Lesson4:
Descriptive Modelling of Similarity of Text
Unit5:
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

rank	set	tf-idf	Smoothed LM
1	Magnus Carlsen	Magnus Carlsen	Mammalogy
2	Viswanathan Anand	Anatoly Karpov	Municipalities in Switzerland
3	Veslin Topalov	Nigel David Short	Wichita
4	Luke McShane	Shakhriyar Mamadyarov	Reichstag
5	Ruslan Ponomariov	Hikaru Nakamura	Yellow

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

rank	TF-IDF	LM
1	carlsen	the
2	chess	in
3	rating	to
4	kasparov	carlsen
5	grandmaster	is
6	2800	he
7	world	a
8	anand	world
9	youngest	this
10	euros	of

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

WeST
People and Knowledge Networks

Copyright: