Seminar: Selected Topics in IR
Presentation of topics

Alexander Albrecht

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH
FB Informationssysteme
Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam
Alexander.Albrecht@hpi.uni-potsdam.de
Models for Document & Query Representation
Models for Document & Query Representation
Models for Document & Query Representation

- **Query**: Information Retrieval
  - Naumann

- **Number of occurrences**

<table>
<thead>
<tr>
<th></th>
<th>d₁</th>
<th>d₂</th>
<th>d₃</th>
<th>d₄</th>
<th>d_Seminar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naumann</td>
<td>231</td>
<td>66</td>
<td>15</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Information</td>
<td>78</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Retrieval</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

- **Challenges**
  - Ranking query results
  - Importance of query terms
  - Finding similar documents
  - ...

- Model for document & query representation needed!
Models for Document & Query Representation

- **(Term) Vector-Space Model**
  - Calculate similarity between query and documents that are represented as weighted term vectors

- **Binary Independence Model**
  - Document $d$ is relevant iff $P(R=1 \mid d, q) > P(R=0 \mid d, q)$
  - Compute $P(R \mid d, q)$ for all documents $d$

- **Language Model**
  - Compute $P(q \mid d)$ – probability that the language model of document $d$ generates query $q$
Text Classification Problem

- **Example: E-Mail Classification**
  - Document space $X$: E-Mails
  - Classes $C = \{\text{Spam, Private, Business, Advertising}\}$
  - Learn a classifier $\gamma$ that maps E-Mails to classes, $\gamma: X \rightarrow C$

- **Learning algorithms**
  - Naive Bayes (NB)
  - Support Vector Machines (SVM)
  - Supervised Learning: A supervisor serves as a teacher directing the learning process based on a training set $D$ of labeled documents $<d, c> \in X \times C$ in order to produce a learned classifier $\gamma$ using NB or SVM.

- Your presentation should include: Experimental results of an exemplary implementation using one of the presented approaches.