Midterm Review

Linguistics 445/515

For the Midterm on Monday, October 20, 2008

1 Topics to be covered

1. Text & Speech encoding
2. Searching
3. Corpus annotation
4. Text Classification
5. Spam Filtering

NB: You will not have any Python questions on the exam.

2 Format of the exam

You will have the entire 1:15 (9:30-10:45) should you need/want it.

1. Matching: 10-20 terms (see list below)
2. Calculations: 5-10 questions
   • Binary numbers, ASCII encoding
   • Boolean expressions
   • Regular expressions
   • Weblinking & webpage ranking
   • Precision/Recall
   • Semantic ontologies
   • Part-of-speech & chunk annotating
   • Stylometric analysis
   • Frequency distributions
   • Rule-based operations
   • Probability calculations (ASR & spam filtering)
3. Short answer: answer 3–5 out of 5–8
   • Types of writing systems, pros & cons
   • Relation of writing systems to languages
• Types of character encoding systems, e.g., ASCII & Unicode
• Challenges of ASR & TTS
• How measurements do & do not correspond to what we hear
• Searching in databases vs. on the web vs. in a corpus
• Improving searching, e.g., semantic web
• Levels of linguistic annotation
• Kinds of information needed for document classification
• N-gram analysis
• Plagiarism detection
• The social context of spam & approaches to fighting spam
• Rule-based spam filters
• Statistical spam filters
• Tokenization & Devious spam

3 Terms to know

3.1 Text/Speech encoding

- text
- speech
- abjad
- alphabet
- syllabary
- syllabic alphabet
- diacritic
- logographic system
- logograph
- pictograph
- ideograph
- semantic-phonetic compound
- bit
- byte
- Big-Endian
- Little-Endian
- ASCII
- Unicode
- Character encoding
- MIME
- meta-information
- continuous
- discrete
- Hertz
- transcribe
- phonetic alphabet
- coarticulation
- articulatory phonetics
- speech flow
- loudness/amplitude
- intonation
- pitch
- fundamental frequency
- intonation
- spectrogram
- sampling rate
- ASR
- TTS
- continuous speech system
- isolated-word system
- acoustic signal processing
- information loss
- irreversible

3.2 Searching

- database
- database frontend
- keyword
- query
- synonym
- boolean expression
- regular expression
- operators
- operator precedence
- escaped character
- counter
- literal strings
<table>
<thead>
<tr>
<th>3.3</th>
<th>Corpus annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- disjunction</td>
<td>- click-through measurement</td>
</tr>
<tr>
<td>- negation</td>
<td>- database</td>
</tr>
<tr>
<td>- counters</td>
<td>- index</td>
</tr>
<tr>
<td>- wildcard</td>
<td>- search engine</td>
</tr>
<tr>
<td>- linking</td>
<td>- relevancy</td>
</tr>
<tr>
<td>- link counting</td>
<td>- precision</td>
</tr>
<tr>
<td>- formal language</td>
<td>- recall</td>
</tr>
<tr>
<td>- regular language</td>
<td>- accuracy</td>
</tr>
<tr>
<td>- meta data/meta tag</td>
<td>- index</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3.4</th>
<th>Text classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>- ontology</td>
<td>- word type</td>
</tr>
<tr>
<td>- corpus/corpora</td>
<td>- word token</td>
</tr>
<tr>
<td>- corpus annotation</td>
<td>- tokenization</td>
</tr>
<tr>
<td>- XML</td>
<td>- lemmatization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3.5</th>
<th>Spam filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>- language identification</td>
<td>- whitelist</td>
</tr>
<tr>
<td>- document classification</td>
<td>- rule-based filtering</td>
</tr>
<tr>
<td>- spam</td>
<td>- weight</td>
</tr>
<tr>
<td>- spam filter</td>
<td>- spam probability</td>
</tr>
<tr>
<td>- blacklist</td>
<td>- statistical filtering</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3.5</th>
<th>Spam filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>- blacklist</td>
<td>- (machine) learning</td>
</tr>
<tr>
<td>- language identification</td>
<td>- false positives</td>
</tr>
<tr>
<td>- document classification</td>
<td>- collaborative filtering</td>
</tr>
<tr>
<td>- spam</td>
<td>- message inoculation</td>
</tr>
<tr>
<td>- spam filter</td>
<td>- structured information</td>
</tr>
</tbody>
</table>