INFORMATION RETRIEVAL AND WEB AGENTS

(601.466/666)

Prof. David Yarowsky

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

January 23, 2024
601.466/666 - Information Retrieval and Web Agents

Instructor:  Prof. David Yarowsky
            Hackerman 324G
            410-516-5372
            yarowsky@jhu.edu

TA:  Niyati Bafna
     nbafna1@jhu.edu

CA:  Shreayan Chaudhary
     schaud31@jhu.edu

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

Classroom:  Hackerman B-17


Office Hours:  Instructor - Thu 2-3, Tuesday/Thursday after class and by appointment
              TAs - TBA, special review sections, and by appointment.
Primary Readings


• A Python language reference book of your choice

• Selected papers distributed in class.
Recommended Readings


- 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 Python (or a willingness to learn the language on your own relatively quickly) is also very important.

The material covered will be complementary to that in 601.465/665 (Natural Language Processing) and 520.666 (Information Extraction). Similarities and differences will be discussed in the first class. No background in NLP is assumed, and although 601.465/665 is helpful, it is not necessary as a prerequisite.
Course Requirements

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

- Assignments (4): 32%
- Comprehensive Exam: 30%
- Final Project: 32%
- Class Participation: 6%
Assignments

1. Machine Learning for preliminary Text Analysis and Corpus Processing
2. A Vector-model 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) Related Text Classification Problems (Gender detection, Authorship ID, Language ID, Sentiment Classification)
4. Build (and unleash) a Web Agent

Considerable infrastructure will be provided in support of each assignment. 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 measure of performance. Code for self-evaluation on a secondary test set will also be provided so students may receive feedback during assignment development and debugging.
Course Requirements

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

Lateness Policy:
Although students have 2-3 weeks to complete most assignments, recognizing that last-minute illness or unplanned events may occur, homework assignments may optionally be handed in late up to a total of 5 days combined across all homeworks without penalty and without the need for permission or excuse.
Each 24 hour period after the due date and time counts as 1 late day, and are counted in granularities of whole days (no partial days).
Sample Final Project Areas (previous years)

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

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tu 1/23</td>
<td>Course Overview. Discussion of problems and issues in Information Retrieval</td>
</tr>
<tr>
<td>Th 1/25</td>
<td>Introduction to IR models and methods (Boolean/vector/probabilistic)</td>
</tr>
<tr>
<td>Tu 1/30</td>
<td>Preliminary stages of text analysis and document processing. Boolean IR models.</td>
</tr>
<tr>
<td>Th 2/01</td>
<td>Inverted files, indexing, signature files, PAT trees, suffix arrays</td>
</tr>
<tr>
<td>Tu 2/06</td>
<td>Inverted files, indexing, signature files, PAT trees, suffix arrays (cont.)</td>
</tr>
<tr>
<td>Th 2/08</td>
<td>Vector-based IR models</td>
</tr>
<tr>
<td>Tu 2/13</td>
<td>Vector-based IR models (cont.) - including term weighting, similarity measures</td>
</tr>
<tr>
<td>Th 2/15</td>
<td>Query expansion, thesaurus creation, clustering algorithms, SVD/LSI</td>
</tr>
<tr>
<td>Tu 2/20</td>
<td>Evaluation metrics, test collections and issues.</td>
</tr>
<tr>
<td>Th 2/22</td>
<td>Relevance Feedback and Probabilistic IR models</td>
</tr>
<tr>
<td>Tu 2/27</td>
<td>(cont.) - including user modeling, automatic feedback acquisition</td>
</tr>
<tr>
<td>Th 2/29</td>
<td>Document routing/filtering/topic-classification; Spam detection</td>
</tr>
<tr>
<td>Tu 3/05</td>
<td>(Large) Language-model-based IR and Industry Standard IR tools</td>
</tr>
</tbody>
</table>
## II. Information Extraction and Visualization

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Th 3/07</td>
<td>Information extraction, Text Classification and Question Answering</td>
</tr>
<tr>
<td>Tu 3/12</td>
<td>IE (cont.) - named entity recognition/tagging, semantic frame analysis</td>
</tr>
<tr>
<td>Th 3/14</td>
<td>IE (cont.) - sense tagging and general semantic disambiguation</td>
</tr>
<tr>
<td>Tu 3/19</td>
<td>Spring Break (no class)</td>
</tr>
<tr>
<td>Th 3/21</td>
<td>Spring Break (no class)</td>
</tr>
<tr>
<td>Tu 3/26</td>
<td>IE (cont.) - Sentiment classification, authorship ID, language ID, gender detection</td>
</tr>
<tr>
<td>Th 3/28</td>
<td>Information visualization - Dotplot, Texttiling, graphical queries</td>
</tr>
</tbody>
</table>

### III. Web Agents and IR on the Web

<table>
<thead>
<tr>
<th>Date</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tu 4/02</td>
<td>Web robots, spiders, crawlers, ants, HTTP, robot exclusion</td>
</tr>
<tr>
<td>Th 4/04</td>
<td>IR on the WWW cont. - Harvest, collection fusion, Metacrawler</td>
</tr>
<tr>
<td>Tu 4/09</td>
<td>IR on the World Wide Web - new technologies and protocols</td>
</tr>
<tr>
<td>Th 4/11</td>
<td>Music Information retrieval and Image Search</td>
</tr>
<tr>
<td>Tu 4/16</td>
<td>Collaborative filtering. Web Agents.</td>
</tr>
<tr>
<td>Th 4/18</td>
<td>Web agents - webshopper, bargainfinder, case studies</td>
</tr>
<tr>
<td>Tu 4/23</td>
<td>Web agents - case studies, economic, ethical, legal and political issues</td>
</tr>
<tr>
<td>Th 4/25</td>
<td>Future directions, overview and conclusion</td>
</tr>
<tr>
<td>We 5/08</td>
<td>Final Examination (9AM-12PM, official T/Th 3pm exam slot)</td>
</tr>
</tbody>
</table>