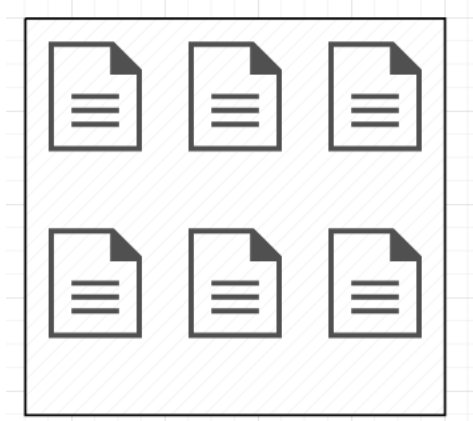
Kontextabhängige Termgewichtungen bei Total-Recall-Suchen

Wilhelm Beiche

12. Februar 2022

Einleitung

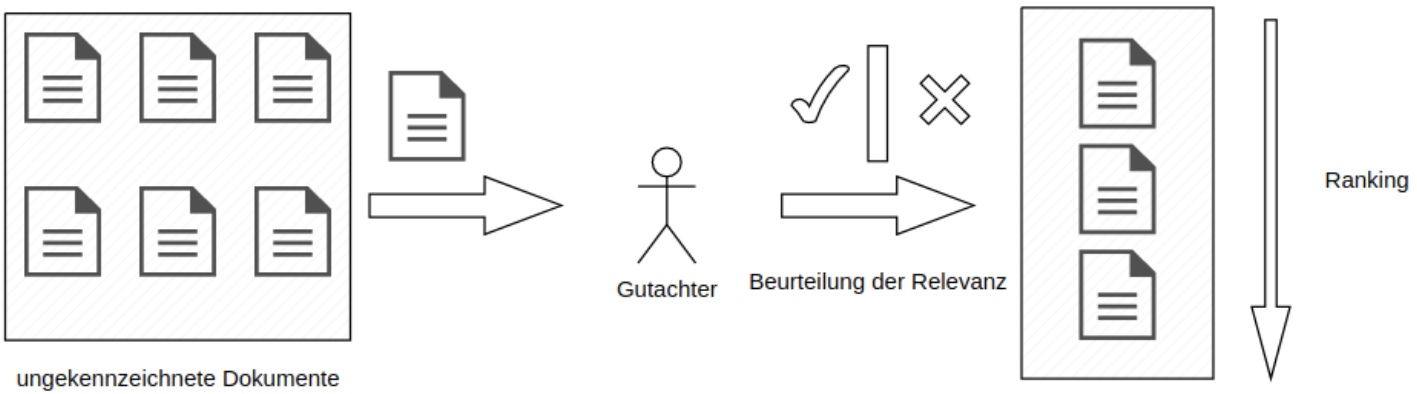
- Suchanfrage:
  Welche Merkmale werden durch Genom Editierung in Modellpflanzen sowie in landwirtschaftlichen Nutzpflanzen verändert?
- Die Suchanfrage wird von speziellen Suchmaschinen genutzt, um alle relevante Dokumente aus dem Korpus zu finden
- Total Recall: Extraktion aller relevanten Dokumente für das Informationsbedürfnis aus dem Korpus



Korpus

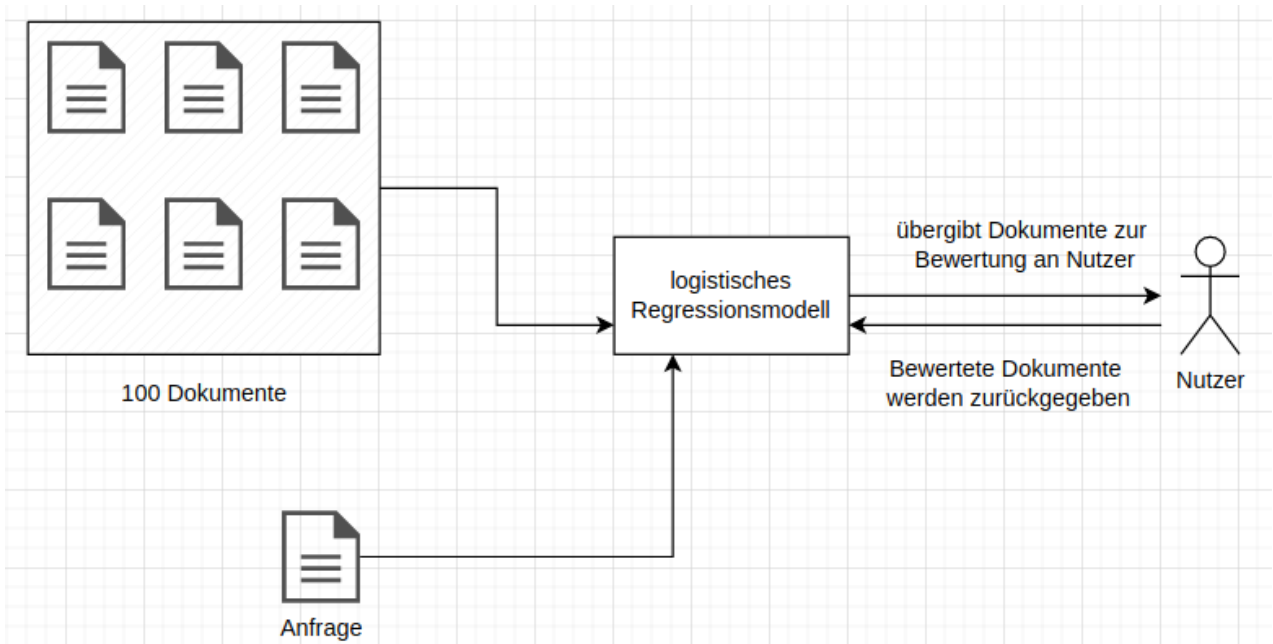
Total-Recall-Suchen mit Relevanzfeedback

- Total-Recall-Suchen können durch Machine Learning unterstützt werden
- Nutzer bewerten Dokumente anhand ihrer Relevanz zur Suchanfrage
- Machine Learning - Modell wird auf dem Relevanzfeedback der Nutzer trainiert



HiCal

- HiCal-Algorithmus der Universität Waterloo
- benutzt ein logistisches Regressionsmodell, welches auf dem Relevanzfeedback des Nutzers trainiert wird
- gilt als State-Of-The-Art Algorithmus für das Screening von Dokumenten bei systematischen Reviews



Ablauf einer Total-Recall-Suche mit HiCal

- Nutzer überprüft die Relevanz der Dokumente
- Modell wird auf den vom Nutzer gekennzeichneten Dokumenten trainiert (TF-IDF-Vektoren der gekennzeichneten Dokumente werden zur erneuten Gewichtung des Modells genutzt)
- Modell klassifiziert die ungekennzeichneten Dokumente
- Dokumente werden anhand ihrer Wahrscheinlichkeit sortiert und dem Nutzer präsentiert
- Loop beginnt von vorn und Nutzer überprüft erneut die Relevanz

Beispiel von TF-IDF und kontextabhängigem Termgewicht

Abstract : The utilisation of the CRISPR/Cas9 technology has sparked a renewed interest in gene drive mechanisms. These mechanisms of biased inheritance may yield promising applications in the fields of vector control and nature conservation. However, the same properties that will enable these applications may also pose a risk if organisms that are equipped with gene drive cassettes are unintentionally released into the environment. Although several groups of scientists and regulators have started to address these safety concerns, there are currently no dedicated guidelines published on the required risk assessment and minimal control measures applicable to gene drive organisms in contained use. To fill this gap, this paper describes a fundamental approach to assessing the risks of these organisms while handled in a contained laboratory environment. Based on the likelihood that an adverse effect will arise from the handling of a gene drive organism and the severity of this effect, three risk classes for contained use activities are presented. Finally, specific minimum requirements regarding physical measures and working practices are proposed according to the presented risk classes and tailored to activities with rodents, insects, and fungi, which are most likely to be used for gene drive applications in the near future.

Kontext: ?

- Terme werden anhand der Häufigkeit im Dokument betrachtet
- Kontext einer Passage kann dabei helfen die zentralen Terme im Dokument stärker zu gewichten

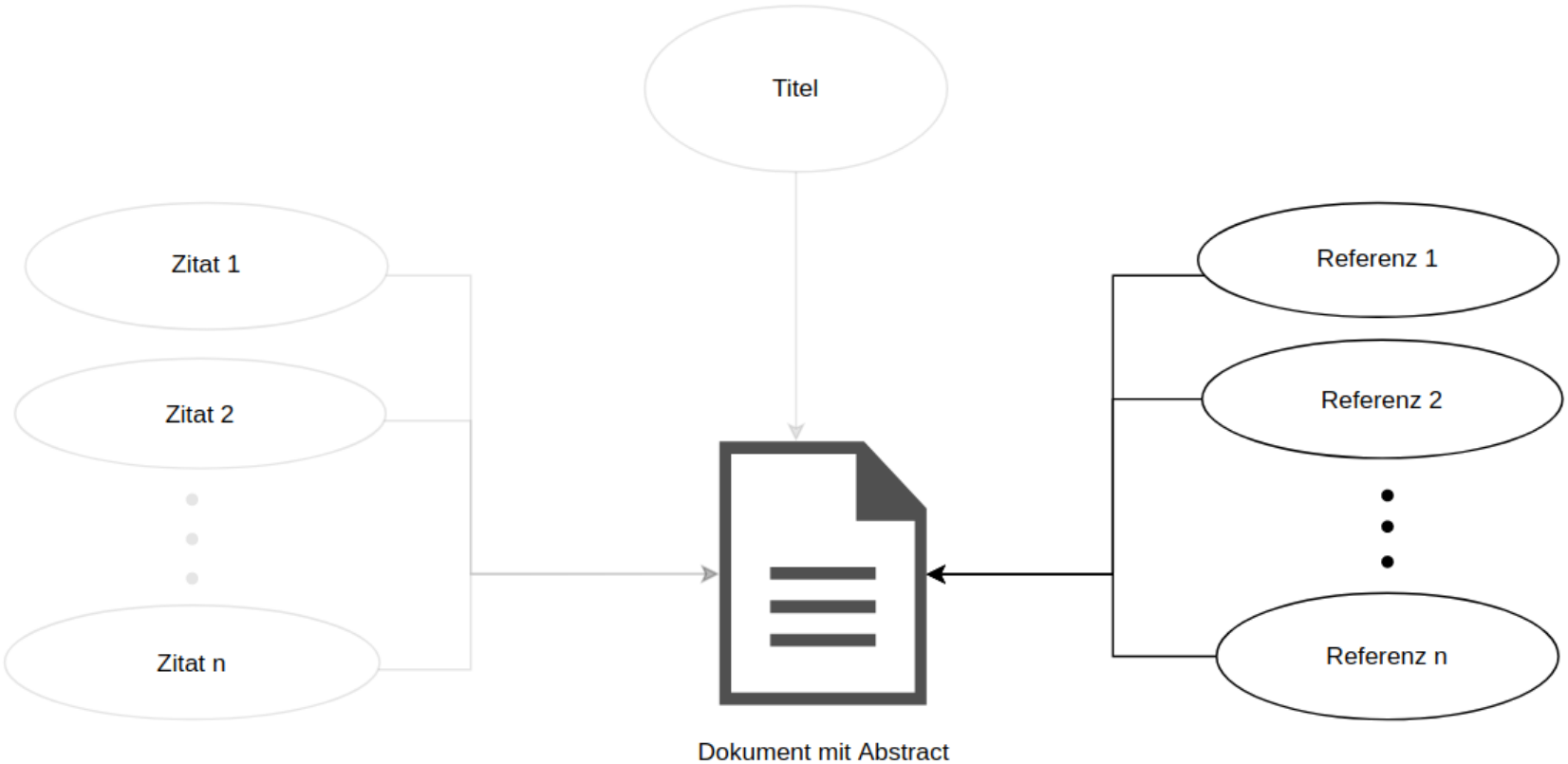
Vergleich von Termgewichten mit Häufigkeit und Kontext

- Termgewicht wird anhand der Bedeutung im Kontext der Passage berechnet
- Vorteile:
  - Erfassung von komplexen Satzstrukturen
  - Terme werden anhand ihrer Relevanz zum Kontext gewichtet
- Nachteile:
  - erneute Gewichtung des Index wird benötigt

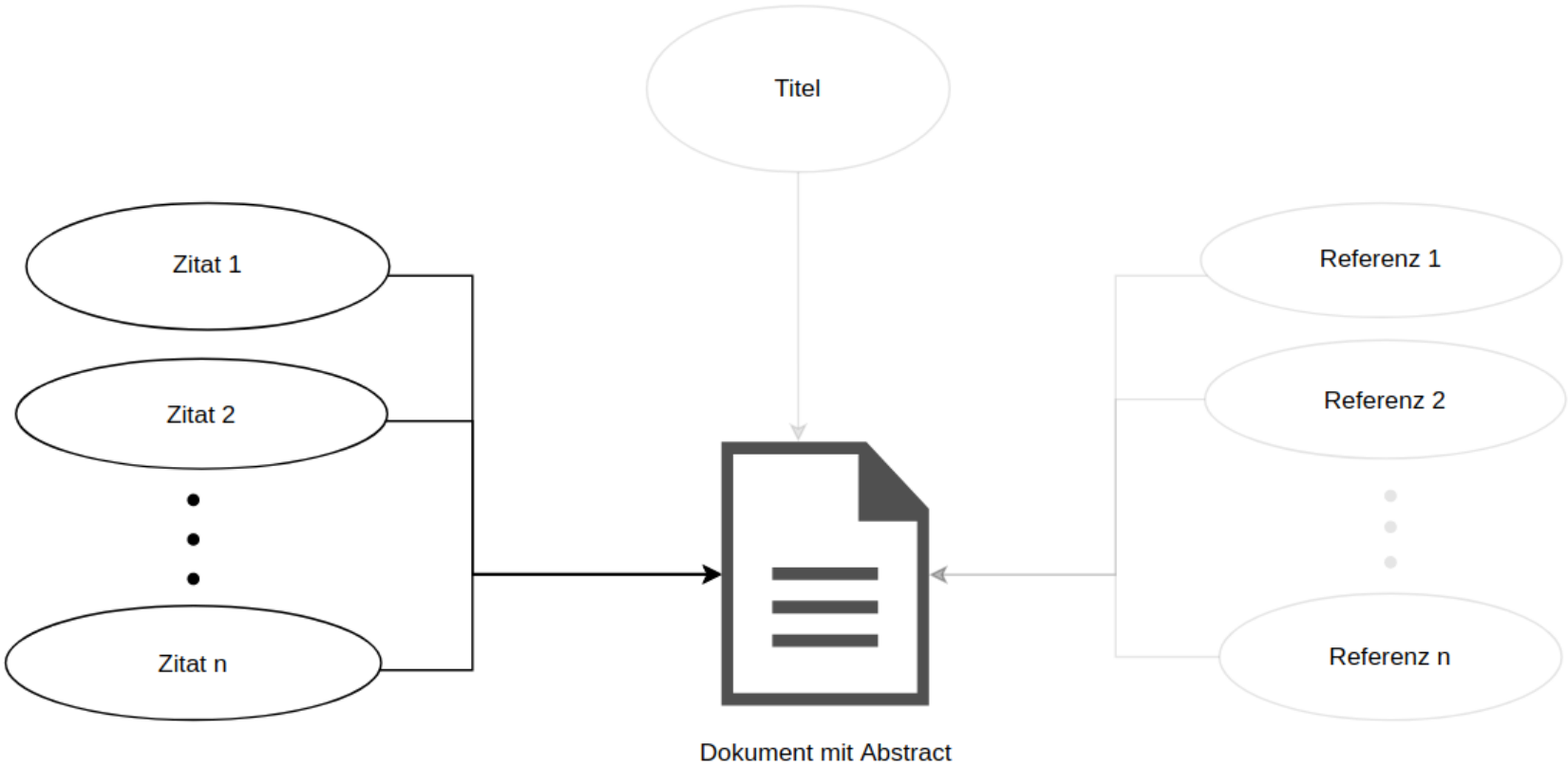
- Termgewicht wird anhand der Häufigkeit im Dokument und im Vergleich zu anderen Dokumenten im Corpus berechnet
- Vorteile:
  - hat sich als stabiles Maß für Retrieval erwiesen
- Nachteile:
  - keine Erfassung von komplexen Sprachmustern oder Grammatik

Kontext für wissenschaftliche Artikel - Titel

Kontext für wissenschaftliche Artikel - Referenzen

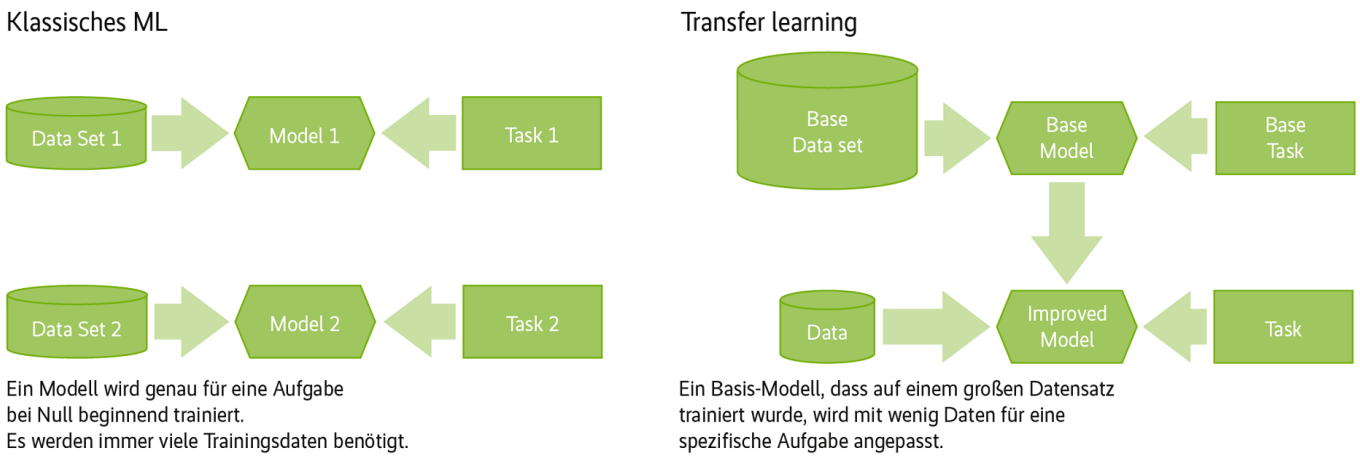


Kontext für wissenschaftliche Artikel - Zitate



BERT - neuronales Sprachmodell

- BERT ist ein neuronales Sprachmodell, welches bei NLP-Problemen eingesetzt wird
- Erzeugen von kontextualisierten Worteinbettungen
- Einsatz von BERT mithilfe von FineTuning auf Klassifikationsprobleme



Finetuning von BERT mit DeepCT

- Umwandlung von den kontextualisierten Worteinbettungen in kontextabhängige Termgewichte
- Training von BERT mit DeepCT zur Vorhersage von zentralen kontextabhängigen Termen im Dokument
- Erhöhung der Termhäufigkeit von wichtigen kontextabhängigen Termen
- Verringerung der Termhäufigkeit von unwichtigen Termen im Kontext

Titel – Strategie für kontextabhängige Termgewichtung

- benötigte Dokumentfelder: Titel, Abstract
- Hyphothese: Titel beschreibt das größere Feld Abstract
- Labeling der Terme, die im Abstract, sowie im Titel vorkommen
- Beispieldokument:

Titel :
A Framework for the Risk Assessment and Management of Gene Drive Technology in Contained Use

Abstract : The utilisation of the CRISPR/Cas9 technology has sparked a renewed interest in gene drive mechanisms. These mechanisms of biased inheritance may yield promising applications in the fields of vector control and nature conservation. However, the same properties that will enable these applications may also pose a risk if organisms that are equipped with gene drive cassettes are unintentionally released into the environment. Although several groups of scientists and regulators have started to address these safety concerns, there are currently no dedicated guidelines published on the required risk assessment and minimal control measures applicable to gene drive organisms in contained use. To fill this gap, this paper describes a fundamental approach to assessing the risks of these organisms while handled in a contained laboratory environment. Based on the likelihood that an adverse effect will arise from the handling of a gene drive organism and the severity of this effect, three risk classes for contained use activities are presented. Finally, specific minimum requirements regarding physical measures and working practices are proposed according to the presented risk classes and tailored to activities with rodents, insects, and fungi, which are most likely to be used for gene drive applications in the near future.

Zitate – Strategie für kontextabhängige Termgewichtung

- benötigte Dokumentfelder: Titel der Zitate, Abstract
- Hyphothese: Titel der Zitate beschreiben das größere Feld Abstract
- Labeling der Terme, die im Abstract, sowie in mindestens einem Titel der Zitate vorkommen

Zitate:

- Adequacy and sufficiency evaluation of existing EFSA guidelines for the molecular characterisation, environmental risk assessment and post\u2010market environmental monitoring of genetically modified insects containing engineered gene drives
- A Cross-Sectional Survey of Biosafety Professionals Regarding Genetically Modified Insects Building biosecurity for synthetic biology
- The value of existing regulatory frameworks for the environmental risk assessment of agricultural pest control using gene drives
- Regulation of GM Organisms for Invasive Species Control
- Population management using gene drive: molecular design, models of\u00a0spread dynamics and assessment of ecological risks
- Assessment of human health and environmental risks of new developments in modern biotechnology : Policy report

Abstract: The utilisation of the CRISPR/Cas9 technology has sparked a renewed interest in gene drive mechanisms. These mechanisms of biased inheritance may yield promising applications in the fields of vector control and nature conservation. However, the same properties that will enable these applications may also pose a risk if organisms that are equipped with gene drive cassettes are unintentionally released into the environment. Although several groups of scientists and regulators have started to address these safety concerns, there are currently no dedicated guidelines published on the required risk assessment and minimal control measures applicable to gene drive organisms in contained use. To fill this gap, this paper describes a fundamental approach to assessing the risks of these organisms while handled in a contained laboratory environment. Based on the likelihood that an adverse effect will arise from the handling of a gene drive organism and the severity of this effect, three risk classes for contained use activities are presented. Finally, specific minimum requirements regarding physical measures and working practices are proposed according to the presented risk classes and tailored to activities with rodents, insects, and fungi, which are most likely to be used for gene drive applications in the near future.

Referenzen – Strategie zum Lernen für kontextabhängige Termgewichtung

- benötigte Dokumentfelder: Titel der Referenzen, Abstract
- Hyphothese: Titel der Referenzen beschreiben das größere Feld Abstract
- Labeling der Terme, die im Abstract, sowie in mindestens einem Titel der Referenzen vorkommen

Titel der Referenzen

- Containment of arthropod disease vectors
- Regulating gene drives
- Contained use of genetically modified micro-organisms
- Concerning RNA-Guided Gene Drives for the Alteration of Wild Populations
- Safeguarding gene drive experiments in the laboratory",
- Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi",
- Opinion: Is CRISPR-based gene drive a biocontrol silver bullet or global conservation threat?"
- Safeguarding CRISPR-Cas9 gene drives in yeast
- The mutagenic chain reaction: A method for converting heterozygous to homozygous mutations
- Can CRISPR-Cas9 gene drives curb malaria?
- Gene drives : Policy report
- A CRISPR-Cas9 Gene Drive System Targeting Female Reproduction in the Malaria Mosquito vector Anopheles gambiae
- Results from the Workshop \u201cProblem Formulation for the Use of Gene Drive in Mosquitoes\u201d
- Daisyfield gene drive systems harness repeated genomic elements as a generational clock to limit spread
- Adaptive Risk Management of Gene Drive Experiments
- Is It Time for Synthetic Biodiversity Conservation?
- Daisy-chain gene drives for the alteration of local populations

Abstract: The utilisation of the CRISPR/Cas9 technology has sparked a renewed interest in gene drive mechanisms. These mechanisms of biased inheritance may yield promising applications in the fields of vector control and nature conservation. However, the same properties that will enable these applications may also pose a risk if organisms that are equipped with gene drive cassettes are unintentionally released into the environment. Although several groups of scientists and regulators have started to address these safety concerns, there are currently no dedicated guidelines published on the required risk assessment and minimal control measures applicable to gene drive organisms in contained use. To fill this gap, this paper describes a fundamental approach to assessing the risks of these organisms while handled in a contained laboratory environment. Based on the likelihood that an adverse effect will arise from the handling of a gene drive organism and the severity of this effect, three risk classes for contained use activities are presented. Finally, specific minimum requirements regarding physical measures and working practices are proposed according to the presented risk classes and tailored to activities with rodents, insects, and fungi, which are most likely to be used for gene drive applications in the near future.

Experimentaufbau - Datensätze

Testdatensatz	Anteil relevant	Informationsbedürfnis	Anwendung
Ceeder'18	5%	Umweltmanagement und Risiken	Screening für systematisches Review
Ceeder'19	5%	Umweltmanagement und Risiken	Screening für systematisches Review
Genom Editierung	26%	Genom Editierung bei Zier und Nutzpflanzen	Screening für systematisches Review
Genom Editierung - Volltext	34%	Genom Editierung bei Zier und Nutzpflanzen	Screening für systematisches Review

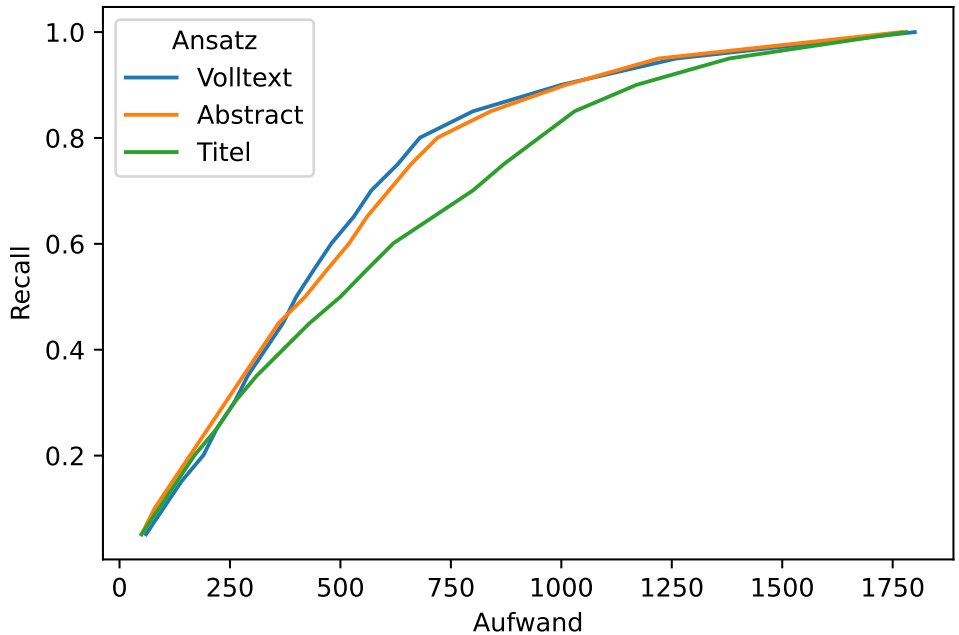
Trainingsdatensatz	Anzahl Dokumente	Modifikation	Dokumentfelder
Ceeder	5000	Kombination von Ceeder '18 und Ceeder '19	Dokumenttitel Abstract Titel der Zitate Titel der Referenzen
Genom Editierung	120000	Erweiterung des Genom Editierung Datensatzes mit Zitaten und Referenzen	Dokumenttitel Abstract Titel der Zitate Titel der Referenzen

Experimentaufbau - Datensätze

- Metrik zur Evaluation: Work saved over sampling
  - Maß zur Bestimmung der eingesparten Arbeit
  - $WSS@Recall = \frac{TrueNegatives + FalseNegatives}{Gesamte\ Anzahl\ an\ Dokumenten} - (1 - Recall)$
  - True Negatives + False Negatives = Dokumente, die noch betrachtet werden

- Plots zur Evaluation: Recall gegen Aufwand
  - Recall : Anteil an gekennzeichneten relevanten Dokumenten
  - Aufwand : Anzahl an betrachteten Dokumenten

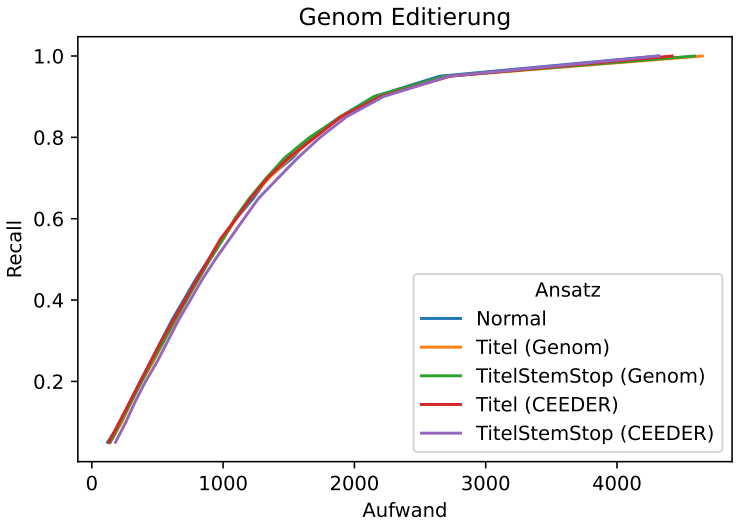
Evaluation II - Vergleich von Dokumentfeldern

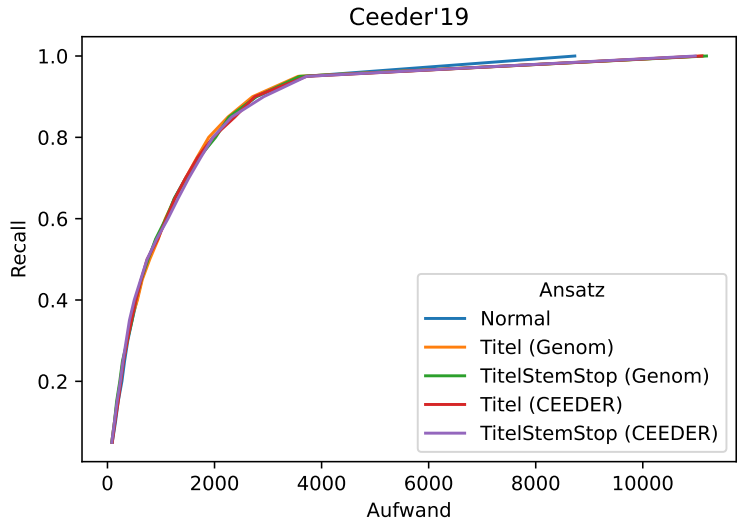


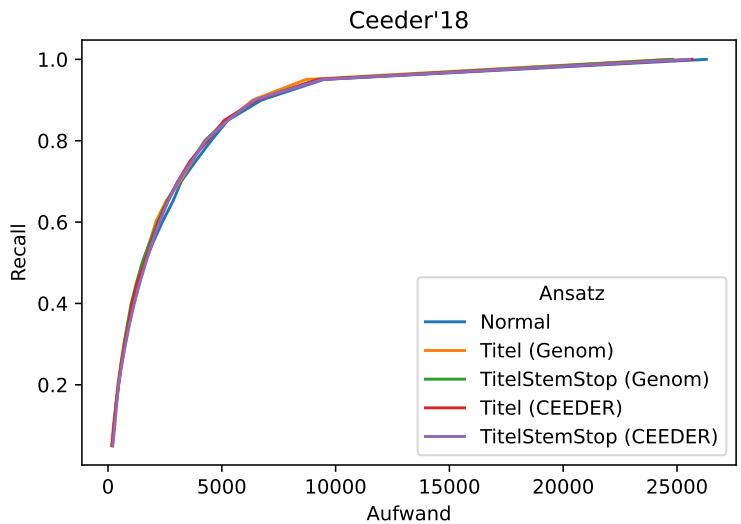
Repräsentation	Datensatz	WSS@85	WSS@90	WSS@95
Abstract	Genom-Volltext	0.389	0.439	0.489
Volltext	Genom-Volltext	**0.408**	**0.458**	**0.508**
Titel	Genom-Volltext	0.281	0.331	0.381

Tabelle: WSS von einzelnen Dokumentfeldern (Titel, Abstract, Volltext) beim Screening für systematische Reviews mit HiCal

Evaluation III - Titel-Strategie Plots





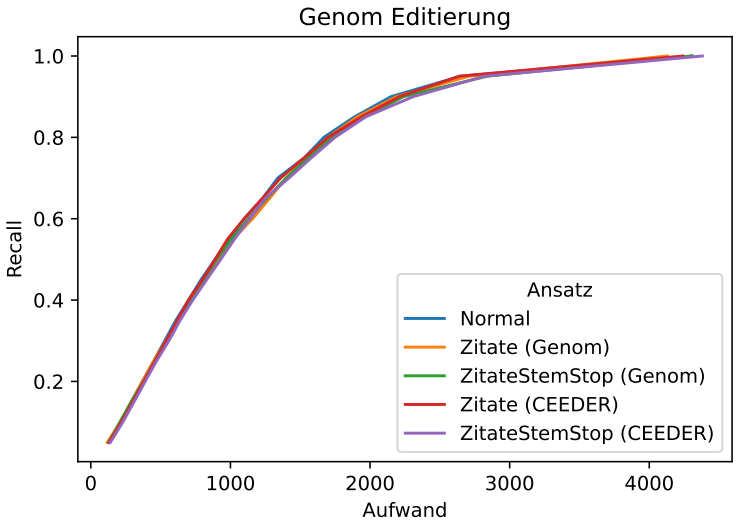


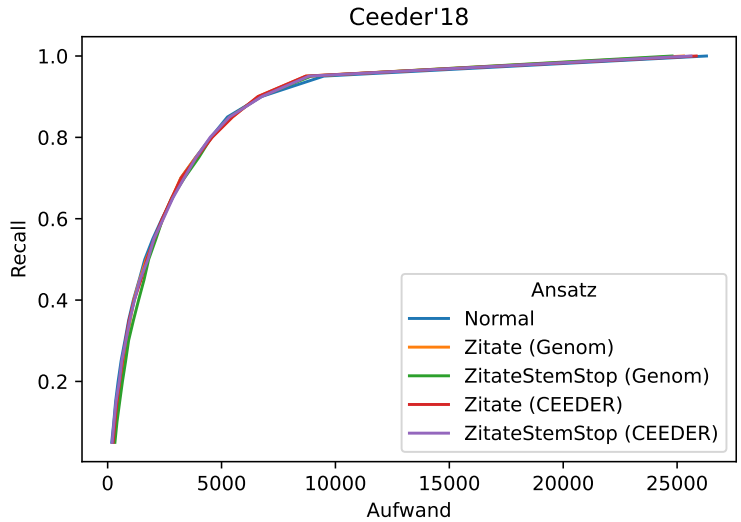
Evaluation III - Titel-Strategie WSS

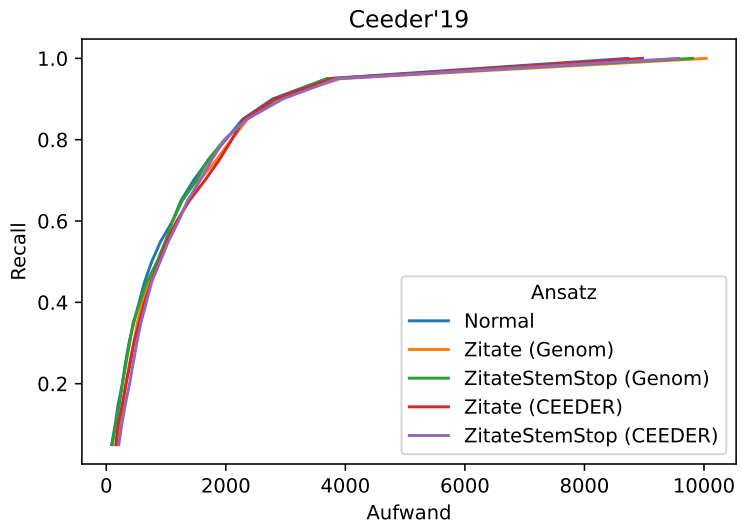
Methode	Trainingsdaten	Datensatz	WSS@85	WSS@90	WSS@95
Normal	-	Ceeder'18	0.6536	0.7036	0.7536
Titel	Genom	Ceeder'18	0.6544	0.7044	0.7544
+ Stem/Stop	Genom	Ceeder'18	0.6562	0.7062	0.7562
Titel	Ceeder	Ceeder'18	**0.6592**	**0.7092**	**0.7592**
+ Stem/Stop	Ceeder	Ceeder'18	0.6536	0.7036	0.7536
Normal	-	Ceeder'19	0.6491	0.6991	0.7491
Titel	Genom	Ceeder'19	**0.6518**	**0.7018**	**0.7518**
+ Stem/Stop	Genom	Ceeder'19	0.6501	0.7001	0.7501
Titel	Ceeder	Ceeder'19	0.6405	0.6905	0.7405
+ Stem/Stop	Ceeder	Ceeder'19	0.6457	0.6957	0.7457
Normal	-	Genom	0.4599	0.5099	0.5599
Titel	Genom	Genom	**0.4605**	**0.5105**	**0.5605**
+ Stem/Stop	Genom	Genom	0.4586	0.5086	0.5586
Titel	Ceeder	Genom	0.4580	0.5080	0.5580
+ Stem/Stop	Ceeder	Genom	0.4485	0.4985	0.5485

Tabelle: Übersicht über den WSS beim Screening-Prozess von systematischen Reviews mit der Titel-Strategie mit Stemming/Stopping. Als Testdatensätze wurden der Ceeder'18, der Ceeder'19 und der Genom-Editierung-Datensatz verwendet.

Evaluation IV – Zitate-Strategie Plots





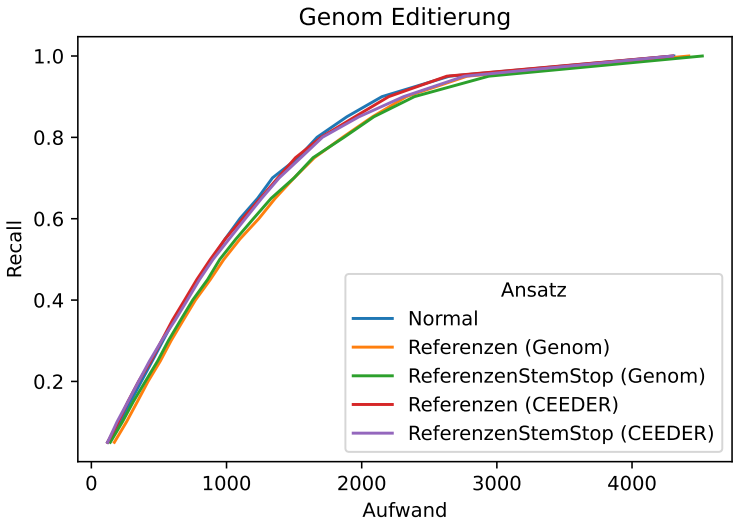


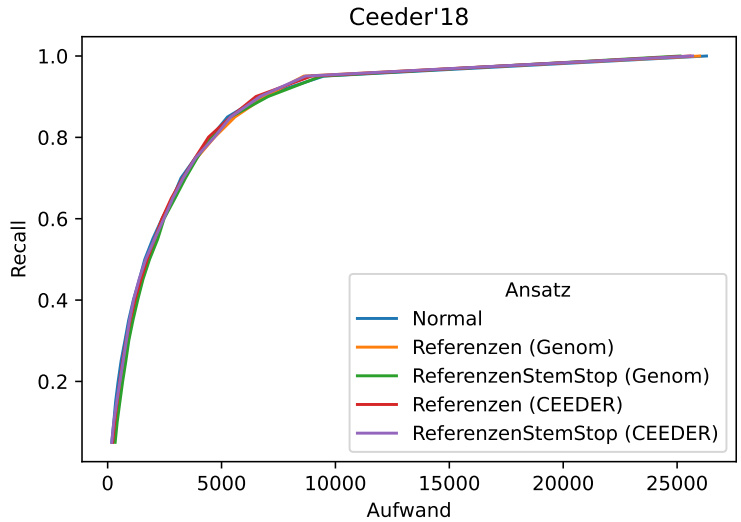
Evaluation IV – Zitate-Strategie WSS

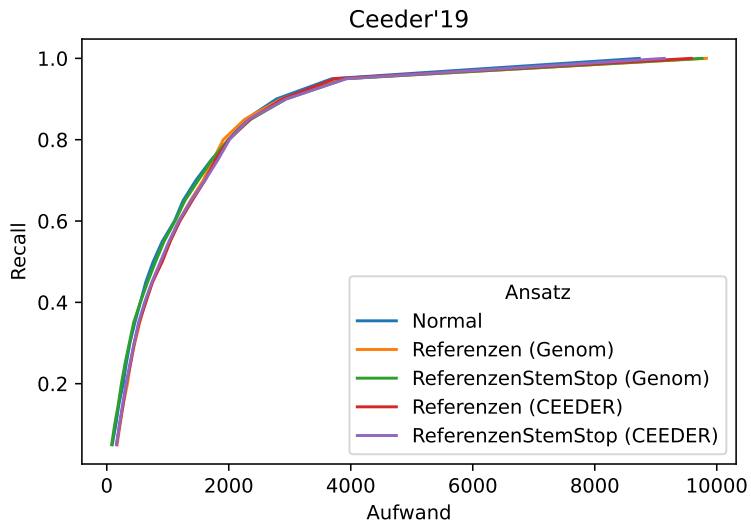
Methode	Trainingsdaten	Datensatz	WSS@85	WSS@90	WSS@95
Normal	-	Ceeder'18	**0.6536**	**0.7036**	**0.7536**
Zitate	Genom	Ceeder'18	0.6453	0.6953	0.7453
+ Stem/Stop	Genom	Ceeder'18	0.6502	0.7002	0.7502
Zitate	Ceeder	Ceeder'18	0.6453	0.6953	0.7453
+ Stem/Stop	Ceeder	Ceeder'18	0.6501	0.7001	0.7501
Normal	-	Ceeder'19	**0.6491**	**0.6991**	**0.7491**
Zitate	Genom	Ceeder'19	0.6421	0.6921	0.7421
+ Stem/Stop	Genom	Ceeder'19	0.6453	0.6953	0.7453
Zitate	Ceeder	Ceeder'19	0.6472	0.6972	0.7472
+ Stem/Stop	Ceeder	Ceeder'19	0.6430	0.6930	0.7430
Normal	-	Genom	**0.4599**	**0.5099**	**0.5599**
Zitate	Genom	Genom	0.4557	0.5057	0.5557
+ Stem/Stop	Genom	Genom	0.4452	0.4952	0.5452
Zitate	Ceeder	Genom	0.4502	0.5002	0.5502
+ Stem/Stop	Ceeder	Genom	0.4433	0.4933	0.5433

Tabelle: Übersicht über den WSS beim Screening-Prozess von systematischen Reviews, mit der Zitate-Strategie, mit Stemming/Stopping. Als Testdatensätze wurden der Ceeder'18, der Ceeder'19 und der Genom-Editierung-Datensatz verwendet.

Evaluation V – Referenzen-Strategie Plots







Evaluation V – Referenzen-Strategie WSS

Methode	Trainingsdaten	Datensatz	WSS@85	WSS@90	WSS@95
Normal	-	Ceeder'18	**0.6536**	**0.7036**	**0.7536**
Referenzen	Genom	Ceeder'18	0.6412	0.6912	0.7412
+ Stem/Stop	Genom	Ceeder'18	0.6485	0.6985	0.7485
Referenzen	Ceeder	Ceeder'18	0.6483	0.6983	0.7483
+ Stem/Stop	Ceeder	Ceeder'18	0.6484	0.6984	0.7484
Normal	-	Ceeder'19	0.6491	0.6991	0.7491
Referenzen	Genom	Ceeder'19	**0.6509**	**0.7009**	**0.7509**
+ Stem/Stop	Genom	Ceeder'19	0.6425	0.6925	0.7425
Referenzen	Ceeder	Ceeder'19	0.6429	0.6929	0.7429
+ Stem/Stop	Ceeder	Ceeder'19	0.6452	0.6952	0.7452
Normal	-	Genom	**0.4599**	**0.5099**	**0.5599**
Referenzen	Genom	Genom	0.4206	0.4706	0.5206
+ Stem/Stop	Genom	Genom	0.4175	0.4675	0.5175
Referenzen	Ceeder	Genom	0.4467	0.4967	0.5467
+ Stem/Stop	Ceeder	Genom	0.4407	0.4907	0.5407

Tabelle: Übersicht über den WSS beim Screening-Prozess von systematischen Reviews mit der Referenzen-Strategie mit Stemming/Stopping. Als Testdatensätze wurden der Ceeder'18, der Ceeder'19 und Genom-Editierung verwendet.

Evaluation VI - Volltext-Dokumente mit kontextabhängiger Termgewichtung Plots

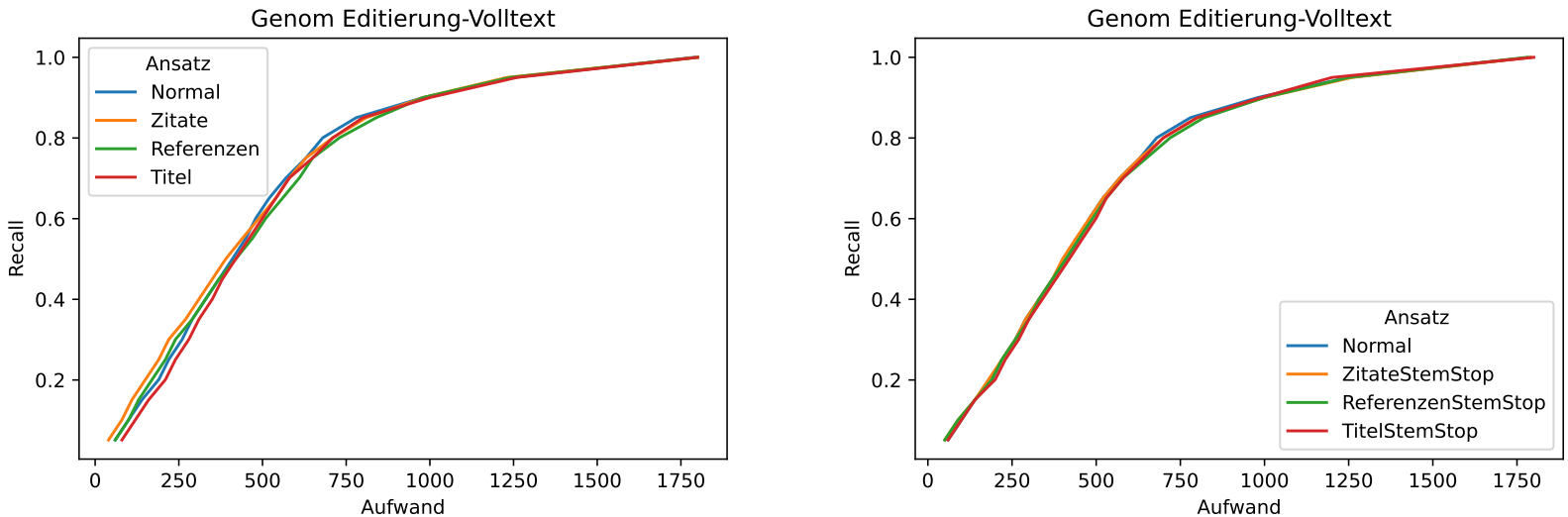


Abbildung: Aufwand gegen Recall bei Volltexten mit kontextabhängiger Termgewichtung

Evaluation VI - Volltext-Dokumente mit kontextabhängiger Termgewichtung WSS

Methode	Datensatz	WSS@85	WSS@90	WSS@95
Normal	Genom-Volltext	**0.4227**	**0.4727**	**0.5227**
Titel	Genom-Volltext	0.4106	0.4606	0.5106
+ Stem/Stop	Genom-Volltext	0.4122	0.4622	0.5122
Zitate	Genom-Volltext	0.4029	0.4529	0.5029
+ Stem/Stop	Genom-Volltext	0.4095	0.4595	0.5095
Referenzen	Genom-Volltext	0.3902	0.4402	0.4902
+ Stem/Stop	Genom-Volltext	0.3968	0.4468	0.4968

Tabelle: WSS beim Screening von systematischen Reviews mit Volltexten

Zusammenfassung

- Kontextabhängige Termgewichtungen können verwendet werden, um die Effektivität von Total-Recall-Suchen zu erhöhen
- Machine Learning kann eingesetzt werden um Total-Recall-Suchen durchzuführen
- Mithilfe der Titel-Strategie können kontextabhängige Termgewichtungen erzeugt werden, welche den Aufwand zum Erreichen von Total Recall verringert

Vielen Dank für Ihre Aufmerksamkeit!