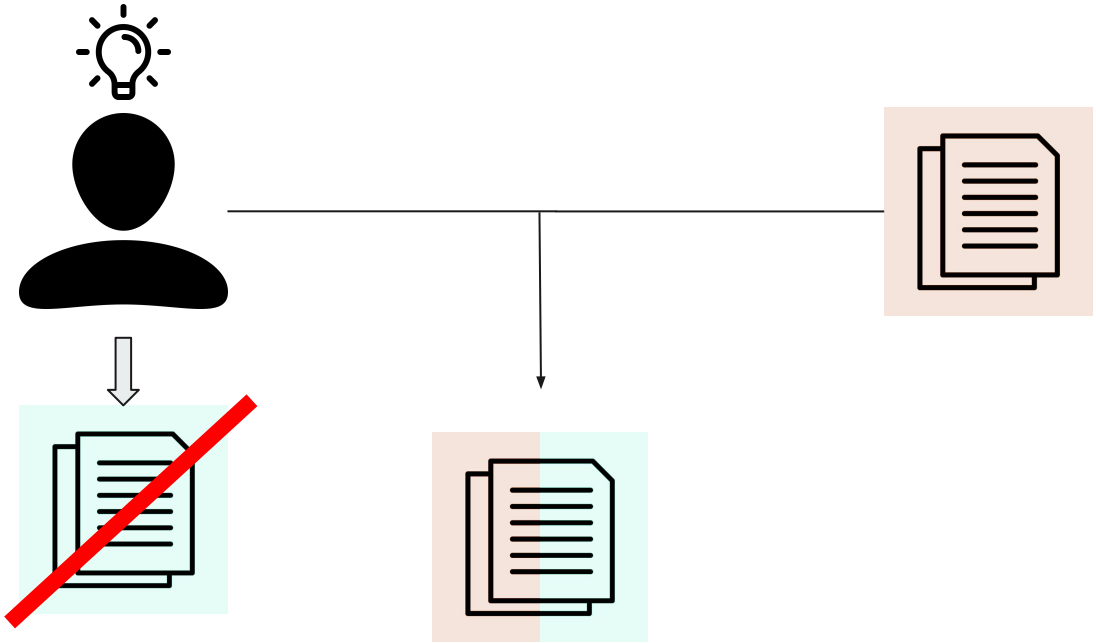
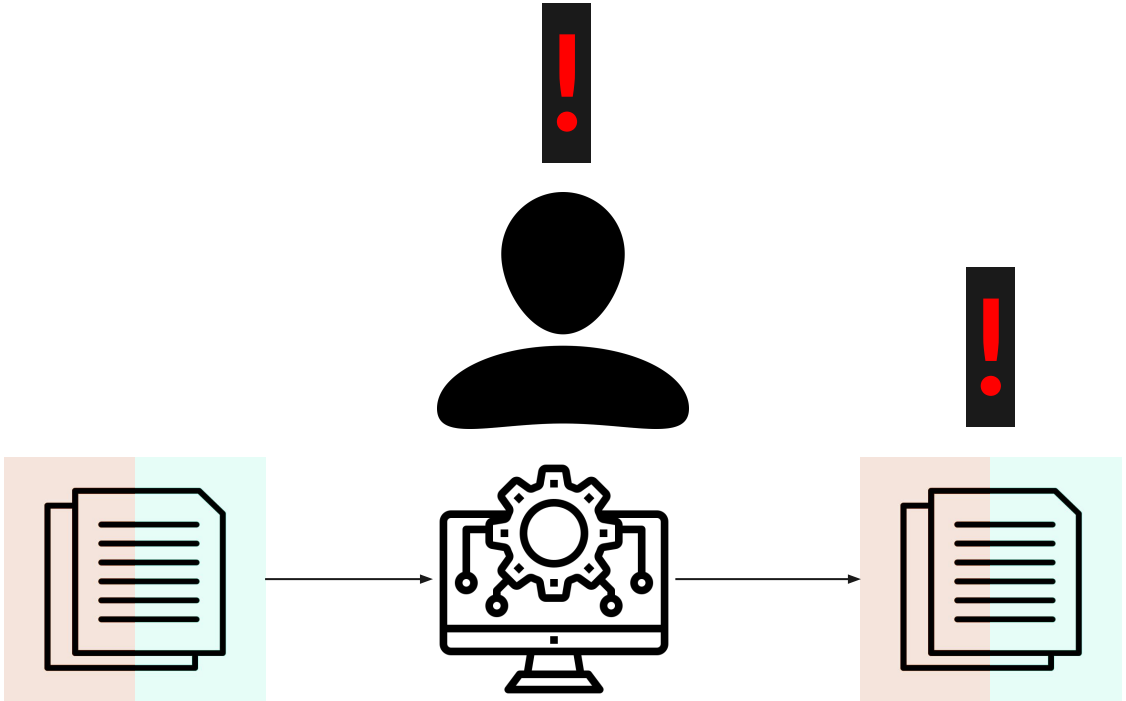
Plagiatserkennung

SoSe 2022 Universität zu Köln
Computerlinguistik II
Präsentation von Mario Müller, David Müller

Gliederung

1. Was sind Plagiate?
2. Plagiatserkennung
   a. Was ist Plagiatserkennung
   b. International Competition on Plagiarism Detection
3. Methoden der Erkennung
   a. Extrinsische Plagiatserkennung
   b. Intrinsische Plagiatserkennung
4. Praxisbeispiel
   a. PlagAware (eigene Hausarbeit)
5. Fazit





Was sind Plagiate?

“Wissenschaftliches Fehlverhalten liegt insbesondere vor, wenn bewusst, willentlich oder grob fahrlässig unter Anmaßung der Autorinnen- oder Autorenschaft (Plagiat) geistiges Eigentum anderer durch die unbefugte Verwertung verletzt wird.” [2]

- Prüfungsordnung für das Bachelorstudium an der Philosophischen Fakultät der Universität zu Köln

Was sind Plagiate?

“Die unrechtmäßige Aneignung von Gedanken, Ideen o. Ä. eines anderen auf künstlerischem oder wissenschaftlichem Gebiet und ihre Veröffentlichung; Diebstahl geistigen Eigentums” [1]

- Oxford Languages Wörterbuch

Was ist Plagiatserkennung?

- Verfahren zum Auffinden von Plagiaten oder Urheberrechtsverletzungen in einem Werk oder Dokument [3]
- Es gibt manuelle und automatisierte Methoden der Plagiatserkennung
  - In unserem Fall geht es um automatisierte Prozesse in maschinenlesbaren Texten
- Alle Plagiatserkennungsansätze folgen einem generischen Suchprozessschema [4]
  - 1. heuristische Suche -> problemlösende Techniken und Suchmethodiken [5]
  - 2. detaillierte Analyse
  - 3. wissensbasierte Nachbearbeitung

International Competition on Plagiarism Detection

"*... when asked to name the best algorithm or the best system for plagiarism detection, hardly any evidence can be found to make an educated guess among the alternatives.*" [4]

- Um das zu ändern wurde die International Competition on Plagiarism Detection ins Leben gerufen
- Bewertungsrahmen für die automatische Plagiatserkenner, der aus einem umfangreichen Plagiatskorpus und Qualitätsmaßen für die Erkennung besteht

Methoden der Erkennung

- Extrinsische Plagiatserkennung
  - Vergleich des Textes mit einer Menge an Texten (Korpus), um das Original zu finden

- Intrinsische Plagiatserkennung
  - Untersuchungen ausschließlich des Textes selbst, um Auffälligkeiten zu finden

Extrinsische Methoden

- Vergleich des Textes mit Korpus - je nach Methode die verwendet werden soll, muss der Korpus annotiert sein
- Möglichkeiten und Probleme
  - Wörtliche Übernahme von Passagen kann mit einfacher Suche und Vergleich gefunden werden
  - Problem: das kann aber auch ein Zitat sein → menschliche Überprüfung notwendig

Extrinsische Methoden

- Verschiedene Methoden (auf Untersuchungstext und Korpus angewendet) können Ergebnisse verbessern:
  - Untersuchung von mehreren Wörtern in Folge (n-grams) und Vergleich der Wahrscheinlichkeit, dass diese gemeinsam auftreten
  - Einbeziehen von syntaktischen Eigenschaften, also Funktion der Wörter im Satz

  → kann Paraphrasierungen und Wortumstellungen erkennen

Extrinsische Methoden

Zitatanalyse:

- Untersuchung nicht des Textes, der paraphrasiert sein kann, sondern der vorkommenden Zitate
- Ähnliche Quellen und Ähnlichkeit der Reihenfolge der Zitate weist auf Plagiat hin
- Kann auch Übersetzungen erkennen

Extrinsische Methoden

Probleme extrinsischer Methoden:

- Nicht alle Originaltexte stehen in Korpora zur Verfügung
- Paraphrasierungen, Umstellung und vor allem Übersetzungen sind nur schwer zu erkennen

Intrinsische Methoden

- Intrinsische Plagiatserkennung sucht nach Unregelmäßigkeiten innerhalb des Textes, ohne diesen mit anderen Texten oder Quellen zu vergleichen
- Es werden verschiedene Methoden in Verbindung angewendet, wie
    - Stilistische Methoden (Stilometrie): Schreibstil soll eindeutig auf Autor hinweisen. Nach Aufteilung des Textes sollten unterschiedliche Schreibstile auffallen.

Mögliche Merkmale:

- Bevorzugtes Vokabular
- Verwendete Phrasen, Satz- und Sonderzeichen, Emoticons
- Rechtschreib-, Grammatik- und Tippfehler
- (unbewusst) verwendete Grammatikstrukturen
- Formatierung: Häufigkeit von Absätzen etc.

Intrinsische Methoden

Probleme intrinsischer Methoden:

- Können nur Hinweise geben auf auffällige Unterschiede im Text mit Wahrscheinlichkeit
- Kein Finden eines kopierten Originaltextes
- Unterschiede im Text könnten auch von einem Autor stammen, z.B. bei mehrjähriger Arbeit an Dissertation
- Überarbeiten, Korrekturen, Lektorat: Texte sind auch Ergebnis von Teamarbeit

Praxisbeispiel:
PlagAware

Fazit

Plagiatserkennungssoftware bietet die Möglichkeit automatisch plagiierte Stellen in Texten zu finden. Sie können jedoch bislang nicht manuelle Plagiatssuche ersetzen, denn die Algorithmen können ausgetrickst werden und bieten bislang nicht immer sichere Ergebnisse.

> Dennoch: Plagiatserkennungssoftware kann die manuelle Plagiatssuche unterstützen und verbessern.

Quellen

[1] Oxford Languages Wörterbuch: Plagiat Definition: https://www.google.com/search?q=Plagiat+Definition (abgerufen am 23.05.22)

[2] Prüfungsordnung für das Bachelorstudium an der Philosophischen Fakultät der Universität zu Köln; abgerufen am 23. Juni 2021: https://phil-fak.uni-koeln.de/sites/phil-fak/lehre_studium/bachelor/Pruefungsordnungen_PO2015/PO-2015-Bachelorstudiengaenge.pdf (abgerufen am 23. Juni 2021)

[3] [2] Oakes, Michael P. (2014): Author Profiling and Related Applications. In Mitkov, Ruslan (Hrsg.): The Oxford Handbook of Computational Linguistics 2nd edition. Oxford University Press.

[4] Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein and Paolo Rosso (2010). 'An evaluation framework for plagiarism detection'. Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, 23–27 Aug.: 997–1005.

[5] Sanjay Jena, Nils Liebelt (2004). 'Heuristische Algorithmen am Beispiel des A*-Algorithmus / 8-Puzzle': http://www.gm.fh-koeln.de/~hk/lehre/ala/ws0506/Praktikum/Projekt/E_gelb/ALA-HeuristischeAlgorithmen-Jena-Liebelt.pdf (abgerufen am 23. Juni 2021)

[6] Tschuggnall, Michael: Automatisierte Plagiatserkennung in Textdokumenten: Was der Schreibstil eines Autors über die Echtheit verrät. In: Sandra Mauler, Heike Ortner, Ulrike Pfeiffenberger (Hg.): Medien und Glaubwürdigkeit. Interdisziplinäre Perspektiven auf neue Herausforderungen im medialen Diskurs. Innsbruck: Innsbruck University Press 2017 (Medien – Wissen – Bildung), S. 131–140. DOI: https://doi.org/10.25969/mediarep/1629.

[7] https://lehre.idh.uni-koeln.de/site/assets/files/2044/plagiatserkennung.pdf

[8] https://blogs.faz.net/deus/2011/08/04/plagiatserkennung-ein-steiniger-weg-fuer-die-computerlinguistik-478/