Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Schweizerische Nationalbibliothek NB

Automatic Classification of e-Dissertations

Dr. Marcel Gygli

Overview

Problem Statements

Large Language Models & Word Embeddings

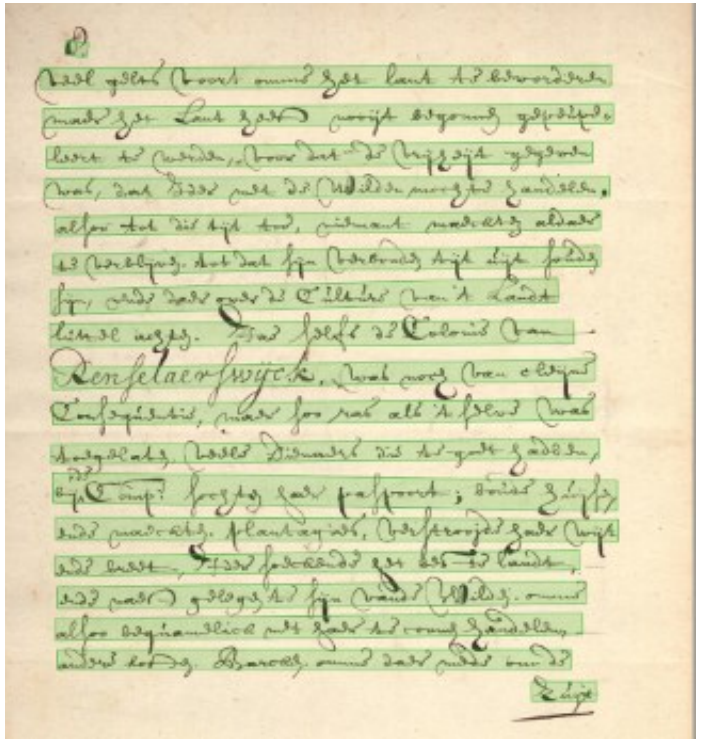
Automatic Classification of e-Dissertations

Annif & Embedding Approach

Results

About Me

- Innovation Fellow Swiss National Library (until 31.07)
- Professor (Tenure Track) at BFH (since 01.05)
  - AI In the Public Sector
- PhD in Computer Science (2015 – 2020)
  - Work on Historical Documents




2022/21
ISSN 1661-8211 | 122. Jahrgang | 15. November 2022
Alle Medien Alle Medien Alle Medien Alle Medien Alle Medien Alle Medien Alle
Das Schweizer Buch Das Schweizer Buch Das
Schweizerische Nationalbibliografie herausgegeben von der Schweizerischen Nationalbibliothek NB Schweizerische Nationalbibliografie herausgegeben
Tous médias Tous médias Tous médias Tous médias Tous médias Tous médias Tous
Le Livre suisse Le Livre suisse Le Livre suisse
Bibliographie nationale suisse éditée par la Bibliothèque nationale suisse BN Bibliographie nationale suisse éditée par la Bibliothèque nationale suisse
Tutti i media Tutti i media Tutti i media Tutti i media Tutti i media Tutti i media Tutti
Il Libro svizzero Il Libro svizzero Il Libro svizz
Bibliografia nazionale svizzera pubblicata dalla Biblioteca nazionale svizzera BN Bibliografia nazionale svizzera pubblicata dalla Biblioteca nazionale
Tut ils meds Tut ils meds Tut ils meds Tut ils meds Tut ils meds Tut ils meds Tut ils
Il Cudesch svizzer Il Cudesch svizzer Il Cudes
Bibliografia naziunala svizra edi da la Biblioteca naziunala svizra BN Bibliografia naziunala svizra edi da la Biblioteca naziunala svizra BN Bibliogra
All media All media All media All media All media All media All media All media
The Swiss Book The Swiss Book The Swiss Bo
Swiss National Bibliography published by the Swiss National Library SNL Swiss National Bibliography published by the Swiss National Library SNL
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Dipartimento federale dell'interno DFI
Departament federal da l'intern DFI
Schweizerische Nationalbibliothek NB
Bibliothèque nationale suisse BN
Biblioteca nazionale svizzera BN
Biblioteca naziunala svizra BN

Reproducible Research (RR) has been a research topic within the computational sciences for a long time. It deals with the problem of ensuring that other researchers can independently verify results generated on a computer.

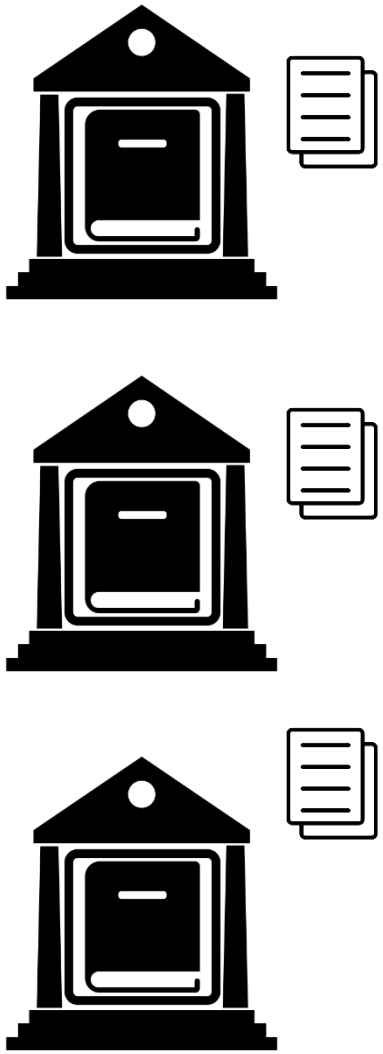
A longstanding problem that reproducible research tries to address is that many published scientific results can not be verified by reading the publication itself. This is due to a set of various reasons: not enough space in the publication to explain all implementation details or aversion to sharing source code publicly. Thus, it becomes ever more difficult for researchers to reproduce results from a presented method, compute results on new data, or use it as part of a new system. In this thesis, we provide an overview of these issues with a particular focus on the area of Document Image Analysis (DIA), where these problems arise as well. We analyze approaches proposed as solutions to these problems, and identify open issues, especially in the area of providing access to executable versions of research code. Using this analysis, we design the features of a system that will be able to support DIA researchers in performing reproducible research. This design is implemented in a Web Service framework called DIVAServices that allows for the publication and execution of research methods through a unified interface.

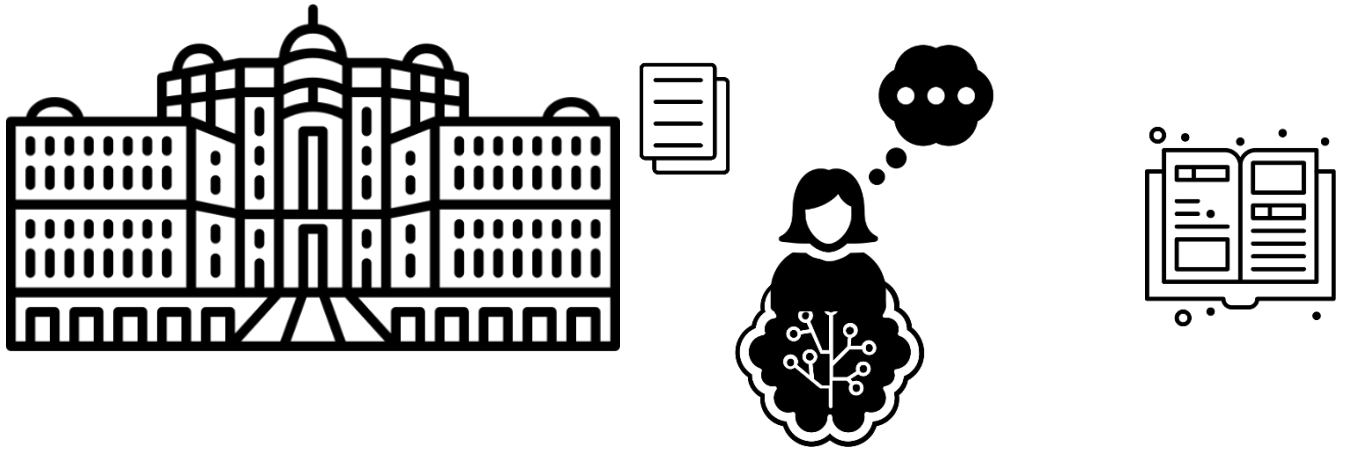
004 Informatik
Informatique
Informatica
Informatica
Data processing and computer science

9 NB 18132627403976

Gygli, Marcel, 1987-. – Web services for reproducible research : building an open source reproducibilty framework for document image analysis / by Marcel Gygli. – Fribourg, [2020]. – xv, 237 Seiten : Illustrationen ; 30 cm

Dissertation No: 2193 University of Fribourg (Switzerland), 2020. – Literaturverzeichnis. – Englischer Text mit englischer und deutscher Zusammenfassung

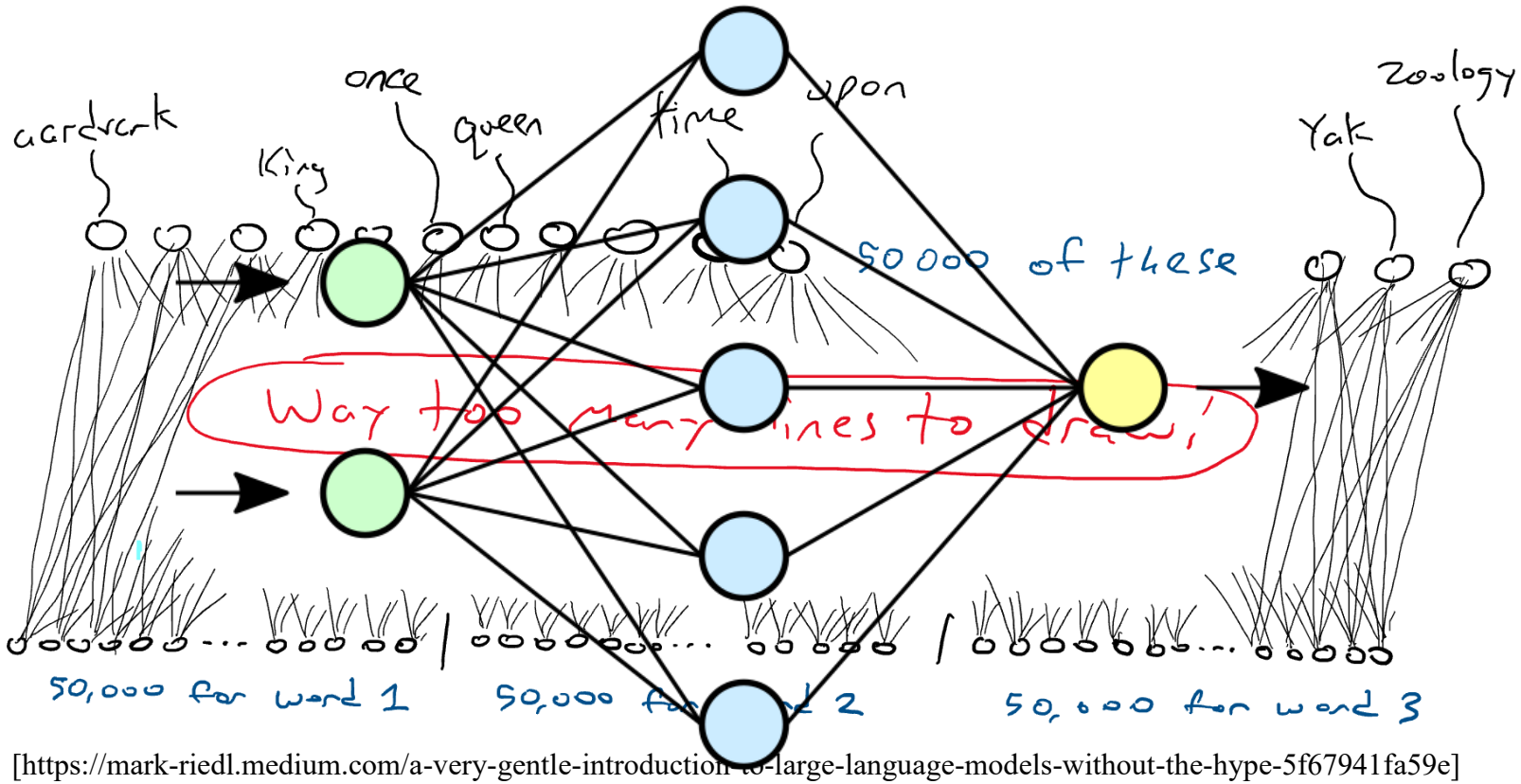




How do Large Language Models work?

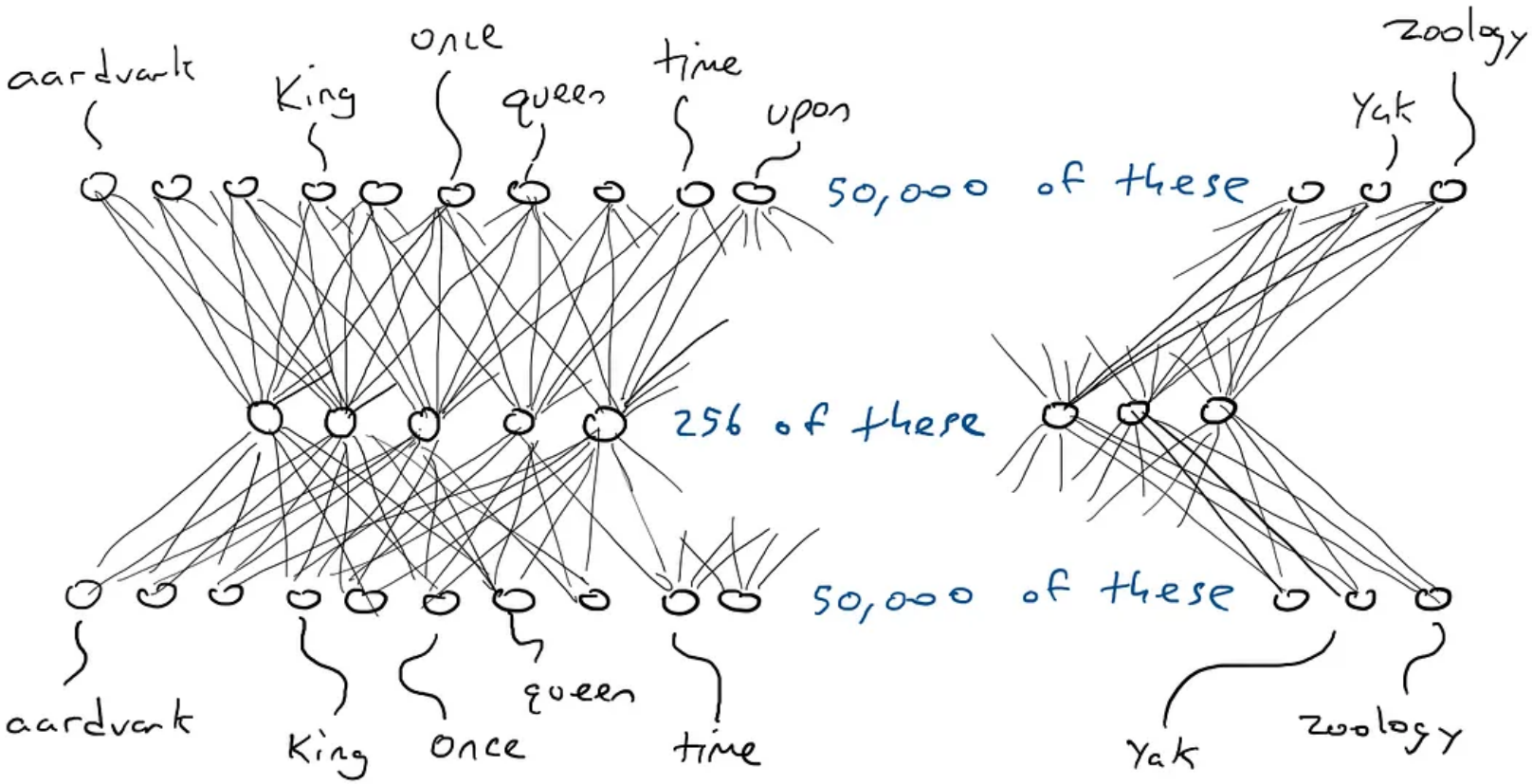
Once upon a ...

Als Neurales Netzwerk nicht umsetzbar

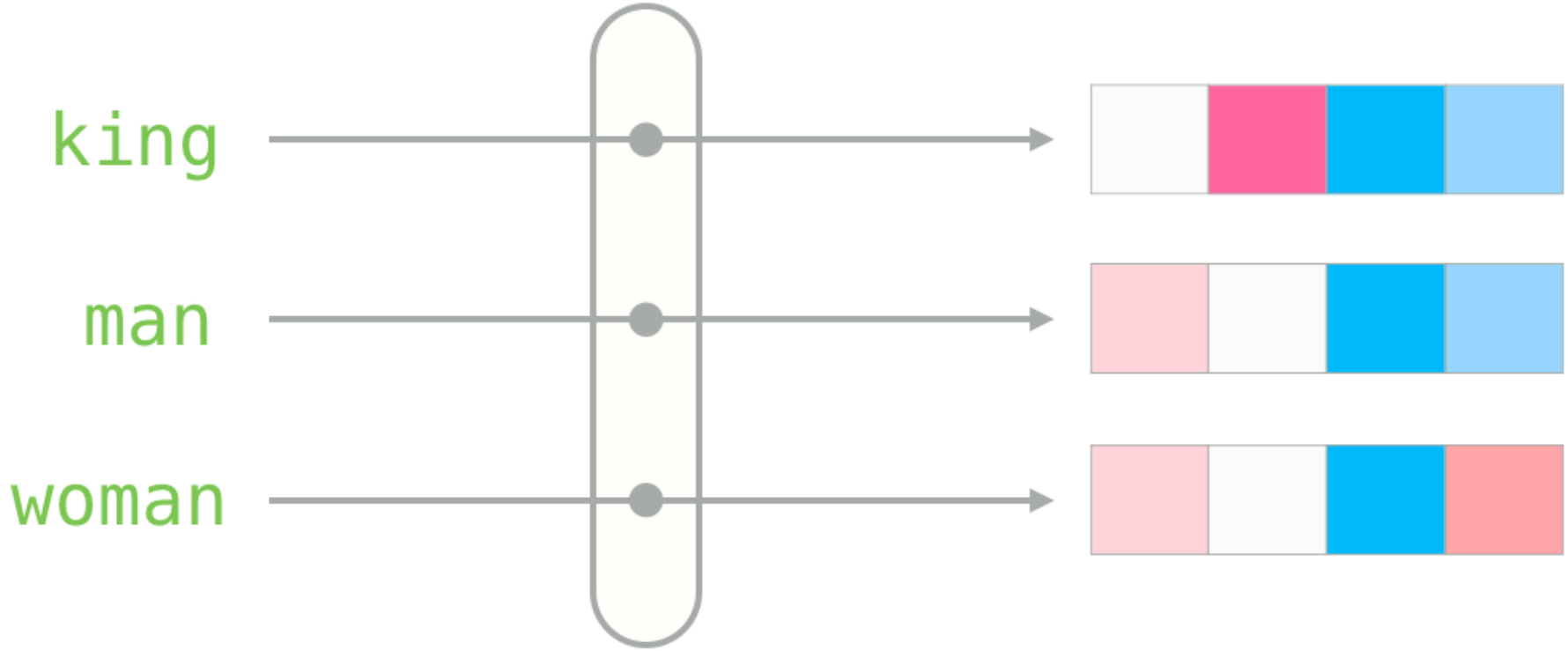


[https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e]

Wir benötigen eine Zwischenräpresentation



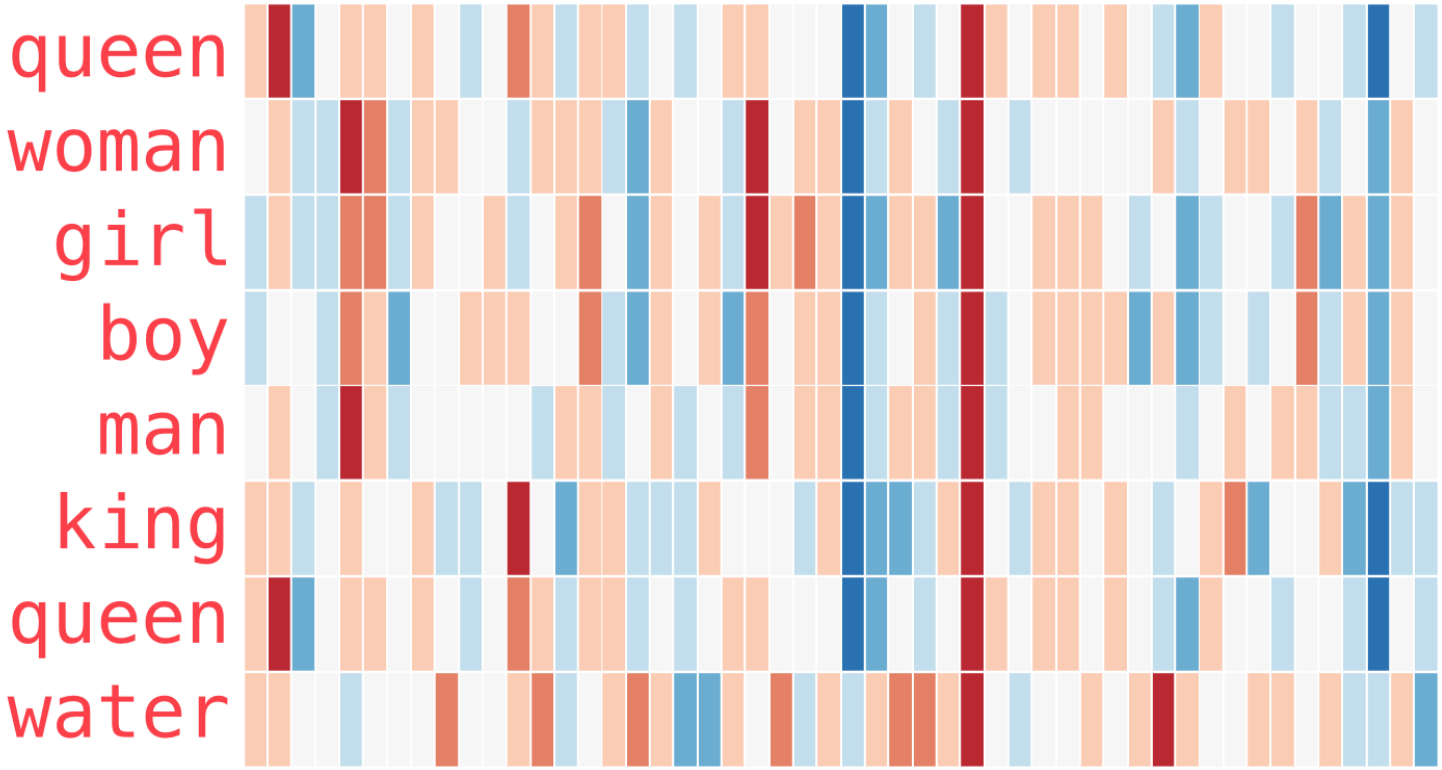
Word-Embeddings als Räpresentation



Word-Embeddings im Details

King

[0.50451 , 0.68607 , -0.59517 , -0.022801, ... , -1.6106 , -0.64426 , -0.51042]



Annif

Open-Source Subject Indexing Tool

Developed by National Library of Finland

Different Algorithms Implemented

Omikuji (Decision Tree Based) worked best

Embedding Based Approach

Used `sentence-transformers/all-mpnet-base-v2`

Hugging Face

Embed all training documents

Inference using Cosine-Similarity

Workflow for PDFs using SpaCy

Results

Dataset	Annif	Embeddings	# Documents
NB	733 / 972 (56.8% / 75.3%)	660 / 920 (41.1% / 71.2%)	1291
NB+PDF	**2025 / 2457 (63.5% / 77.1%)**	1635 / 2152 (51.3% / 67.6%)	3185
NB+PDF (GER)	993 / 1277 (55.0% / 70.7%)	**1342 / 1482 (74.3% / 82.1%)**	1806

Discussion

The appropriate Dewey Decimal Classification (DDC) number for this abstract would likely fall within the range of 000-099, which encompasses Computer Science and Information.

A specific DDC number within this range that could be relevant to the topic of reproducible research in the computational sciences and Document Image Analysis (DIA) would be DDC 004.019, which focuses on software engineering and development.

It's important to note that the DDC system is extensive, and the appropriate classification may depend on the specific focus and content of the thesis. Consulting the complete DDC system or seeking guidance from a librarian specializing in computer science or information science may help in selecting a more precise DDC number for this thesis.