Foundations of Information Retrieval

Exercise 1

Exercise session: 03.11.2016
Questions to: Markus Rokicki, rokicki@L3S.de

1. Query optimization

1. A document collection with 125,000 documents contains film reviews. Given is the following query:

   \[(\text{NOT horror}) \land (\text{film} \lor \text{animation}) \land (\text{action} \lor \text{comedy})\]

Specify the most efficient order of execution for this query that can be determined from the following table:

<table>
<thead>
<tr>
<th>Term</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>horror</td>
<td>75,000</td>
</tr>
<tr>
<td>film</td>
<td>62,000</td>
</tr>
<tr>
<td>animation</td>
<td>3,000</td>
</tr>
<tr>
<td>action</td>
<td>41,000</td>
</tr>
<tr>
<td>comedy</td>
<td>40,000</td>
</tr>
</tbody>
</table>

Is the order you proposed always optimal?

2. Inverted index

Given is the following document collection:

\textbf{D1:}
\textit{Ice Age 4 was released in 34 territories.}
\textbf{D2:}
\textit{Ice Age 4 (original Ice Age: Continental Drift) is a 2012 American computer-animated comedy film.}

Create an inverted index for this document collection. \textbf{Tokenization rules}: word wise, case-folding, ignore punctuation. \textbf{Stop list}: was, in, is, a. Include TF and DF values at a suitable position in the index.

Which search results can be obtained from this index for the following queries?

\(Q_1=\text{Ice Age}\)
\(Q_2=\text{Tokyo Drift}\)