Personalized recommendation of interesting texts

Michal Kompan
supervisor: prof. Mária Bieliková

Contribution
- Short and high representative article vectors
- Real-time content-based recommendation
- Fast similarity computation

Recommendation

- Article list
  1. L
  2. M
  3. G
  4. J

- Similar articles for recommended and visited before
- Similar articles for visited and not recommended before

Evaluation
- Recommendation: Synthetic tests – SME.SK (3 days)
- Article similarity: SME.SK (Author’s choice, manually annotated)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles</td>
<td>9</td>
<td>63</td>
<td>43.23</td>
<td>64.05</td>
<td>50.28</td>
<td>36.27</td>
<td>46.49</td>
<td>46.31</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>48</td>
<td>40.26</td>
<td>63.26</td>
<td>50.94</td>
<td>37.44</td>
<td>44.97</td>
<td>47.04</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>39</td>
<td>39.73</td>
<td>62.12</td>
<td>51.36</td>
<td>39.92</td>
<td>48.80</td>
<td>48.63</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>24</td>
<td>38.02</td>
<td>59.91</td>
<td>59.95</td>
<td>40.23</td>
<td>46.53</td>
<td>48.14</td>
</tr>
</tbody>
</table>

| Articles   | 9               | 63              | 1.43           | 1.83            | 0.84        | 0.77        | 1.06           | 1.08            |
|            | 24              | 48              | 0.76           | 1.81            | 0.47        | 0.80        | 0.58           | 1.11            |
|            | 33              | 39              | 0.67           | 1.68            | 0.49        | 0.85        | 0.57           | 1.13            |
|            | 48              | 24              | 0.5            | 1.53            | 0.64        | 1.34        | 0.56           | 1.43            |

Article similarity

1. Article preprocessing
   - stopwords elimination
   - lemmatization
   - name and place extraction
   - keywords extraction

2. Vector representation
   - Title – lemmatized words from article title (approx 5)
   - TF of title words in the content – normalized
   - Names/Places – words starting with uppercase and no sentence end before
   - Keywords – 10 most relevant keywords (TF-IDF)
   - Category – “tree-based” category vector
   - CLI – Coleman-Liau readability index

3. Similarity computation
   - cosine similarity
   - real-time

\[
\text{similarity} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}
\]