

„ALL THE WORLD'S A VECTOR“

Experimente zur Detektion von intertextuellen Shakespeare-Referenzen mithilfe von Word Embeddings

Bernhard Liebl¹, Manuel Burghardt¹
¹Computational Humanities, Universität Leipzig

1. Motivation: Quantitative Detektion von intertextuellen Shakespeare-Referenzen

2. Ansatz: Effiziente Suche optimaler Alignments mittels weicher Constraints und parametrisierter Ähnlichkeitsmetriken auf Basis von Word Embeddings

The Vectorian

By the pricking of my thumbs, something wicked this way comes.

Search

this this 100% way way 100% dies comes 66% A hiss in her mind faded away and left
 DET DET fasttext NOUN NOUN fasttext VERB VERB fasttext

88.1%

Terry Pratchett, *I Shall Wear Midnight*, Chapter 14, par. 80

she thought as she stared into the night-time gloom. By by 100% the the 100% stinking pricking 55% of of 100% my my 100% nose thumbs 59% something something 100% evil wicked 75%
 ADP ADP fasttext DET DET fasttext NOUN NOUN fasttext ADP ADP fasttext PRON PRON fasttext NOUN NOUN fasttext PRON PRON fasttext ADJ ADJ fasttext

this this 100% way way 100% goes comes 84% she added, to stop herself gibbering as she scanned the
 DET DET fasttext NOUN NOUN fasttext VERB VERB fasttext

74.7%

Jasper Fforde, *Thursday Next 4 - Something Rotten*, par. 3748

wood. 'By the pricking of my thumbs,' remarked Shakespae in by 64% an the 59% ominous tone pricking 50% of of 100% voice thumbs 57% something something 100% wicked wicked 100%
 ADP ADP fasttext DET DET fasttext -12.9% rel. NOUN NOUN fasttext ADP ADP fasttext NOUN NOUN fasttext PRON PRON fasttext ADJ ADJ fasttext

this this 100% way way 100% comes comes 100% 'There!' yelled Millon, pointing a quivering finger out of
 DET DET fasttext NOUN NOUN fasttext VERB VERB fasttext

3. Parameter

- Gewichtung übereinstimmender POS tags
- Penalties für nicht übereinstimmende Tokens
- Wahl des Word Embeddings
- Ähnlichkeitsmetrik für Vektoren des Embeddings

Alignment

Mismatch Length Penalty 50% rel. after 5 t.

Similarity Threshold 10%

Sub Match Weight 0

IDF Weight 0

Basis-Metrik

Fasttext WordNet

Similarity Measure

Cosine quantiles

NICDM

APSynP

Maximum

Similarity Falloff 1.00

POS

Ignore Determiners

Universal POS Mismatch Penalty 50%

POST STSS Weighting 50%

4. Optimale Alignments

Alignment und Scoring von Satzpaaren via Dynamic Programming:

