Test Collection for Legal IR From Online Discussion Forums

Ambedkar Kanapala
Sukomal Pal

Department of Computer Science & Engineering
Indian School of Mines, Dhanbad
Contents

• History of Legal Text Collections
• Data Acquisition
• Corpus
• Statistics
• Conclusion
History of Legal Text Collections

• 2006-2012: TREC
  – Enron email messages

• 2013: FIRE
  – Documents: verdicts from Supreme Court, various acts of Parliament of India
Contents

- Brief History of Legal Text Collections
- Data Acquisition
- Corpus
- Statistics
- Conclusion
Data Acquisition

• Collected the data from online discussion forums

• Acquisition of the corpus took three major steps
  – Obtain the raw data using WebCrawler
  – Extracting queries and answers from raw data
  – Dividing queries and answers into TREC-format
• Downloaded webpages using Downloadthemall which is an addon for Firefox browser
Cont..

- Used Jsoup, a java library to extract queries and answers from HTML page

<Question>
what is about children after divorce who can keep children mother or father.
</Question>

<Ans>
custody of child below 5 is generally given to mother.
</Ans>
Cont..

- Used java program to divide the queries and answers into TREC-format
Contents

• Brief History of Legal Text Collections
• Data Acquisition
• Corpus
• Statistics
• Conclusion
Corpus

- Collection of 15 different categories of legal subdomains
- Consists of legal queries raised in a legal forum and their threads answered by legal experts
- Each question is having a unique queryid
- Each answer is stored in a separate file with proper name
- Topics(queries) - 20784
- Documents(answers) - 42708
what is about children after divorce who can keep children mother or father.
child remain with mother.
Contents

• Brief History of Legal Text Collections
• Data Acquisition
• Corpus
• Statistics
• Conclusion
<table>
<thead>
<tr>
<th>Categories</th>
<th>Documents</th>
<th>No of Words</th>
<th>No of Queries</th>
<th>No of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>896</td>
<td>57517</td>
<td>645</td>
<td>50998</td>
</tr>
<tr>
<td>Admiralty</td>
<td>18</td>
<td>1254</td>
<td>18</td>
<td>1496</td>
</tr>
<tr>
<td>Appeals Writs</td>
<td>396</td>
<td>24830</td>
<td>201</td>
<td>21293</td>
</tr>
<tr>
<td>Aviation</td>
<td>12</td>
<td>708</td>
<td>11</td>
<td>972</td>
</tr>
<tr>
<td>Banking</td>
<td>6255</td>
<td>421697</td>
<td>2815</td>
<td>235309</td>
</tr>
<tr>
<td>Bankruptcy</td>
<td>226</td>
<td>15693</td>
<td>177</td>
<td>13735</td>
</tr>
<tr>
<td>Civil Rights</td>
<td>6163</td>
<td>392491</td>
<td>3946</td>
<td>385295</td>
</tr>
<tr>
<td>Computer Technology</td>
<td>393</td>
<td>26471</td>
<td>257</td>
<td>18784</td>
</tr>
<tr>
<td>Constitutional</td>
<td>1030</td>
<td>63242</td>
<td>587</td>
<td>38937</td>
</tr>
<tr>
<td>Construction</td>
<td>588</td>
<td>37687</td>
<td>429</td>
<td>31091</td>
</tr>
<tr>
<td>Consumer</td>
<td>2261</td>
<td>148758</td>
<td>1104</td>
<td>118668</td>
</tr>
<tr>
<td>Criminal</td>
<td>14273</td>
<td>810440</td>
<td>5908</td>
<td>596509</td>
</tr>
<tr>
<td>Family Divorce Child Custody Adoption</td>
<td>8649</td>
<td>635549</td>
<td>4212</td>
<td>634230</td>
</tr>
<tr>
<td>Labour</td>
<td>56</td>
<td>3611</td>
<td>43</td>
<td>10013</td>
</tr>
<tr>
<td>Tax</td>
<td>30</td>
<td>1634</td>
<td>22</td>
<td>3335</td>
</tr>
</tbody>
</table>
Contents

• Brief History of Legal Text Collections
• Data Acquisition
• Corpus
• Statistics
• Conclusion
Conclusion

• The collection can be used for evaluating various legal search activities from online discussion forums.
• Can be used for adhoc retrieval task
• It can also be used for automatic query and categorization task
Thank you!!