Deliverable #3: Passage Retrieval

Thomas Phan
Jason Shaw
Varden Wang
Lucene

- Followed standard Lucene practices and examples
- Index of TREC 2005 documents was 3.3 GB
  - Stored entire document body in index
MultiText

- Implemented from scratch
- Based on the description in Clarke et al. (TREC 9 submission)
- True IDF measure used
- Find "covers" -- passages bounded by query terms
  - Length of "cover" not accounted for
  - Stopping performed by Lucene
  - Return 500- or 1000-character passage centered on cover
Query Expansion

- Based on Xu & Croft (1996)
- Queries unfocused, ambiguous, or lack detail
- Local Context Analysis
  - Co-occurrence of query terms with passage terms
  - Unable to tag corpus to discover NPs
  - Added terms with score > 5
    - In practice, about 4.8 terms per query
Results

- Document Retrieval

<table>
<thead>
<tr>
<th>QE Level</th>
<th>None</th>
<th>S &gt; 5</th>
<th>+20</th>
<th>+70</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2981</td>
<td>0.2115</td>
<td>0.1735</td>
<td>0.1327</td>
</tr>
</tbody>
</table>

- Passage Retrieval (MRR)

<table>
<thead>
<tr>
<th></th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before QE</td>
<td>0.285/0.145</td>
<td>0.380/0.183</td>
</tr>
<tr>
<td>After QE</td>
<td>0.255/0.114</td>
<td>0.336/0.141</td>
</tr>
</tbody>
</table>