Lesson 4: Descriptive Modelling of Similarity of Text
Unit 5: Comparing similarity measures

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany
Completing this unit you should …

• Understand that different modeling choices can produce very different results.

• Have a feeling how you could statistically compare the differences of the models.

• Know how you could extract keywords from documents with the tf-idf approach.

• Try to argue which model you like best in a certain scenario.
Examining an exemple query (set based)

- Query = “magnus carlsen chess norwegian grandmaster elo”
- First sentence of resulting articles:
  - **Norwegian** might mean:
  - In 2011, Nakamura won the Tata Steel **Grandmaster** A tournament in Wijk aan Zee.
  - Krunoslav Hulak (25 May 1951 – 23 October 2015) was a Croatian **chess** master.
  - Royal Caribbean International () is a **Norwegian**-American cruise ship company based in Miami, Florida.
  - Dmitry Jakovenko (born 1983) is a Russian **chess** grandmaster.
Examining an exemple query (tfidf)

• Query = “magnus carlsen chess norwegian grandmaster elo”
• First sentence of resulting articles:
  – **Norwegian** might mean:
  – Maxime Vachier-Lagrave (born 21 October 1990) is a French **chess Grandmaster** and the 2009 World Junior **Chess** Champion.
  – Wesley So (born 9 October 1993) is a Filipino **chess grandmaster**.
  – Levente Lengyel (13 June 1933 – 18 August 2014) was a Hungarian **chess** player.
  – Shakhriyar Mamadyarov (born 12 April 1985 in Sumgayit, Azerbaijan), is a **chess grandmaster**.
Examining an exemple query (LM)

- Query = “magnus carlsen chess norwegian grandmaster elo”
- First sentence of resulting articles:
  – Magnus Carlsen (born Sven Magnus Øen Carlsen, 30 November 1990) is the World Chess Champion.
Examining an exemple query (+1 LM)

• Query = “magnus carlsen chess norwegian grandmaster elo”

• First sentence of resulting articles:
  – **Norwegian** might mean:
  – A **chess** club is a place where people come to play **chess**.
  – A **chess** tournament is a between **chess** players.
  – Fast **chess** is similar to a normal game of **chess**, but played at a faster than usual rate.
  – The Manhattan **Chess** Club is the second-oldest **chess** club in the United States (after the Mechanics' Institute **Chess** Club in San Francisco).
## Similar articles to Magnus Carlsen

<table>
<thead>
<tr>
<th>rank</th>
<th>set</th>
<th>tf-idf</th>
<th>Smoothed LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Magnus Carlsen</td>
<td>Magnus Carlsen</td>
<td>Mammalogy</td>
</tr>
<tr>
<td>2</td>
<td>Viswanathan Anand</td>
<td>Anatoly Karpov</td>
<td>Municipalities in Switzerland</td>
</tr>
<tr>
<td>3</td>
<td>Veslin Topalov</td>
<td>Nigel David Short</td>
<td>Wichita</td>
</tr>
<tr>
<td>4</td>
<td>Luke McShane</td>
<td>Shakhriyar Mamadyarov</td>
<td>Reichstag</td>
</tr>
<tr>
<td>5</td>
<td>Ruslan Ponomariov</td>
<td>Hikaru Nakamura</td>
<td>Yellow</td>
</tr>
</tbody>
</table>
Finding the characteristic words for a document

• Set based
  – Not applicable since all words are “equal”

• Unigram Language Model
  – Basically this is a normalized term frequency
  – In all articles words like “the, a, of,…” will be most probable

• TF-IDF
  – Inverse document frequency is a nice weighting scheme to boost words that occur just in this or a few documents
## Characteristic words for Magnus Carlsen

<table>
<thead>
<tr>
<th>rank</th>
<th>TF-IDF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>carlsen</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>chess</td>
<td>in</td>
</tr>
<tr>
<td>3</td>
<td>rating</td>
<td>to</td>
</tr>
<tr>
<td>4</td>
<td>kasparov</td>
<td>carlsen</td>
</tr>
<tr>
<td>5</td>
<td>grandmaster</td>
<td>is</td>
</tr>
<tr>
<td>6</td>
<td>2800</td>
<td>he</td>
</tr>
<tr>
<td>7</td>
<td>world</td>
<td>a</td>
</tr>
<tr>
<td>8</td>
<td>anand</td>
<td>world</td>
</tr>
<tr>
<td>9</td>
<td>youngest</td>
<td>this</td>
</tr>
<tr>
<td>10</td>
<td>euros</td>
<td>of</td>
</tr>
</tbody>
</table>
Take away

• Different modeling choices can lead to very different results

• Modeling similarity is often a core step for a predictive model

• For a **reasonable comparison** we need
  – a **gold standard**
  – **Metrics for evaluation** from predictive models

• More about these ideas and topics in **Web information retrieval** next semester
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de
Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license.
- [https://commons.wikimedia.org/wiki/File:Synoptic_word-for-word.png](https://commons.wikimedia.org/wiki/File:Synoptic_word-for-word.png) By Alecmonroy (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons