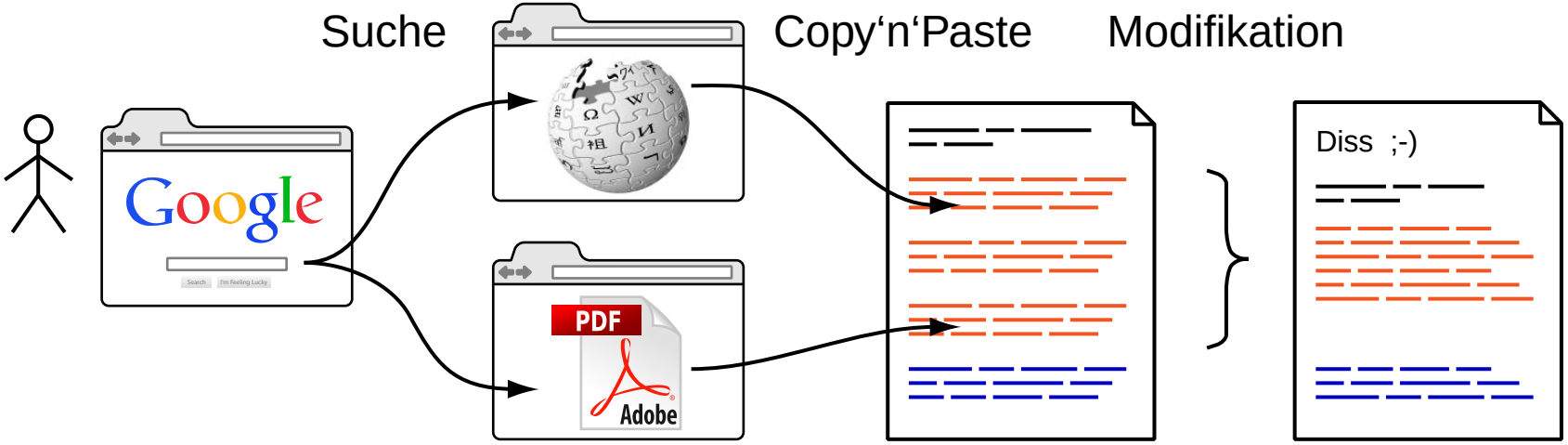
Plagiate, und wie (gut) man sie erkennen kann

Technologien zur Wiederverwendung von Texten aus dem Web

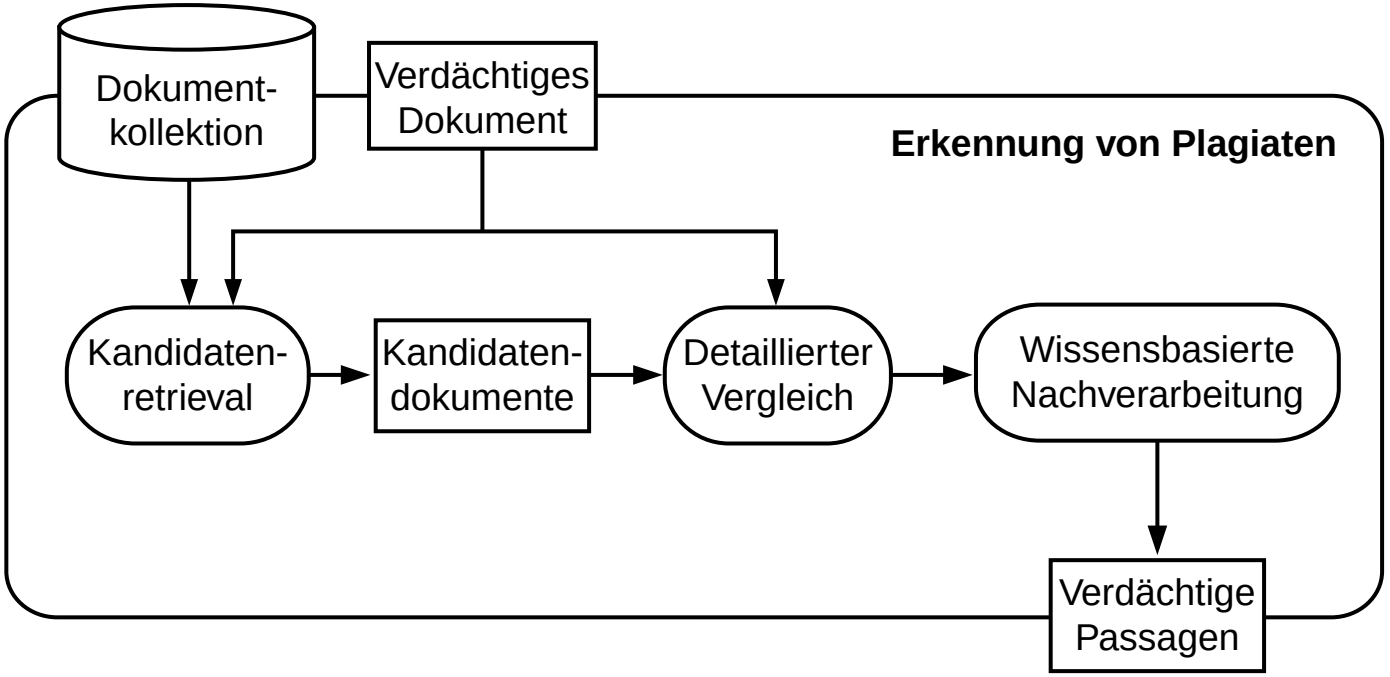
Martin Potthast

Bauhaus-Universität Weimar
www.webis.de

Wie Menschen plagiieren

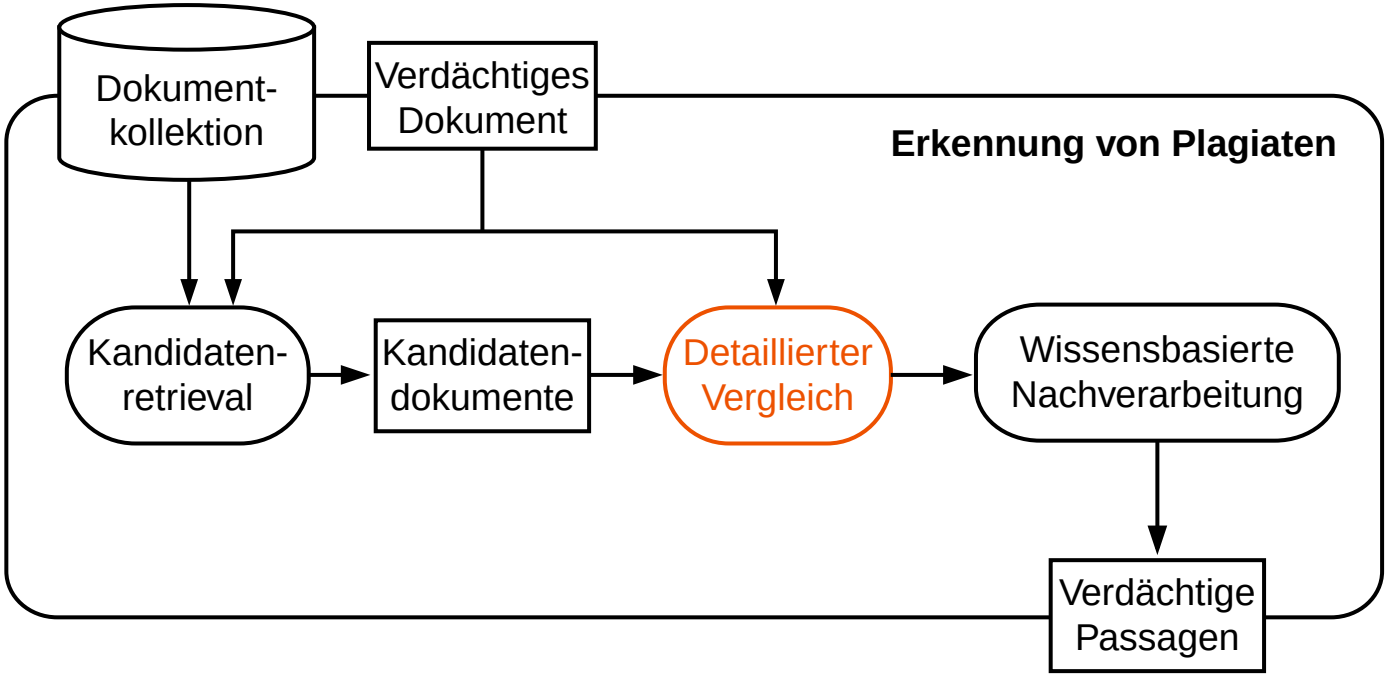


Wie Maschinen Plagiate erkennen



Wie Maschinen Plagiate erkennen

- Wie können sprachübergreifende Plagiate erkannt werden?
- Wie verlässlich sind Algorithmen zur Plagiatserkennung?



Messung sprachübergreifender Textähnlichkeit

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

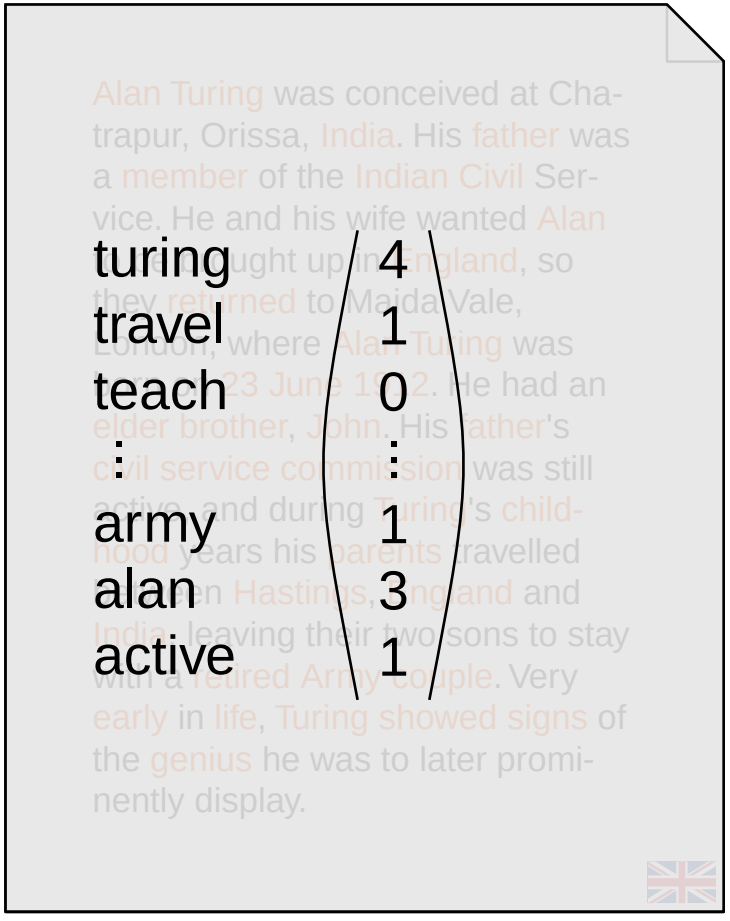
Alan Mathison Turing was born on 23 June 1912. His father was Julius Mathison Turing, member of the civil service in India, and his mother Ethel Sara Turing, the daughter of Edward Waller Stoney. Alan's childhood was spent with his elder brother John, living with a retired Army couple near Hastings, England. His parents returned to India until the end of his father‘s civil service commission, and visited when they could. Signs of Turing‘s genius showed early in his life. It is reported that he taught himself reading in less than three weeks.

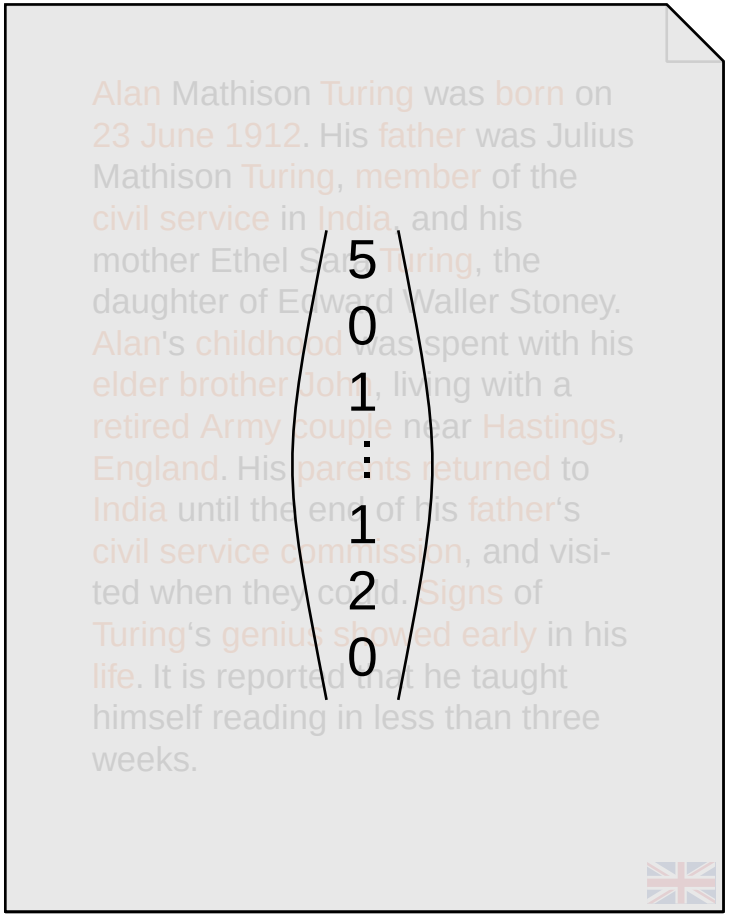
Messung sprachübergreifender Textähnlichkeit

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

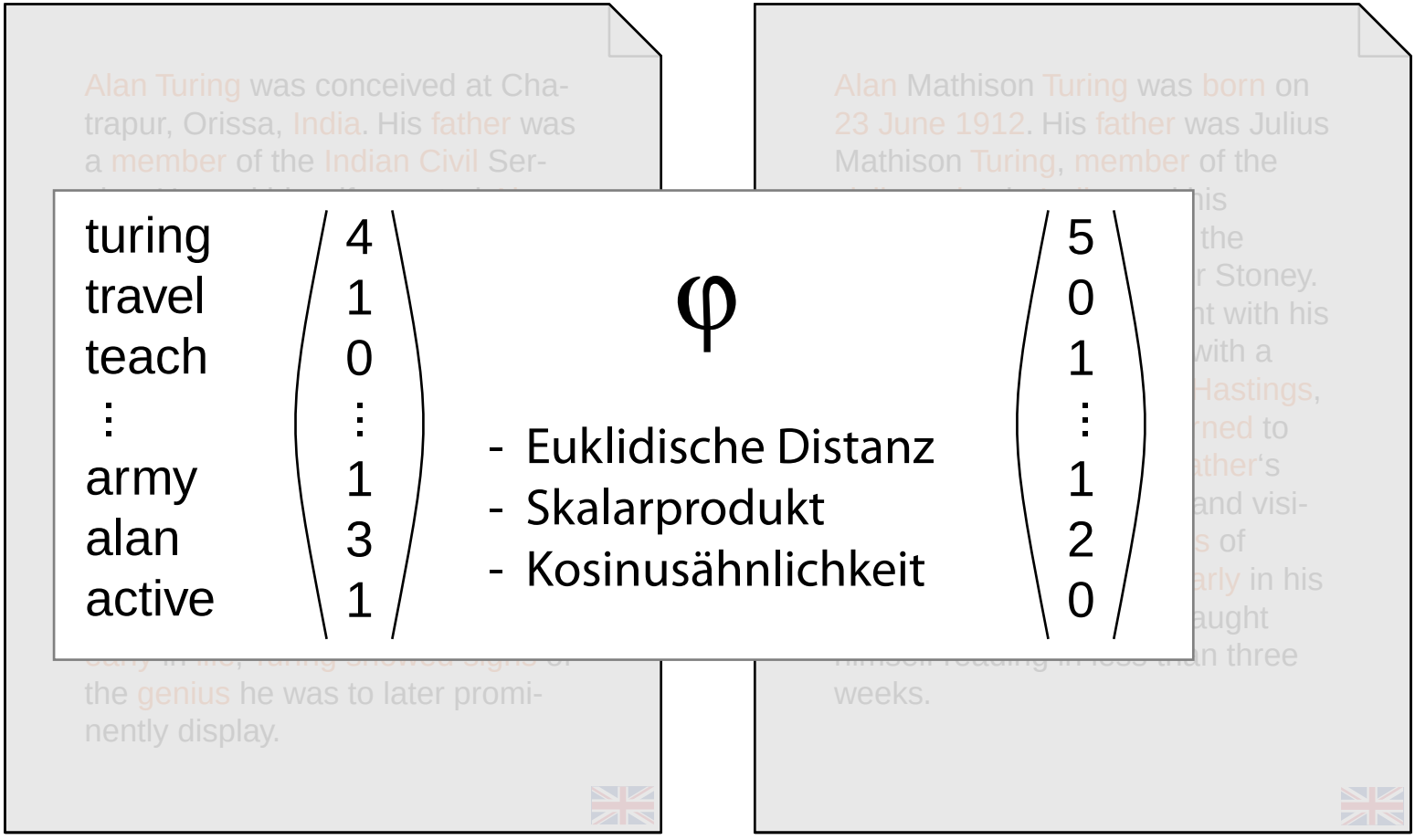
Alan Mathison Turing was born on 23 June 1912. His father was Julius Mathison Turing, member of the civil service in India, and his mother Ethel Sara Turing, the daughter of Edward Waller Stoney. Alan's childhood was spent with his elder brother John, living with a retired Army couple near Hastings, England. His parents returned to India until the end of his father‘s civil service commission, and visited when they could. Signs of Turing‘s genius showed early in his life. It is reported that he taught himself reading in less than three weeks.

Messung sprachübergreifender Textähnlichkeit





Messung sprachübergreifender Textähnlichkeit



Messung sprachübergreifender Textähnlichkeit

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

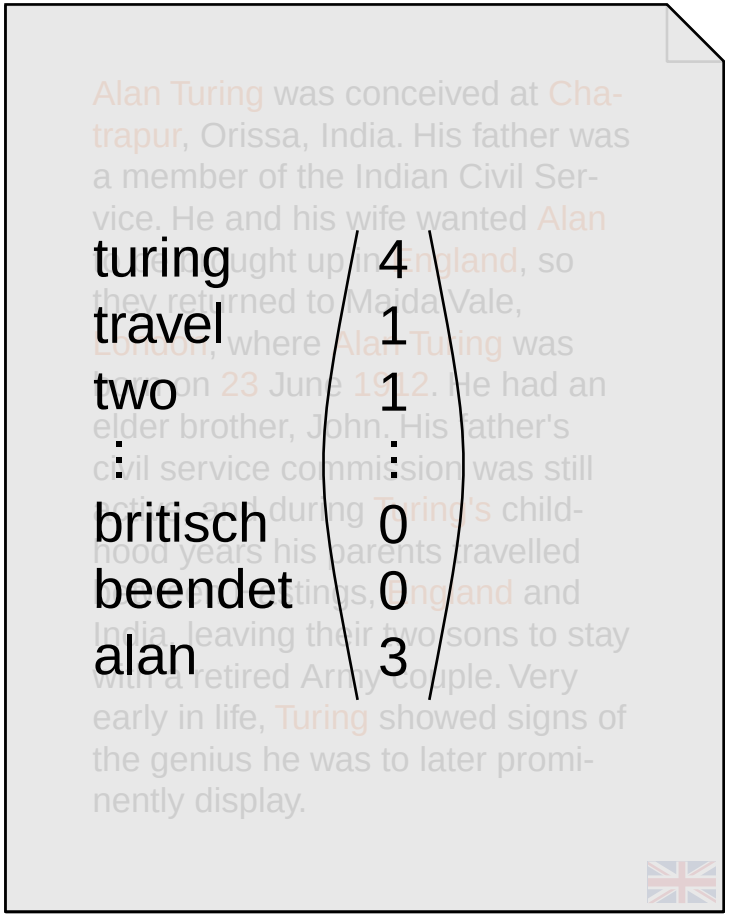
Turings Vater Julius Mathison Turing, ein britischer Staatsdiener in Chatrapur, Indien, und dessen Frau Ethel Sara wollten, dass ihr Kind in Großbritannien geboren wird. Deshalb kehrten sie nach London-Paddington zurück, wo Alan Turing am 23. Juni 1912 zur Welt kam. Da der Staatsdienst seines Vaters noch nicht beendet war, pendelte dieser während Turings Kindheit zwischen England und Indien. Seine Familie ließ er aus Furcht vor Gefahren in der britischen Kolonie bei Freunden in England zurück. Schon in frühester Kindheit zeigte sich die hohe Begabung und Intelligenz Turings.

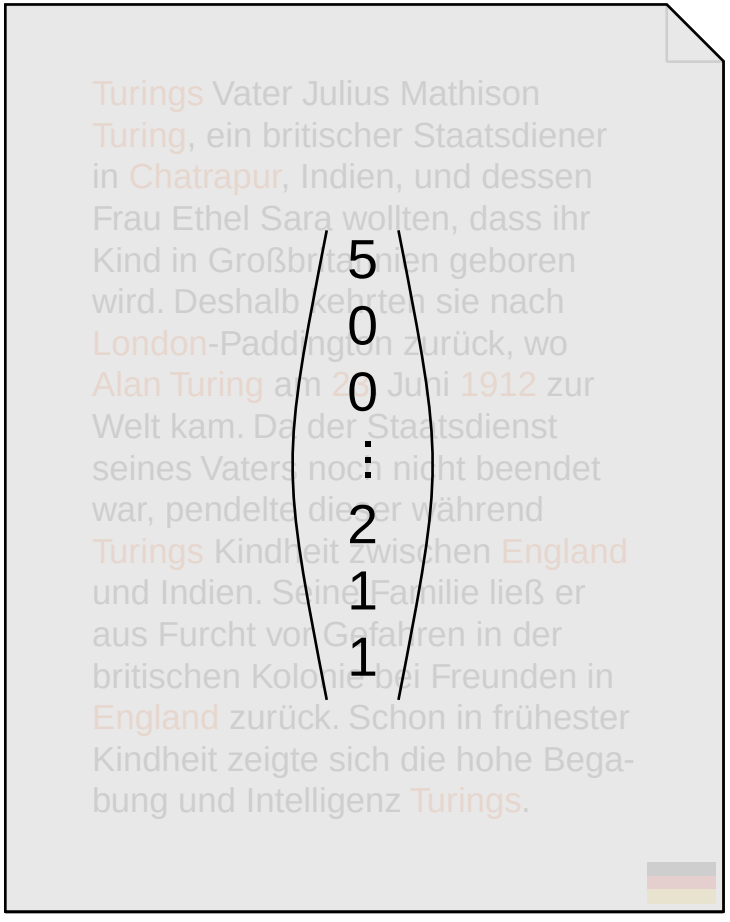
Messung sprachübergreifender Textähnlichkeit

Alan Turing was conceived at Chatrapur, Orissa, India. His father was a member of the Indian Civil Service. He and his wife wanted Alan to be brought up in England, so they returned to Maida Vale, London, where Alan Turing was born on 23 June 1912. He had an elder brother, John. His father's civil service commission was still active, and during Turing's childhood years his parents travelled between Hastings, England and India, leaving their two sons to stay with a retired Army couple. Very early in life, Turing showed signs of the genius he was to later prominently display.

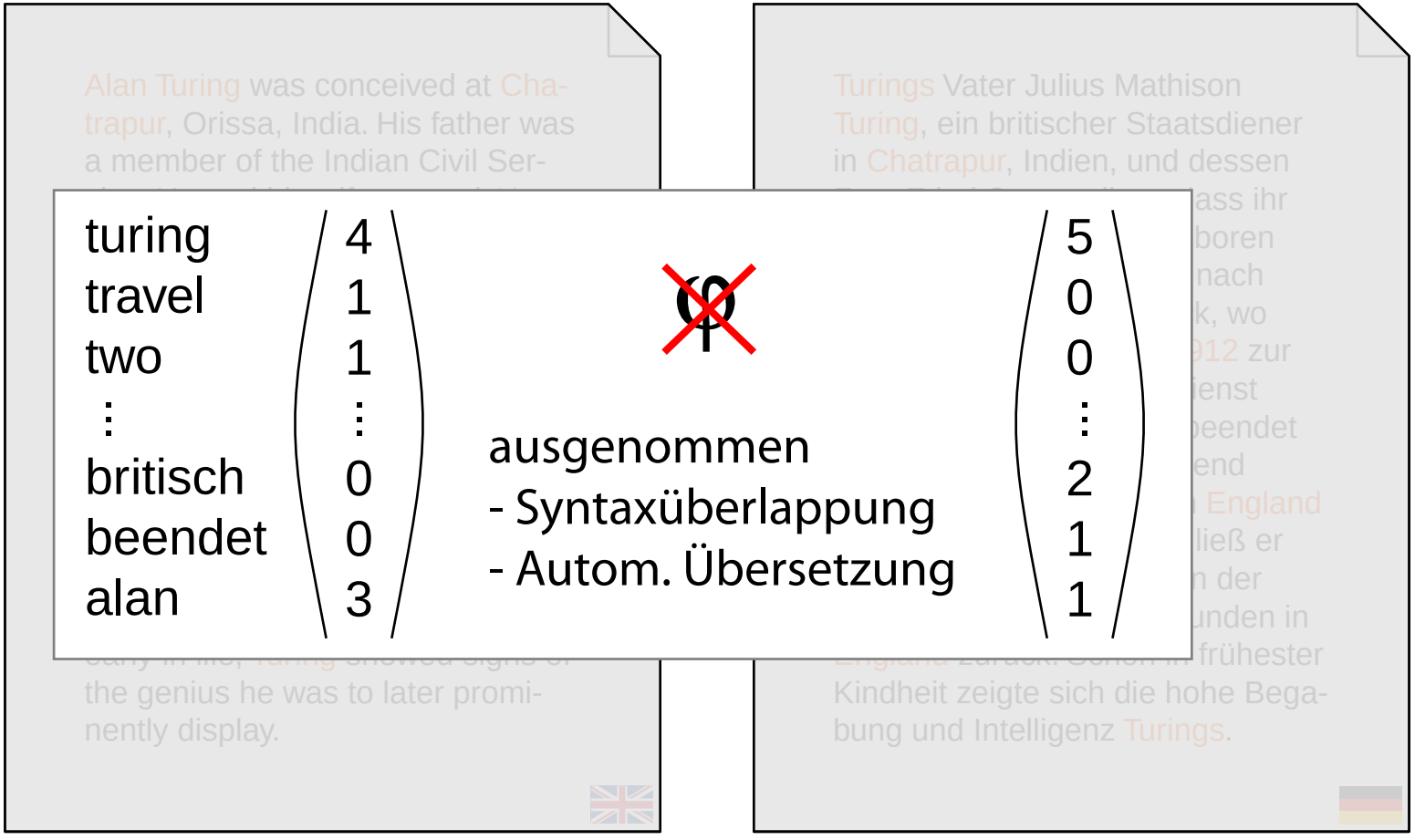
Turings Vater Julius Mathison Turing, ein britischer Staatsdiener in Chatrapur, Indien, und dessen Frau Ethel Sara wollten, dass ihr Kind in Großbritannien geboren wird. Deshalb kehrten sie nach London-Paddington zurück, wo Alan Turing am 23. Juni 1912 zur Welt kam. Da der Staatsdienst seines Vaters noch nicht beendet war, pendelte dieser während Turings Kindheit zwischen England und Indien. Seine Familie ließ er aus Furcht vor Gefahren in der britischen Kolonie bei Freunden in England zurück. Schon in frühester Kindheit zeigte sich die hohe Begabung und Intelligenz Turings.

Messung sprachübergreifender Textähnlichkeit



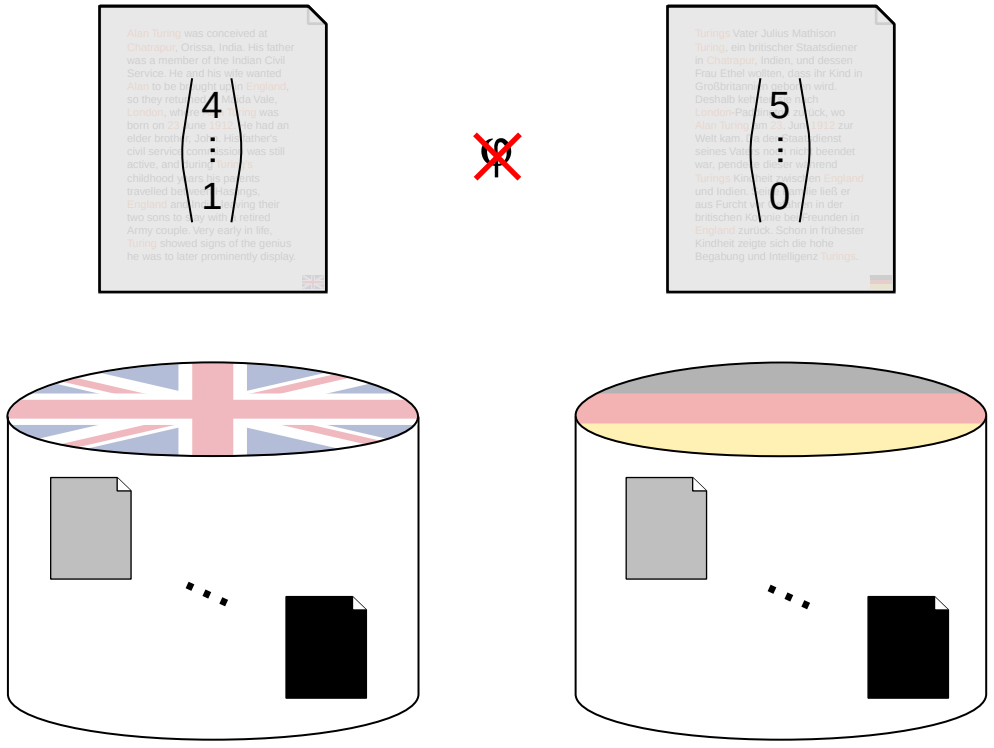


Messung sprachübergreifender Textähnlichkeit

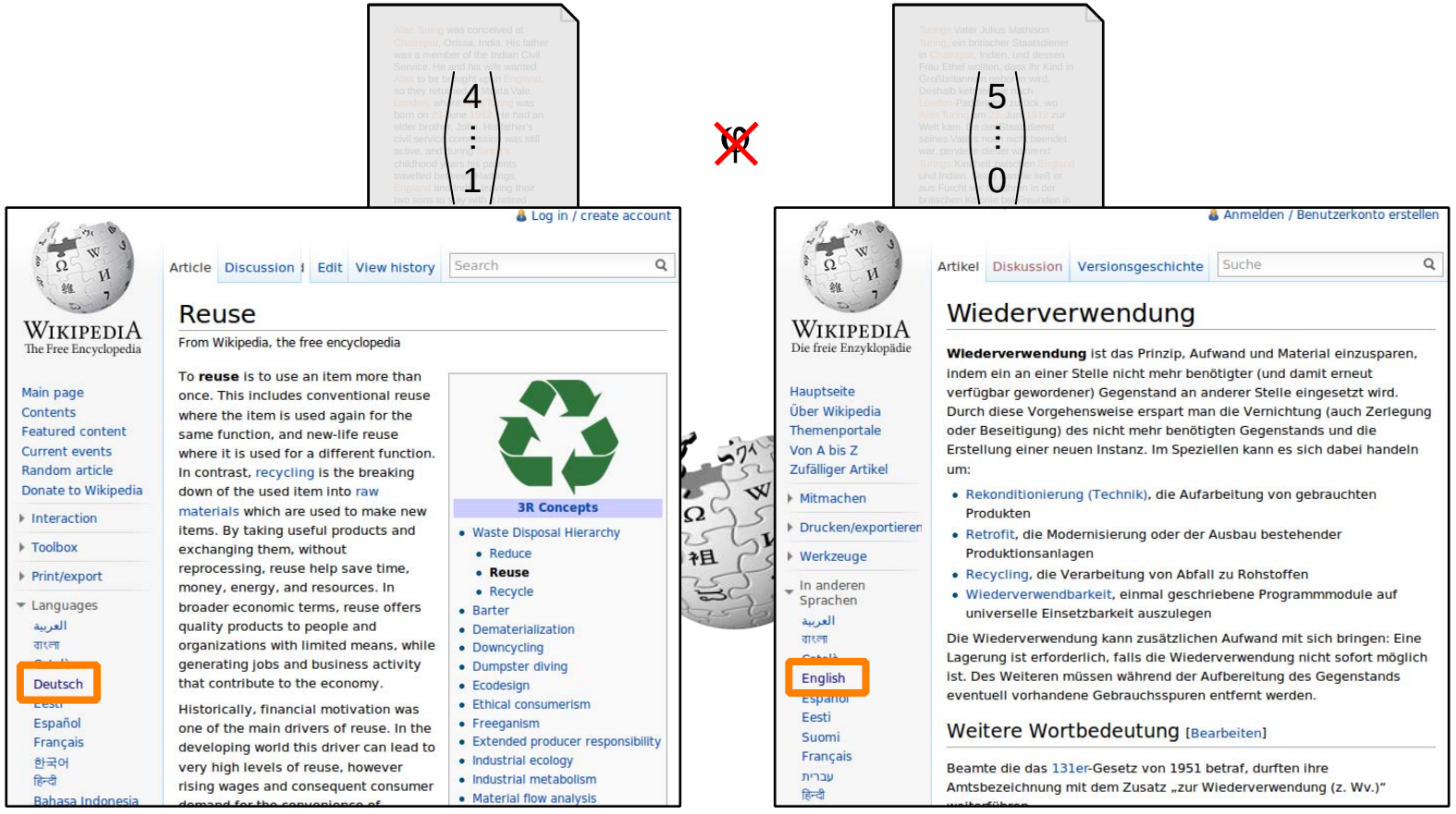


Cross-language Explicit Semantic Analysis (CL-ESA)

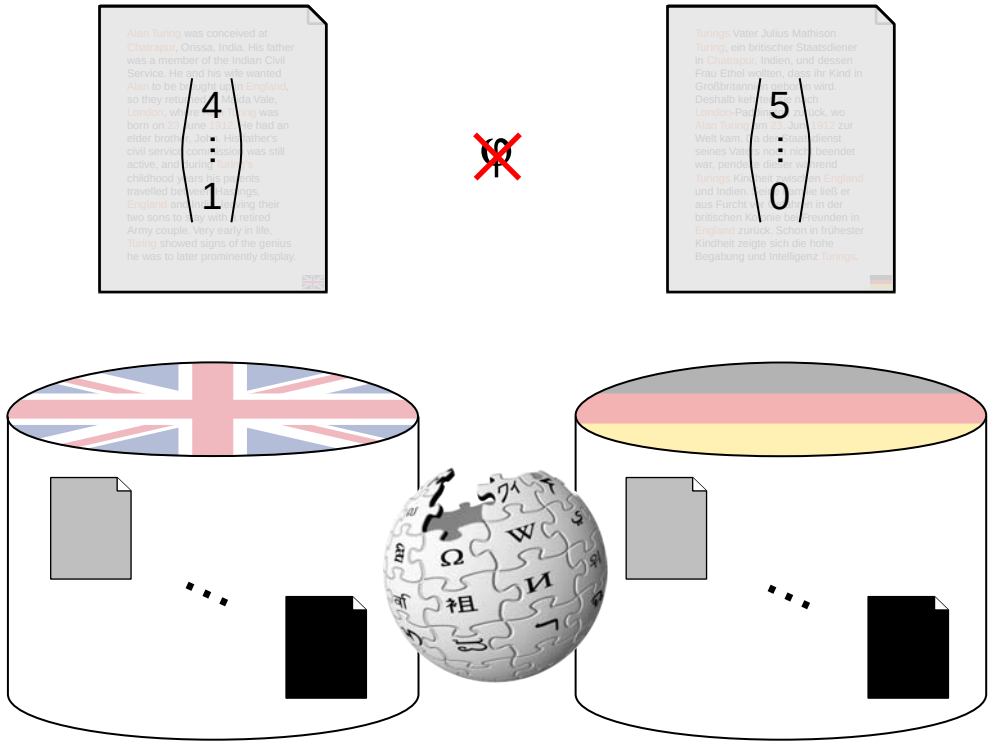
Cross-language Explicit Semantic Analysis (CL-ESA)



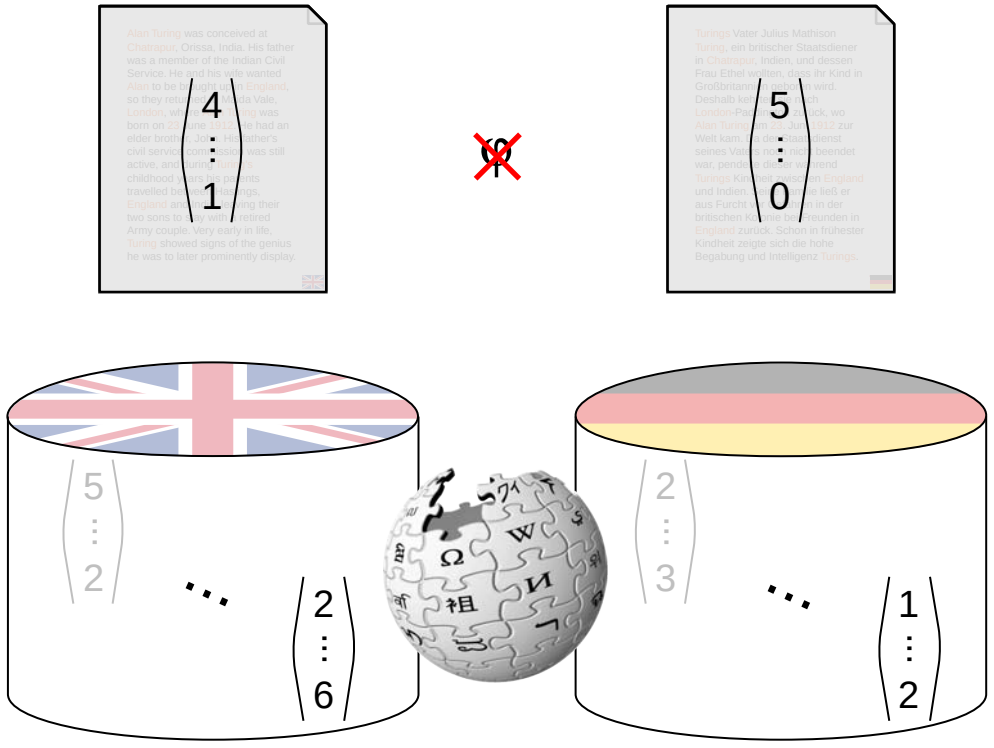
Cross-language Explicit Semantic Analysis (CL-ESA)



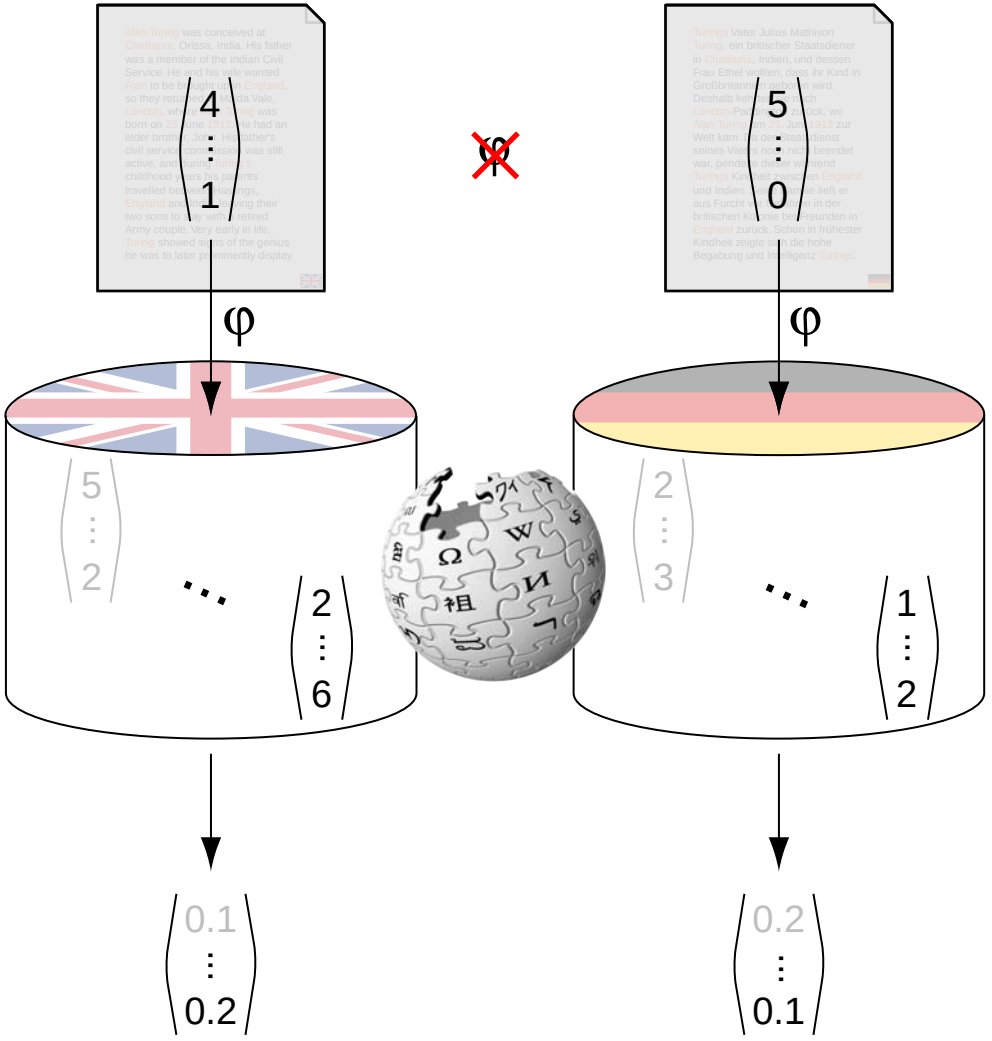
Cross-language Explicit Semantic Analysis (CL-ESA)



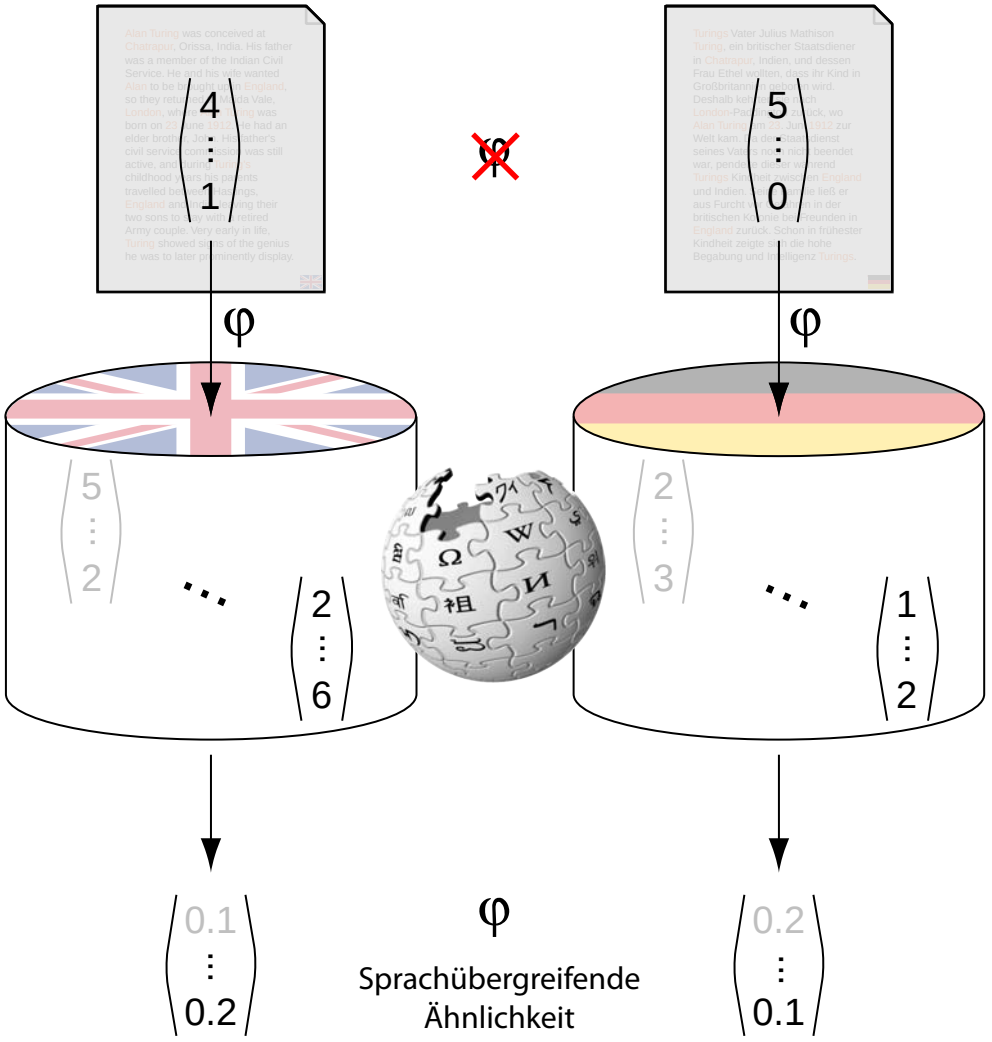
Cross-language Explicit Semantic Analysis (CL-ESA)



Cross-language Explicit Semantic Analysis (CL-ESA)



Cross-language Explicit Semantic Analysis (CL-ESA)



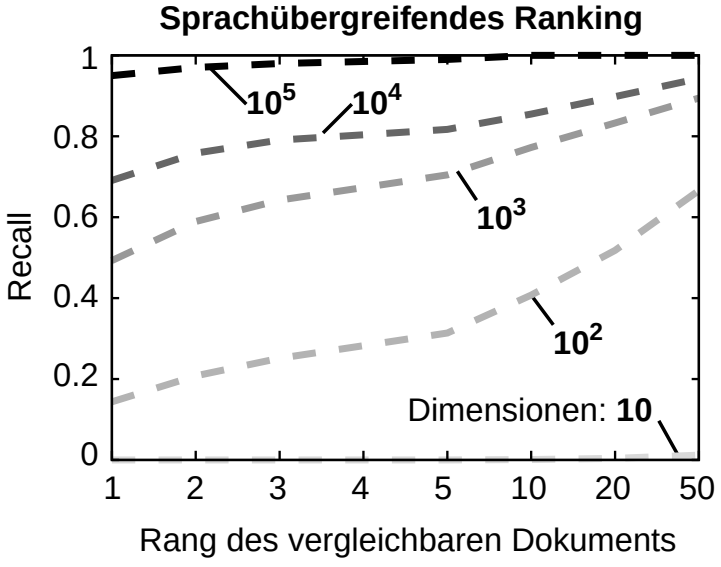
Cross-language Explicit Semantic Analysis (CL-ESA)

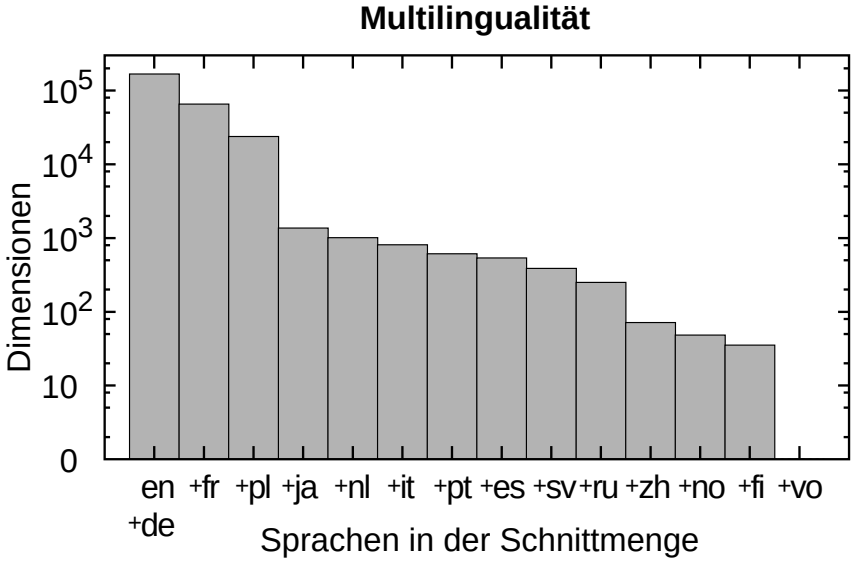
Experimente

1. Sprachübergreifendes Ranking
2. Bilinguale Rangkorrelation
3. Sprachübergreifende Ähnlichkeitsverteilung
4. Tradeoff Effektivität und Effizienz (Dimensionalität)
5. Multilingualität (Zahl gleichzeitig repräsentierbarer Sprachen)
6. Laufzeit

- Vergleich mit zwei State-of-the-Art-Modellen
- Verwendung zweier multilingualer Testkollektionen
- Vergleich auf 6 Sprachpaarungen
- > 100 Millionen Ähnlichkeiten berechnet

Cross-language Explicit Semantic Analysis (CL-ESA)

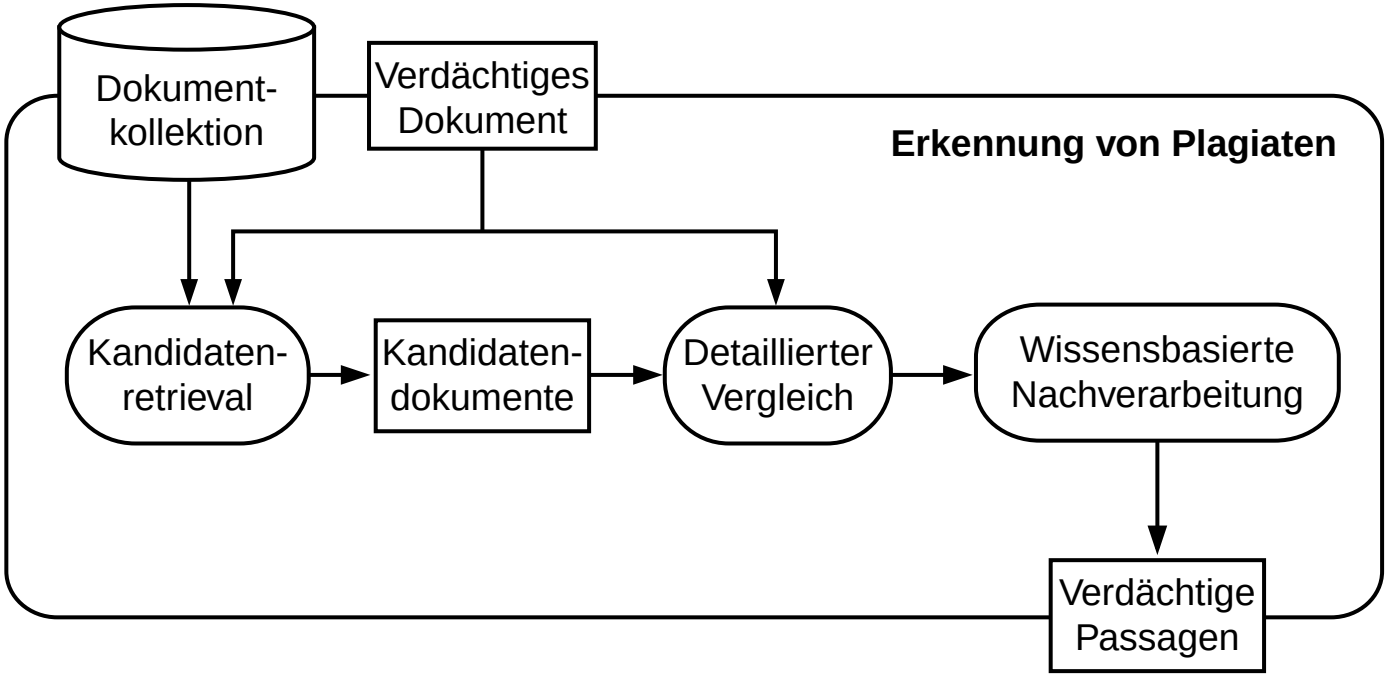




Wie Maschinen Plagiate erkennen

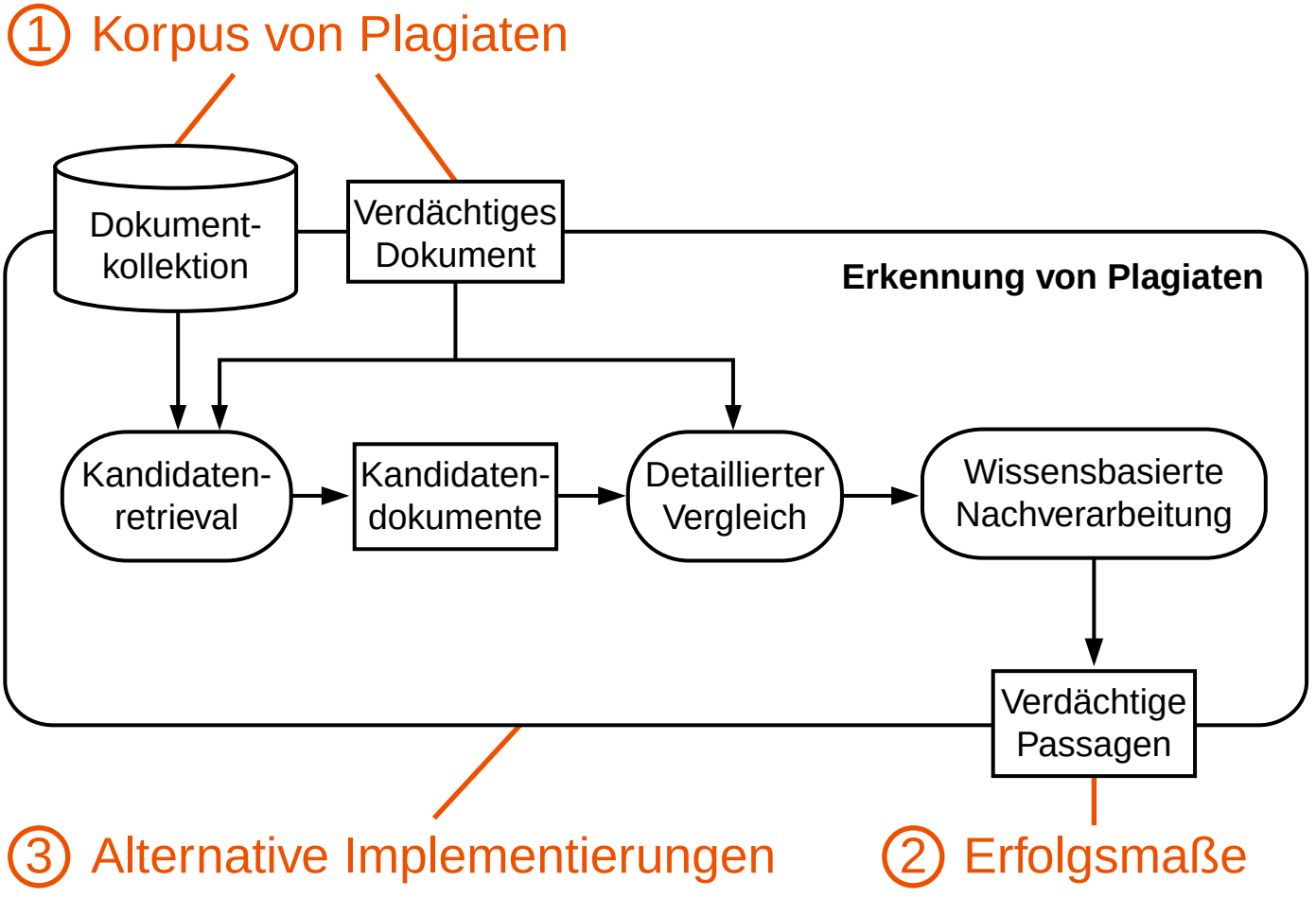
Wie Maschinen Plagiate erkennen

- Wie können sprachübergreifende Plagiate erkannt werden?
- Wie verlässlich sind Algorithmen zur Plagiatserkennung?



Wie Maschinen Plagiate erkennen

- Wie können sprachübergreifende Plagiate erkannt werden?
- Wie verlässlich sind Algorithmen zur Plagiatserkennung?



Studie zur Evaluierung von Plagiatserkennungsalgorithmen

Evaluierungsaspekt	Text	Code
① *Korpusakquise*		
Verfügbares Korpus	20%	18%
Eigenes Korpus	80%	82%
Korpusgröße [# Dokumente]		
$[1, 10)$	11%	10%
$[10, 10^2)$	19%	30%
$[10^2, 10^3)$	38%	33%
$[10^3, 10^4)$	8%	11%
$[10^4, 10^5)$	16%	4%
$[10^5, 10^6)$	8%	0%

Evaluierungsaspekt	Text	Code
② *Erfolgsmaße*		
Precision, Recall	43%	18%
Manuell, Ähnlichkeit	35%	69%
Laufzeit	15%	1%
Andere	7%	12%
③ *Vergleichende Experimente*		
Keine	46%	51%
Parameter	19%	9%
Andere Algorithmen	35%	40%

- 205 Papiere wurden analysiert
- Ist es möglich, eine Trendwende herbeizuführen?

Internationaler Wettbewerb zur Plagiatserkennung

- Erste standardisierte Testumgebung
- Ausgerichtet jährlich seit 2009

① Große neue Korpora von Plagiaten (PAN Plagiarism Corpus 2009-2012)
② Maßgeschneiderte Erfolgsmaße (Plagdet, Precision, Recall, Granularity)
③ 32 teilnehmende Forschergruppen (35 Anmeldungen für 2012)

PAN: Korpuskonstruktion

Ziel:

- Eine möglichst große, repräsentative Auswahl von plagiierten Dokumenten sowie den verwendeten Quelldokumenten.

Probleme:

- Reale Plagiatsfälle nicht im großen Maßstab verfügbar
- Abhängige Variablen unbekannt

Möglichkeiten:

- Plagiate automatisch generieren (seit 2009 im PAN Plagiatskorpus)
- Plagiate halbautomatisch generieren (seit 2010 via Mechanical Turk)
- Plagiate manuell erstellen (seit 2012 via oDesk)
- Ähnliche Textwiederverwendungen anderer Domänen sammeln
- Reale Plagiate suchen / sammeln (seit 2011 für deutsche Dissertationen)

➜ Crowdsourcing als Schlüssel für große Skalen, Realismus und Diversität

Zusammenfassung & Fazit

Sprachübergreifende Textähnlichkeit

- CL-ESA benötigt keine Wörterbücher, parallele Korpora, oder Übersetzer
- Vergleichbare Korpora sind für viele Sprachpaarungen leicht erstellbar

Evaluierung von Plagiatserkennungsverfahren

- Ergebnisse von PAN zeigen, dass noch viel Luft nach oben ist
- Das beste, was ein Plagiatserkenner in der Praxis leisten kann ist, das Durchkommen mit plumpen Plagiaten schwieriger zu machen
- Eigentlich werden nicht Plagiate, sondern bloß verdächtige Textwiederverwendungen erkannt

Wettbewerb zur Plagiatserkennung PAN

- Webseite: http://pan.webis.de
- Korpora: http://www.webis.de/research/corpora
- Community: http://groups.google.com/group/pan-workshop-series