



Lesson4:
Descriptive Modelling of Similarity of Text
Unit2:
Set theoretic Models

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties





Completing this unit you should ...

- Understand how text documents can be modeled as sets
- Know the Jaccard coefficient as a similarity measure on sets
- Know a trick how to remember the formula
- Be aware of the possible outcomes of the Jaccard index
- As always be able to criticize your model



A set based Model for documents

- For a given Document $D_i = w_1 w_2 \dots w_n$
- We can define its word set by setting

$$W_i = \{w | w \in D_i\}$$

- Realize $|W_i| \leq n$
- Quiz: Why not equal to n?



A Simple Example

- $D_i =$ Magnus Carlsen is a chess player. He is from Norway.
- $W_i = \{$ Magnus, Carlsen, is, a, chess, player, he from, Norway $\}$



Boolean operations lead to Jaccard

- Intersection $|W_i \cap W_j|$ gives us the number of common words in the word sets of D_i and D_j
- Can this be a similarity measure?
- Seems good. The more words in common the more similar the documents would be.



Warning! Intersection is not a similarity

- D1 = I love Web Science
- D2 = Magnus Carlsen is a chess player.

$$|W_1 \cap W_1| = 4$$

$$|W_2 \cap W_2| = 6$$

$$|W_1 \cap W_1| \neq |W_2 \cap W_2|$$

- No equal self similarity!
- Can this be fixed?



Jaccard coefficient: Normalizing with Union

$$s(D_i, D_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$$

- s is always between 0 and 1
- Self similarity for all documents is 1
- Symmetry is given
- Maximality is given



How to remember which one is it?

- Is it $\frac{|W_i \cap W_j|}{|W_i \cup W_j|}$ or $\frac{|W_i \cup W_j|}{|W_i \cap W_j|}$?
- I had students failing exams because they could not remember.
- Key Idea: **Don't** learn the formula by heart
 - Chances are high you will mix it up
- Generally better: Understand where the formula comes from!



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:

- 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 By Alecmconroy (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
- <https://commons.wikimedia.org/wiki/File:Inner-product-angle.png> CC-BY-SA by CSTAR & Oleg Alexandrov