resume finder robot (locates, classifies, extracts into)
- FriendFinder - builds profiles of likes/dislikes and classifies homepages
- Email filter and classifier
- Comparative Shopping Bot (online bookstores, music stores, etc.)
# I. Core Information Retrieval

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/29</td>
<td>Course Overview. Discussion of problems and issues in Information Retrieval</td>
</tr>
<tr>
<td>1/31</td>
<td>Introduction to IR models and methods (Boolean/Vector/Probabilistic)</td>
</tr>
<tr>
<td>2/05</td>
<td>Preliminary stages of text analysis and document processing. Boolean IR models.</td>
</tr>
<tr>
<td>2/07</td>
<td>Inverted files, indexing, signature files, PFA, n-grams, suffix arrays.</td>
</tr>
<tr>
<td>2/12</td>
<td>Inverted files, indexing, signature files, PFA, n-grams, suffix arrays (cont.)</td>
</tr>
<tr>
<td>2/14</td>
<td>Vector-based IR models (cont.) - including term weighting, similarity measures.</td>
</tr>
<tr>
<td>2/19</td>
<td>Vector-based IR models (cont.) - including term weighting, similarity measures, SVD/LSI.</td>
</tr>
<tr>
<td>2/21</td>
<td>Query expansion, thesaurus creation, clustering algorithms, SVD/LSI.</td>
</tr>
<tr>
<td>2/24</td>
<td>Evaluation metrics, test collections and issues.</td>
</tr>
<tr>
<td>2/26</td>
<td>Inverted files, indexing, signature files, PFA, n-grams, suffix arrays (cont.)</td>
</tr>
<tr>
<td>3/05</td>
<td>Industry Standard IR tools (e.g. Lucene, Solr, ElasticSearch).</td>
</tr>
<tr>
<td>3/07</td>
<td>Relevance Feedback and Probabilistic IR models (cont.) - including user modeling, automatic feedback acquisition.</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3/14</td>
<td>Information Extraction, Text Classification, Authorship ID, Language ID, Gender detection</td>
</tr>
<tr>
<td>3/19</td>
<td>NO CLASS - Spring Break</td>
</tr>
<tr>
<td>3/21</td>
<td>NO CLASS - Spring Break</td>
</tr>
<tr>
<td>3/26</td>
<td>Music Information Retrieval and Image Search</td>
</tr>
<tr>
<td>3/28</td>
<td>IE (cont.) - Sentiment classification, authorship ID, language ID, gender detection</td>
</tr>
<tr>
<td>4/02</td>
<td>IE (cont.) - Sense tagging and general semantic disambiguation</td>
</tr>
<tr>
<td>4/04</td>
<td>IE (cont.) - Named entity recognition/tagging, semantic frame analysis</td>
</tr>
<tr>
<td>4/09</td>
<td>Information visualization - Dotplot, Texttiling, Graphical queries</td>
</tr>
<tr>
<td>4/04</td>
<td>Information Visualization</td>
</tr>
</tbody>
</table>

II. Information Extraction and Visualization
III. Web Agents and IR on the Web

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th 4/11</td>
<td>Web robots, spiders, crawlers, ants, HTTP, Robot exclusion</td>
</tr>
<tr>
<td>Th 4/16</td>
<td>IR on the World Wide Web - new technologies and protocols</td>
</tr>
<tr>
<td>Th 4/18</td>
<td>IR on the WWW: IR and IR, Harvest, collection fusion, Metacrawler</td>
</tr>
<tr>
<td>Th 4/23</td>
<td>Collaborative filtering, Web Agents</td>
</tr>
<tr>
<td>Th 4/25</td>
<td>Web agents - webshopper, bargainhunter, case studies</td>
</tr>
<tr>
<td>Th 4/30</td>
<td>Web agents - case studies, economic, ethical, legal and political issues</td>
</tr>
<tr>
<td>Th 5/02</td>
<td>Future directions, overview and conclusion</td>
</tr>
</tbody>
</table>

Th 4/25

Web Agents and IR on the Web