Plagiatserkennung

Referat von William Hesse und Eduardo C. Schneider
im Seminar Computerlinguistik II
Universität zu Köln
SoSe2020

Warum (maschinelle) Plagiatserkennung?

- Plagiat – Diebstahl geistigen Eigentums
- Internet begünstigt Plagiieren
- Internet macht manuelle Plagiatserkennung quasi unmöglich
- Einsatzbereiche:
  - Wissenschaft (Universitäten, Magazine, Erfindungen,…)
  - Kunst (Literatur, Fotografie, Musik,…)

Arten von Plagiaten

- **Word for Word (C&P)**
- Word-switch
- Paraphrasierung

Plagiat	Original
Ein Plagiat ist die Aneignung fremder geistiger Leistungen und kann sich unter anderem auf die Übernahme fremder Texte oder Ideen beziehen.	Ein Plagiat ist die Aneignung fremder geistiger Leistungen und kann sich unter anderem auf die Übernahme fremder Texte oder Ideen beziehen.

Arten von Plagiaten

- Word for Word (C&P)
- **Word-switch**
- Paraphrasierung

Plagiat
Ein Plagiat ist die *Inanspruchnahme* fremder geistiger *Errungenschaften* und kann sich unter anderem auf die Übernahme fremder Texte oder Ideen beziehen.

Original
Ein Plagiat ist die Aneignungfremder geistiger Leistungen und kann sich unter anderem auf die Übernahme fremder Texte oder Ideen beziehen.

Arten von Plagiaten

- Word for Word (C&P)
- Word-switch
- **Paraphrasierung**

Plagiat
Die Anmaßung fremder geistiger Leistungen – unter anderem die Übernahme fremder Ideen oder Texte – wird als Plagiat bezeichnet.

Original
Ein Plagiat ist die Aneignung fremder geistiger Leistungen und kann sich unter anderem auf die Übernahme fremder Texte oder Ideen beziehen.

Arten von Plagiaten

<u>Komplexere Formen:</u>

- Übersetzung
- Idea plagiarism
- Organisational/Style plagiarism

Verfahrensweisen der PE

Menschliche Identifikation

- traditionelle Form der Identifizierung von Plagiaten in schriftlichen Arbeiten
- langwierige und zeitaufwändige Lese- und Suchaufgabe
- Inkonsistenzen bei der Identifizierung von Plagiaten innerhalb einer Untersuchung

Text-Matching-Software (TMS) / "Plagiat-Erkennungssoftware" PDS (Plagiate detection software)

- existiert Open-Source und kommerziell
- erkennt an sich keine Plagiate per se, sondern findet in einem Dokument bestimmte Textpassagen, die mit dem Text in anderen Dokumenten übereinstimmen.

Methoden: Generelle Ansätze

Externe Erkennungssysteme

- vergleichen ein verdächtiges Dokument mit einer referentiellen Sammlung von Dokumenten
- ausgewähltes Dokumentmodell + vordefinierten Ähnlichkeitskriterien
- -> alle Dokumente abrufen, die Text enthalten, der bis zu einem Grad über einem ausgewählten Schwellenwert für Text im verdächtigen Dokument liegt.

Intrinsische PDS

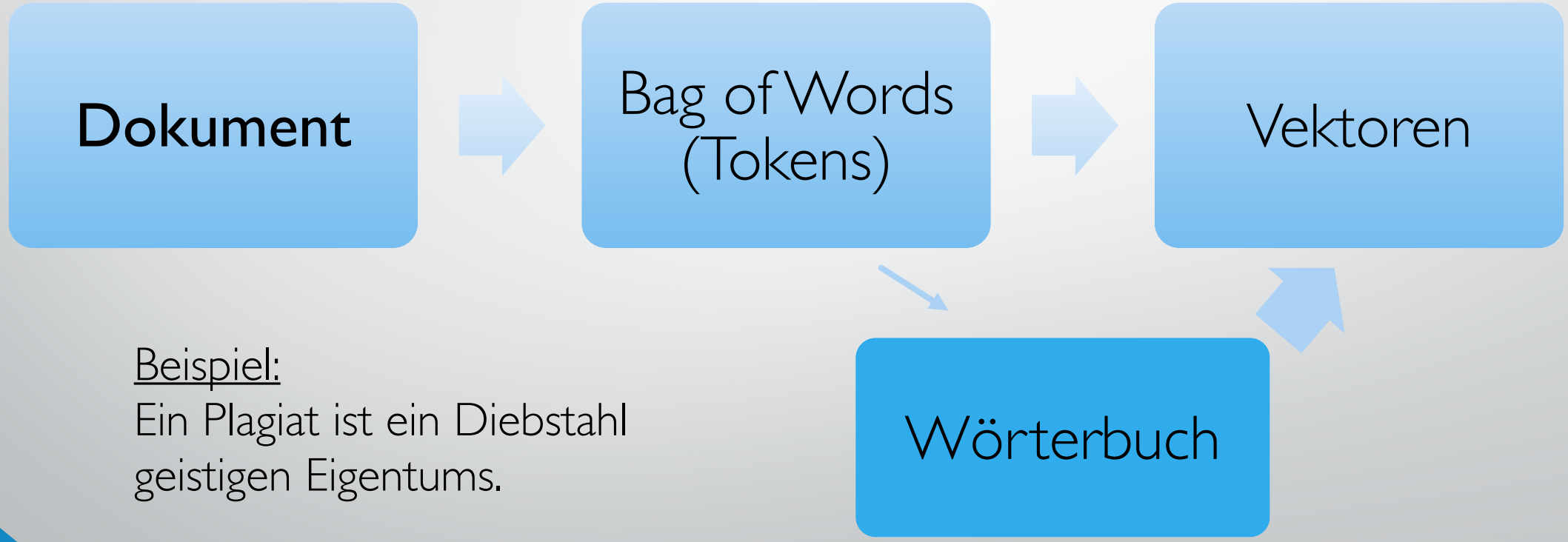
- analysieren ausschließlich den auszuwertenden Text, ohne Vergleiche mit externen Dokumenten
- -> Änderungen im einzigartigen Schreibstil eines Autors als Indikator für potenzielles Plagiat erkennen.
- ohne menschliches Urteilsvermögen unmöglich Plagiate zuverlässig zu identifizieren.
- Ähnlichkeiten werden mit Hilfe vordefinierter Dokumentmodelle berechnet und ergeben eventuell. sogenannte false positives, falsch positive Ergebnisse
- -> Kombination von Methoden basierend auf beiden Ansätzen empfohlen

Methode: String matching

- In der Informatik weit verbreiteter Ansatz
- Strings = Zeichenketten
- Direkter Vergleich auf wörtliche Textüberschneidungen (-> Vergleich von Teilstrings)
- Problem: erfordert enorme Rechenleistung, vor allem bei der Überprüfung großer Dokumentenansammlungen

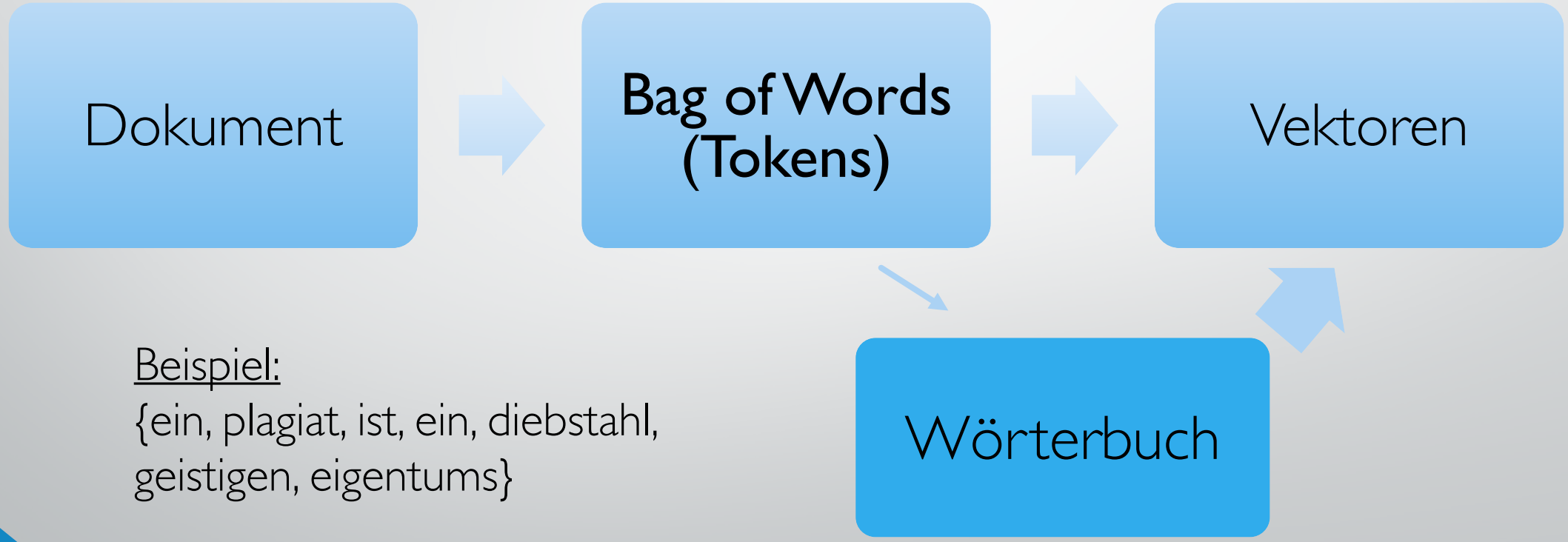
=> Sehr effektiv zur Erkennung von Copy&Paste, aber bei komplexeren Formen von Plagiaten relativ nutzlos

Methode: Bag of Words



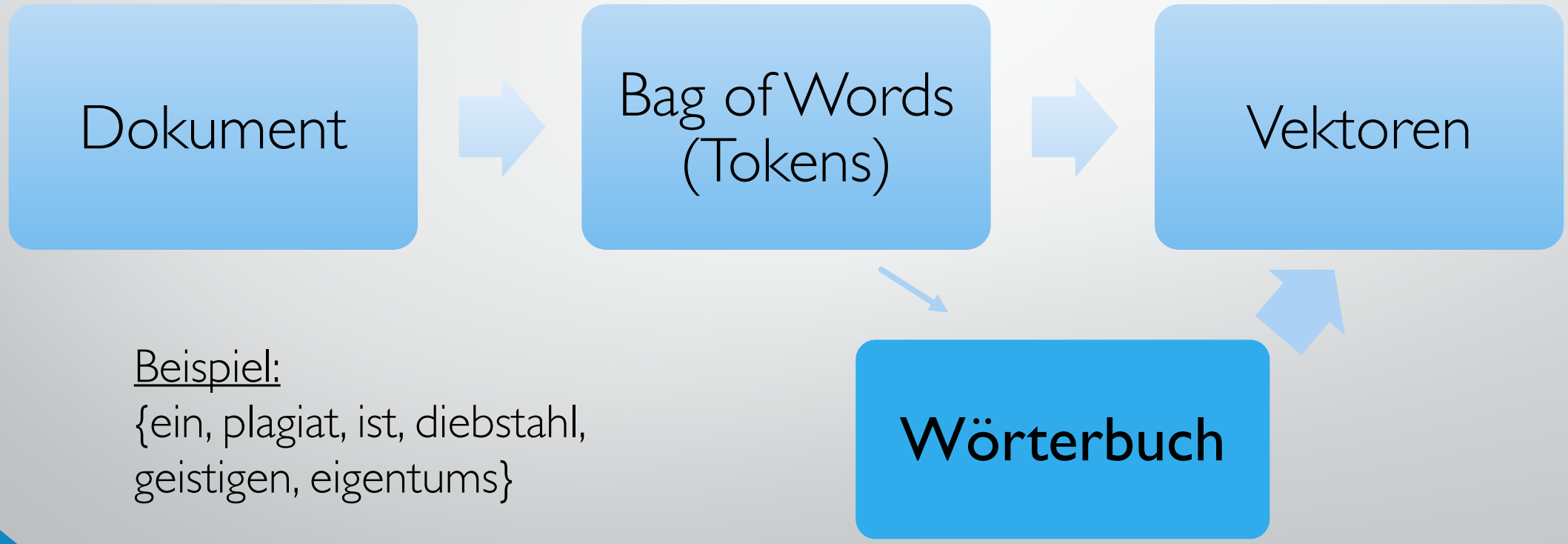
<u>Beispiel:</u>
Ein Plagiat ist ein Diebstahl geistigen Eigentums.

Methode: Bag of Words



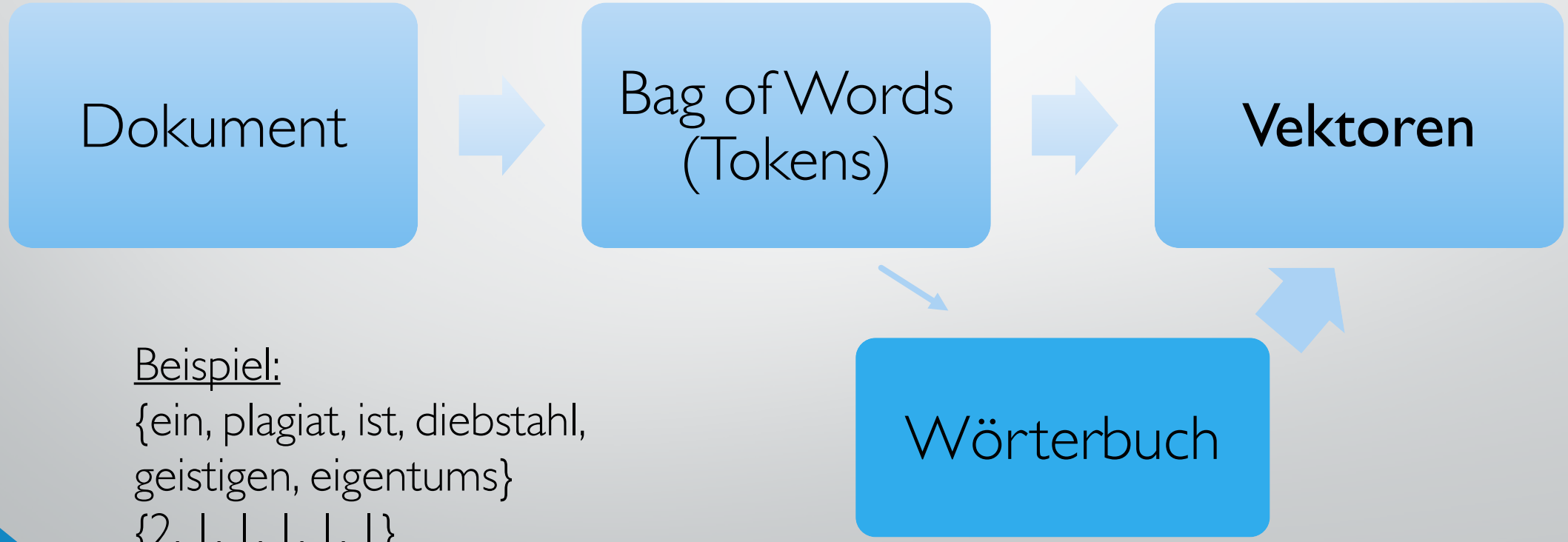
Beispiel:
{ein, plagiat, ist, ein, diebstahl, geistigen, eigentums}

Methode: Bag of Words



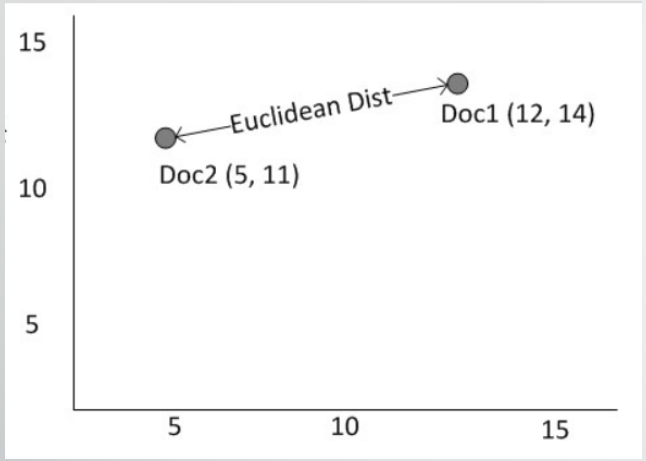
Beispiel:
{ein, plagiat, ist, diebstahl, geistigen, eigentums}

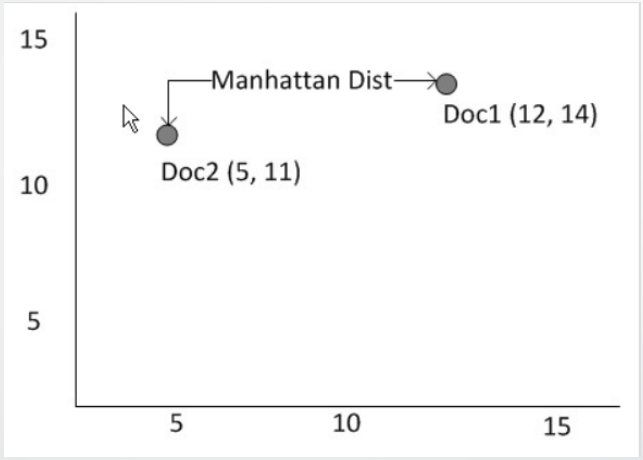
Methode: Bag of Words

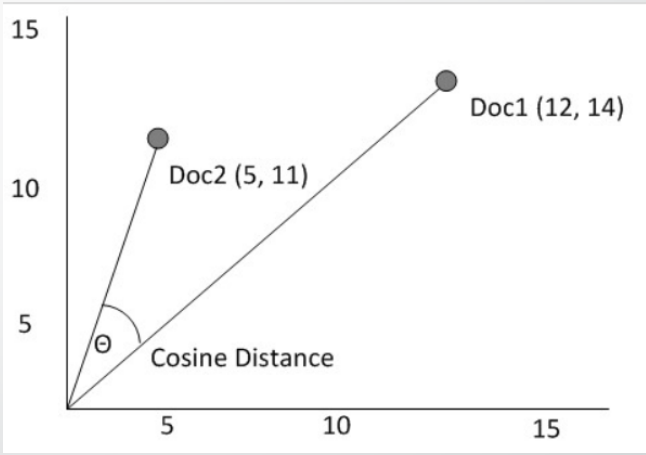


Beispiel:
{ein, plagiat, ist, diebstahl, geistigen, eigentums}
{2, 1, 1, 1, 1, 1}

Methode: Bag of Words







<u>Beispiel:</u>
{ein, plagiat, ist, diebstahl, geistigen, eigentums}
{2, 1, 1, 1, 1, 1}

{5, 11, 2, 1, 3}

{4, 9, 5, 2, 2}

{12, 14, 9, 4, 23}

Algorithmen: Bag of Words

Problem	Lösungansatz
Grundannahme: Gleiche Worte <-> gleicher Inhalt	Stopwords filtern
Auflösung von zusammenhängenden Worten (z.B. New York)	n-grams
Konjugation/Deklination (z.B. des Plagiats)	Stemmer
Auflösung von Satzstrukturen	(n-grams)

=> Effektiv zur Erkennung von: Paraphrasierung, Copy&Paste, (Idea)

Methode: Fingerprinting

- Am weitesten verbreiteter Ansatz zur Erkennung von Inhaltsähnlichkeiten
- repräsentative Zusammenfassungen von Dokumenten, indem ein Set mehrerer Teilzeichenfolgen (n-Gramm) aus ihnen ausgewählt wird
- Sets stellen Fingerabdrücke dar und ihre Elemente werden als Minutien (eng. : "minutiae") bezeichnet
- Verdächtiges Dokument wird auf Plagiate überprüft, indem sein Fingerabdruck berechnet und Minutien mit einem vorberechneten Index von Fingerabdrücken für alle Dokumente einer referenziellen Sammlung abgefragt werden.
- Minutien, die mit denen anderer Dokumente übereinstimmen, weisen auf gemeinsame Textsegmente hin und deuten auf ein potenzielles Plagiat hin, wenn sie eine gewählte Ähnlichkeitsschwelle (z.B. bestimmte übereinstimmende Satzlänge und Struktur) überschreiten.
- Rechenressourcen und Zeit sind einschränkende Faktoren:
- normalerweise wird nur eine Teilmenge von Minutien verglichen, um Berechnung zu beschleunigen und Überprüfungen in sehr großen Sammlungen z.B. Internet zu ermöglichen.

Methode: Zitatanalyse - zitatorientierte Plagiatserkennung (CbPD)

- einziger Ansatz zur Plagiatserkennung, der nicht auf der Ähnlichkeit des Textes beruht.
- CbPD untersucht die Zitier- und Referenzinformationen in Texten, um ähnliche Muster in den Zitiersequenzen zu identifizieren.
- Ansatz für wissenschaftliche Texte oder andere akademische Dokumente, die Zitate enthalten.
- Relativ jung und neuartig zur Erkennung von Plagiaten
- Ähnliche Reihenfolgen und Nähen der Zitate in den untersuchten Dokumenten sind die Hauptkriterien für die Berechnung der Ähnlichkeiten von Zitiermustern.
- Zitiermuster stellen Teilsequenzen dar, die nicht ausschließlich Zitate enthalten, die von den verglichenen Dokumenten geteilt werden.
- Anzahl oder relativer Anteil gemeinsamer Zitate im Muster sowie Wahrscheinlichkeit, dass Zitate in einem Dokument gleichzeitig auftreten, werden auch berücksichtigt, um den Ähnlichkeitsgrad der Muster zu quantifizieren.

Methode: Stilometrie

- Intrinsische Erkennung von Plagiaten mithilfe des Vergleichs sprachlicher Ähnlichkeit
- statistische Methoden zur Quantifizierung des einzigartigen Schreibstils eines Autors
- hauptsächlich für Zuweisung von Autoren oder die intrinsische PD verwendet.
- Erstellen und Vergleichen von stilometrischen Modellen für verschiedene Textsegmente -> Erkennung v. Passagen , die sich stilistisch von anderen unterscheiden und daher möglicherweise plagiiert wurden
- Stilistischen Unterschiede zwischen plagiierten und ursprünglichen Segmenten ermöglichen Identifizierung von getarnten und paraphrasierten Plagiaten
- Stilometrische Vergleiche schlagen fehl, wenn Segmente so stark umschrieben sind, dass sie dem persönlichen Schreibstil des Plagiators ähnlicher sind oder wenn ein Text von mehreren Autoren zusammengestellt wurde.
- Aufgrund des menschlichen Arbeitsaufwandes eher bei moderaten Textmengen geeignet.

Praxisbeispiel

Fazit

- Die Leistung von PDS hängt stark von der Art des vorhandenen Plagiats ab
- Mit Ausnahme der Zitiermusteranalyse beruhen alle Erkennungsansätze auf Textähnlichkeit.
- Daraus folgt, dass die Erkennungsgenauigkeit auch abnimmt, je mehr Plagiatsfälle z.b. paraphrastisch verschleiert werden.
- Wörtliche Kopien (Copy-and-Paste-Plagiate) bzw. offensichtliche Urheberrechtsverletzung oder leicht getarnte Plagiatsfälle, können von aktuellen externen PDS mit hoher Genauigkeit erkannt werden, wenn die Quelle für die Software zugänglich ist.
- Sprachübergreifende Plagiatserkennung (CLPD) NOCH nicht als ausgereifte Technologie angesehen
  -> in der Praxis bisher keine zufriedenstellenden Erkennungsergebnisse erzielt

Fazit

- Insbesondere Teilstring-Matching-Verfahren erzielen eine gute Leistung bei Copy & Paste -Plagiaten, da sie häufig verlustfreie Dokumentmodelle wie Suffixbäume verwenden.

- Leistung von Systemen, die beim Erkennen von Kopien Fingerprinting oder Wortbeutelanalysen verwenden, hängt vom Informationsverlust ab, der durch das verwendete Dokumentationsmodell entsteht.

- Durch die Anwendung flexibler Auswahlstrategien sind sie im Vergleich zu Teilstring-Matching-Verfahren besser in der Lage, moderate Formen von verschleiertem Plagiat zu erkennen.

- zitatbasierte Plagiaterkennung mithilfe der Zitiermusteranalyse kann stärkere Paraphrasen und Übersetzungen mit höheren Erfolgsraten identifizieren, da sie unabhängig von Textmerkmalen arbeitet

- von der Verfügbarkeit ausreichender Zitierinformationen abhängig -> auf akademische Texte beschränkt

- bleibt textbasierten Ansätzen bei der Erkennung kürzerer plagiierter Passagen unterlegen, die für C&P oder Schütteln und Einfügen (shake-and-paste plagiarism -> Mischen leicht veränderter Fragmente aus verschiedenen Quellen) schnell und effektiv verlässliche Ergebnisse erzielen.