Information Retrieval
Course presentation
João Magalhães
Relevance vs similarity

What is the best [search space + dissimilarity function] to compute the relevance of documents for a given user information need?
What makes a good search application?

- **Efficiency**: application replies to user queries without noticeable delays.
  - 1 sec is the “limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer”

- **Effectiveness**: application replies to user queries with relevant answers.
  - This depends on the interpretation of the user query and the stored information.
The tasks of a search application

• **Collect** data for storage
  • Crawler

• Analyse collected data and compute the **relevant information**
  • Information analysis

• Store data in an **efficient** manner
  • Indexing

• Process **user** information needs
  • Querying

• Find the documents that best **match** the user information need
  • Ranking
Web crawling

- Begin with known “seed” URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch “robots.txt”
- Fetch each URL on the queue and repeat
Information analysis

• This stage deals with the extraction of the information to be made searchable

• Extract meaningful words, pairs of words or n-grams

• Extract images and their main characteristics

• Link visual characteristics and text data
Indexing

• This stage creates an index to quickly locate relevant documents

• An index is an aggregation of several data structures (e.g. several B-trees)

• Index compression is used to reduce the amount of space and the time needed to compute similarities

• The distribution of the index pages across a cluster improves the search engine responsiveness
Querying

- Conversion of the user query into the internal search space
  - Parsing

- Usage history
  - Cookies, profiles, etc.

- User intention
  - What type of task is the user doing?
Ranking

• Once the user query is converted into the internal search space...
  • The ranking function sorts the information according to its relevance to the user query

• Ranking functions should model the human notion of relevance
  • We don’t really know the mathematical form of the human notion of similarity...
Putting all together...
References

• Slides and articles provided during classes.

• Books:


Course grading

• The course has **two mandatory components**:  
  • Theoretical part (1 test or 1 exam): 40% \(^{\text{(minimum grade > 9.0)}}\)  
  • Labs (groups of 3 students): 60% \(^{\text{(minimum grade > 9.0)}}\)  

• Theory test/exam:  
  • Test: 12 December  
  • Exam: date to be defined  

• Additional rules:  
  • You may use one sided A4 sheet handwritten by you with your notes.  
  • It must be handed at the end of the test.  

• Individual mini-lab grading (minimum grade > 8.0)  
  • 30% implementation + 20% report + 20% questions + 30% discussion
Laboratories: News search

• Implement a search engine to search online news.

• Understand the roles of each component of a search engine in the performance of the search results.

• Labs are done incrementally. Each week new functionalities will be added to the initial implementation.

• There will be 4 mini-labs throughout the semester.
  • The submission date of each mini-lab is three days after the last lab class of the corresponding mini-lab.
<table>
<thead>
<tr>
<th>Week</th>
<th>#</th>
<th>Lectures</th>
<th>In-class labs</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-Sep-18</td>
<td>1</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>19-Sep-18</td>
<td>2</td>
<td>Basic techniques (Lucene examples)</td>
<td>Environment setup</td>
</tr>
<tr>
<td>26-Sep-18</td>
<td>3</td>
<td>Evaluation</td>
<td>Text pre-processing, VSM</td>
</tr>
<tr>
<td>03-Oct-18</td>
<td>4</td>
<td>Retrieval models: LM + BIM + BM25</td>
<td>Evaluation scripts</td>
</tr>
<tr>
<td>10-Oct-18</td>
<td>5</td>
<td>Implementation of Ret Models</td>
<td>Retrieval models</td>
</tr>
<tr>
<td>17-Oct-18</td>
<td>6</td>
<td>Query processing and taxonomies</td>
<td>Retrieval models</td>
</tr>
<tr>
<td>24-Oct-18</td>
<td></td>
<td>Reports discussion</td>
<td>Query expansion</td>
</tr>
<tr>
<td>31-Oct-18</td>
<td>7</td>
<td>Information duplicates</td>
<td>Query expansion</td>
</tr>
<tr>
<td>07-Nov-18</td>
<td>8</td>
<td>Multiple fields and rank fusion</td>
<td>Query expansion</td>
</tr>
<tr>
<td>14-Nov-18</td>
<td>9</td>
<td>Rank expansion</td>
<td>Ranking multiple fields</td>
</tr>
<tr>
<td>21-Nov-18</td>
<td>10</td>
<td>Static and distributed indexing</td>
<td>Ranking multiple fields</td>
</tr>
<tr>
<td>28-Nov-18</td>
<td>11</td>
<td>Efficient query processing</td>
<td>Ranking multiple fields</td>
</tr>
<tr>
<td>05-Dec-18</td>
<td>12</td>
<td>Elasticsearch vs Lucene</td>
<td>Ranking multiple fields</td>
</tr>
<tr>
<td>12-Dec-18</td>
<td></td>
<td>Test + Reports discussion</td>
<td></td>
</tr>
</tbody>
</table>
Summary

• “Information Retrieval” course context
• Course objectives and plan
• Grading
• Labs