TU Wien Informatics

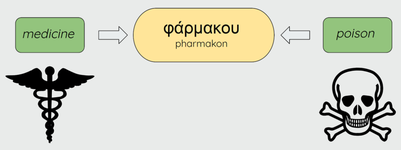
Text analysis using colexification networks

Armin Gander
Data Science

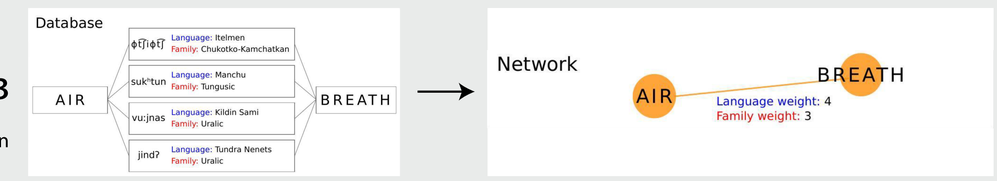
TU Wien Informatics
Institute of Information Systems Engineering
Electronic Commerce
Supervisor: Univ. Prof. Dr. Allan Hanbury
Contact: armingander@gmail.com

Colexification

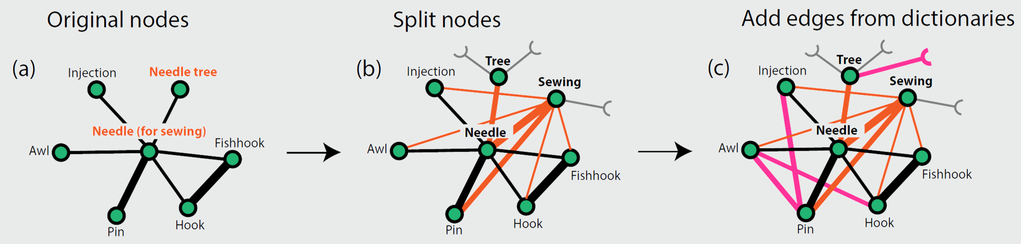
Colexification is a linguistic idea. One colexification instance happens when two different concepts are expressed using the same word in a language.



Colexification network construction



Colexification network modifications



Problem statement

Motivation

Machine learning perspective:	Black box NLP models lack interpretability
NLP perspective:	Corpus-based NLP models are inherently domain-specific
Linguistic perspective:	Idea that colexification occurrences hint to similarity in meaning has not been validated yet

Research questions

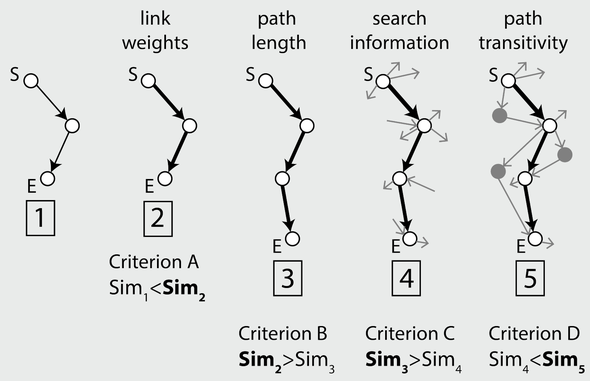
Do colexifications **encode meaning similarity** between concepts?

To which extent can colexification networks be used as a basis for **knowledge-based text analysis**?

What do the results tell us about the **historical development** of natural language?

Word similarity

The word similarity measure built on top of the colexification network fulfills several criteria.



Mathematically, the similarity between the nodes is computed as the stationary visiting distribution of each colexification networks node.

$$S_{ij} = U_{ij} + \beta \sum_{k \epsilon N_i} U_{ik} S_{ik}$$

U= normalized adjacency matrix
S = similarity between nodes
β = damping factor

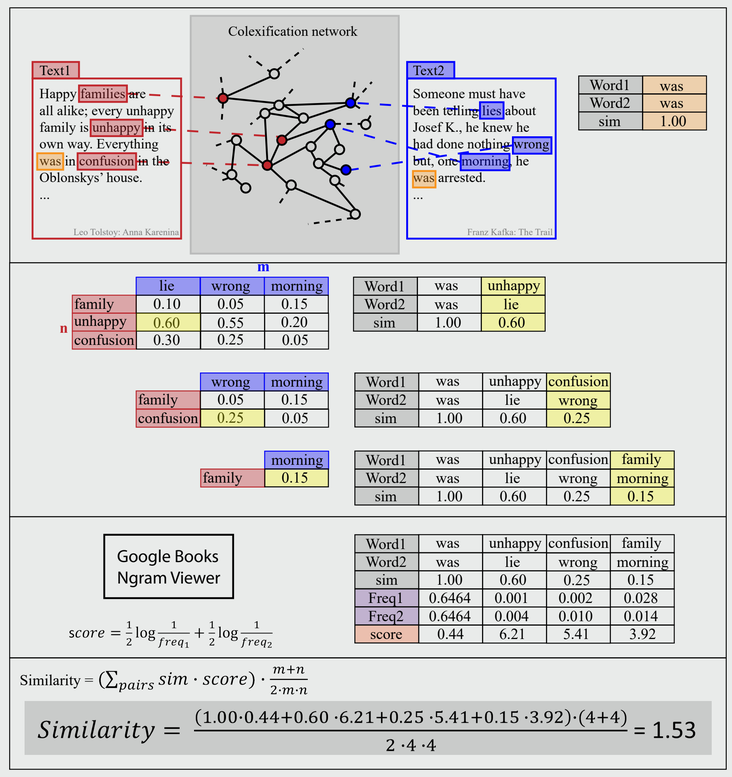
Rewrite as $S = U + \beta US$

Solve for S: $S = (\mathrm{I} - \beta U)^{-1} U$

Text similarity

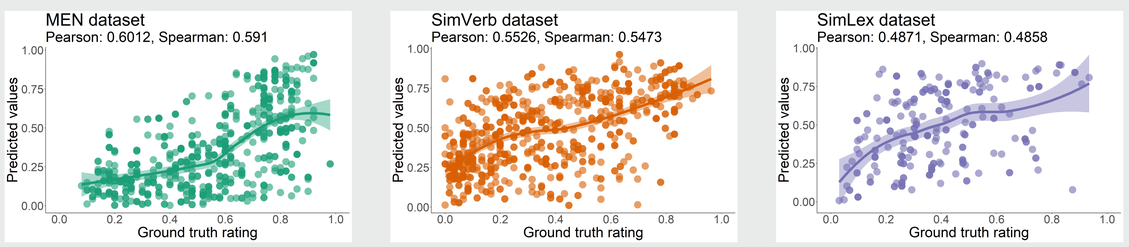
4 steps:

1. Identify exact matches, map words onto colexification network and retrieve similarity values
2. Set up similarity matrix and apply greedy elimination procedure to matching problem
3. Retrieve Google Ngram word frequency data and compute frequency scores
4. Compute final text similarity value



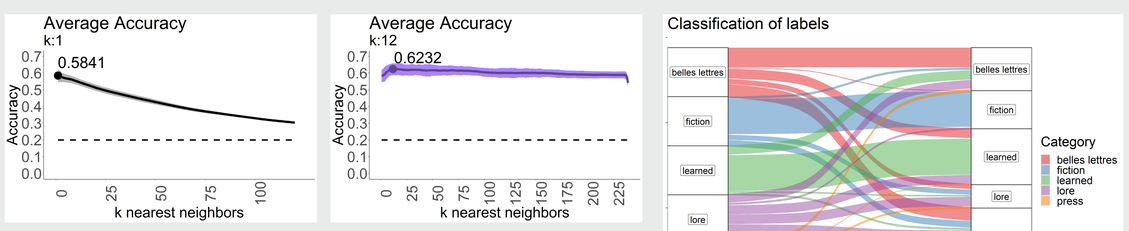
Validation I

Word similarity measure validation



Text similarity measure validation

Brown Corpus genre classification task



Classic kNN approach:
predicts genre as majority class genre of neighboring books

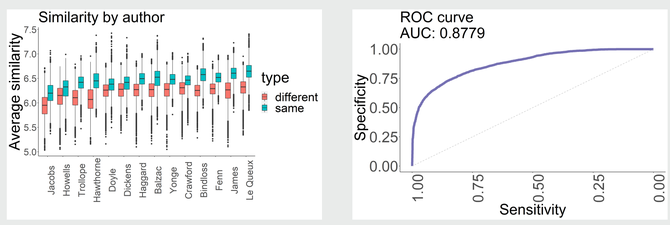
Weighted kNN approach:
predicts genre as weighted combination of mean class similarity and logarithmic class frequency

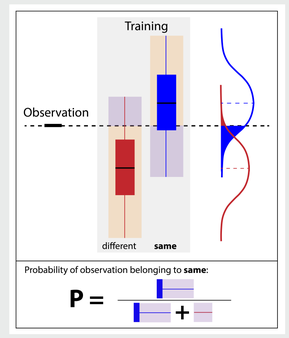
Fine-grained analysis book genre prediction errors

Validation II

Text similarity measure validation

Gutenberg corpus authorship prediction

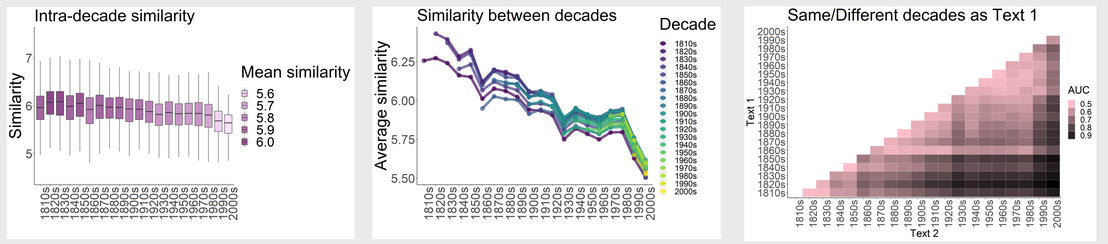




Predcition probability computation for binary classification task

Exploration: Historical Analysis

Historical analysis of American English used in literary fiction from 1810s to 2000s using the Corpus of Historical American English (COHA)



Intra-decade similarity distributions

Inter-decade similarity trend

Book publication binary classification task