Visualise a web site with tag clouds generated by R

Sigbert Klinke\textsuperscript{1,2}

\textsuperscript{1} Institute for Statistics and Econometrics, School of Business and Economics, Humboldt-Universit\"at zu Berlin
\textsuperscript{2} Business and Human Resource Education, Dept. of Law and Economics, Johannes-Gutenberg-Universit\"at Mainz

useR! 2009
Session: Textmining
08-10 Jul 2009, Rennes, France
Introduction

Problem: Redirection of web users

- Changes to web site structure produces errors on access
- How can we redirect the users to a large number of pages?
- Solution: Use a tag cloud where the size of an entry corresponds to the number of visits in the past year

Visualise a web site with tag clouds generated by R

Humboldt-Universität zu Berlin
Problem: Teaching statistics

- Wikipedia is often a (starting) source for students
- Dictionary structure does not allow for an overview of a topic
- Solution: Use a tag cloud to visualise the neighbourhood of a page

Links to Moment, Wahrscheinlichkeitsverteilung, ...
Wikipedia structure

- Expected value
- Moment
- Moment generating function
- Inverse gaussian distribution
- Normal distribution
- Category: Stochastics
  - Category: Statistics
  - Category: Geostatistics
- Wiki page: Category: Stochastic process
- Wiki page: Category: Probability distributions

Visualise a web site with tag clouds generated by R

Humboldt-Universität zu Berlin
**Introduction**

**Work flow**

- PHP script crawls Wikipedia and stores the link structure
  - crawler from http://w-shadow.com using cURL
  - store in csv format: fromPage ; toPage
- R generates a tag cloud for each page
  - load linkstructure read.csv
  - build link network: igraph by Gabor Csardi
    - for importance compute pagerank page.rank (font size)
    - extract neighbourhood graph.neighborhood (of distance 1)
    - compute (bivariate) positions layout.mds (location)
igraph (layout.mds)

- create HTML tag clouds
- create dendrogram from positions (table-based)
- use a top/bottom - left/right approach (compact)
- use one dimensional MDS (oneliner)
Tag cloud: table-based

- Most page titles are long (e.g. Moment (mathematics))
- Take hyphenation into account
Introduction

\textbf{TEX hyphenation}

- utilise the \texttt{TEX hyphenation}
- Perl program available
  - \texttt{TeX::hyphen} by Jan Pazdziora
  - \texttt{hyphen.pl} with german hyphenation by Tilman Kranz
- add &\#8203; (zero width space)
algorithm needs some more polishing
Tag cloud: one liner

Visualise a web site with tag clouds generated by R
**createTagCloud parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>g</code></td>
<td>igraph object</td>
</tr>
<tr>
<td><code>graph.order</code></td>
<td>size of neighbourhood (currently only 1)</td>
</tr>
<tr>
<td><code>graph.layout</code></td>
<td>layout function from igraph (<code>layout.mds</code>)</td>
</tr>
<tr>
<td><code>fontsize.method</code></td>
<td>method to compute the font size (<code>page.rank.vector</code>)</td>
</tr>
<tr>
<td><code>fontsize.transform</code></td>
<td>transformation method for font size (log10)</td>
</tr>
<tr>
<td><code>fontsize.min</code></td>
<td>font size minimum (7.5)</td>
</tr>
<tr>
<td><code>fontsize.max</code></td>
<td>font size maximum (20.5)</td>
</tr>
<tr>
<td><code>buildHTML.method</code></td>
<td>method to build tag cloud(s) (<code>one</code>)</td>
</tr>
<tr>
<td><code>buildHTML.landscape</code></td>
<td>landscape format (<code>T</code>)</td>
</tr>
<tr>
<td><code>buildHTML.hyphenate</code></td>
<td>should \TeX\ hyphenation be applied (<code>TRUE</code>)</td>
</tr>
<tr>
<td><code>file.html</code></td>
<td>name(s) of HTML/PNG file(s)</td>
</tr>
<tr>
<td><code>file.png</code></td>
<td>(<code>vertex%i.html</code>, <code>vertex%i.png</code>)</td>
</tr>
<tr>
<td><code>no</code></td>
<td>index of vertices for which tag clouds are generated (<code>NA</code>)</td>
</tr>
<tr>
<td>...</td>
<td>further parameters</td>
</tr>
</tbody>
</table>

Visualise a web site with tag clouds generated by R

Humboldt-Universität zu Berlin
Outlook

- Use Wikipedia XML dump instead own web crawler
- Account for redirects in Wikipedia
- Add “virtual” links
  - Analyse text (TreeTagger)
- Colour links in tag cloud (Inbound, Outbound, Bidirectional)
- Increase neighbourhood
- Add MediaWiki output
- Improve hyphenations?
Visualise a web site with tag clouds generated by R

Literature/Links