Kontrastive Textanalyse mit *Pydistinto*

ein Python-Paket zur Nutzung unterschiedlicher Distinktivitätsmaße

Scan me

Distinktivitätsmaße ermöglichen es Forschern, Merkmale (z. B. Wörter oder Wortarten) zu extrahieren, die für eine bestimmte Textgruppe im Vergleich zu einer anderen Textgruppe charakteristisch oder „unterscheidungskräftig“ sind. Um den Einsatz relevanter Maße für die kontrastive Textanalyse zu erleichtern und das Bewusstsein für die Vielfalt der Maße zu schärfen, entwickeln wir ein Python-Paket mit dem Namen ***Pydistinto***.

Pydistinto ist einfach!

Kontrastive Textanalyse in drei Schritte:

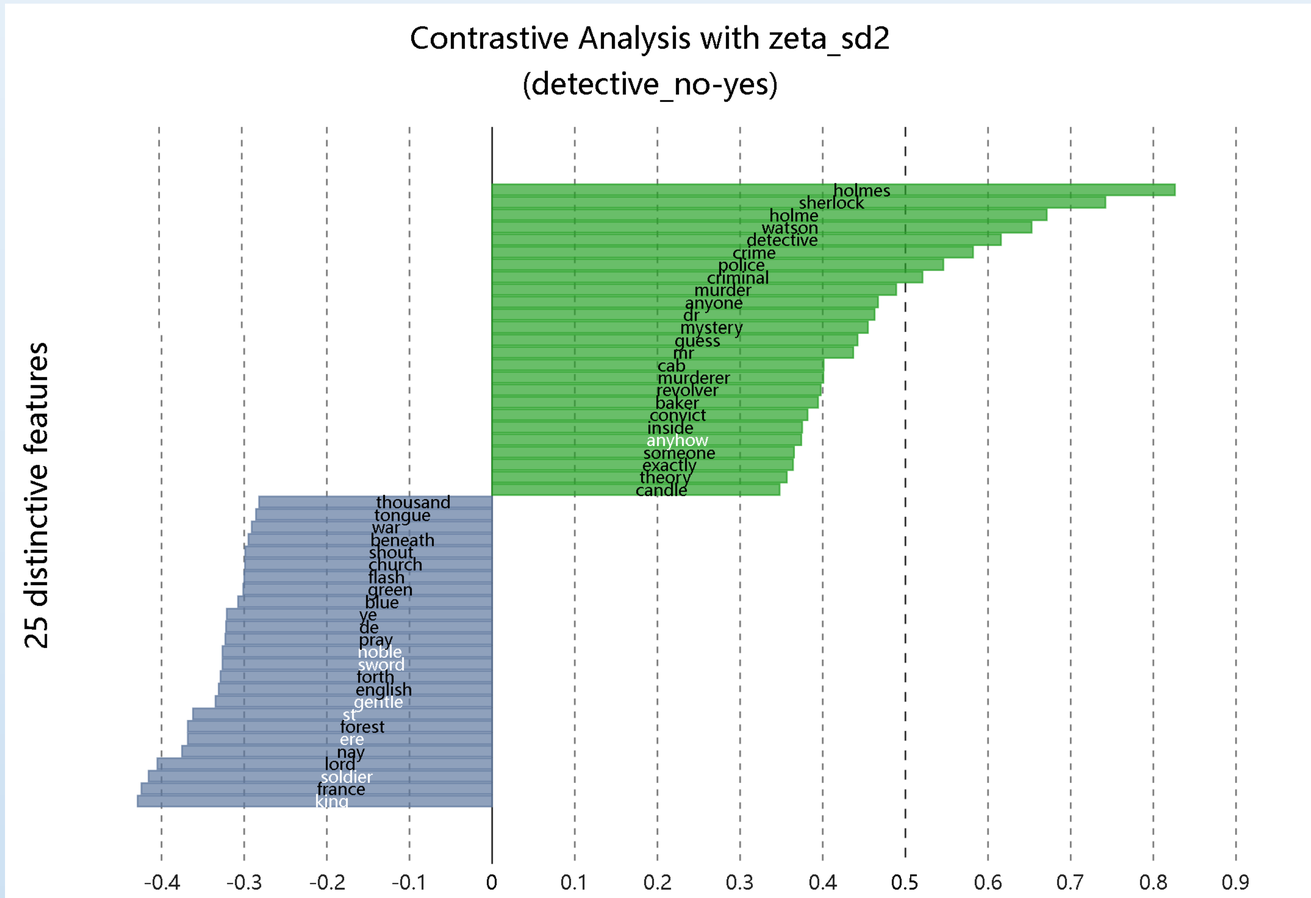
1. Parameter (Korpus, Sprache, Segmentlänge etc.) anpassen
2. preprocessing_before_running_pydistinto.py ausführen
3. run_pydistinto_beginners.py ausführen

Zielkorpus (Detektivroman_yes)

vs.

Vergleichskorpus (Detektivroman_no)

Distinktive Features

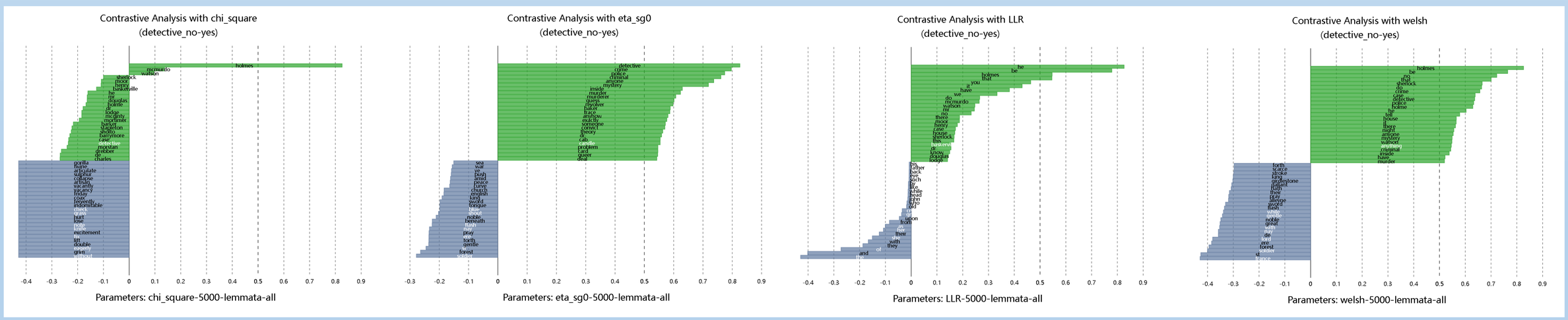


Pydistinto ist vielfältig!

Insgesamt werden **9 unterschiedliche Distinktivitätsmaße** implementiert (Und es werden noch mehr Maße hinzukommen):

1. Chi-Squared Test
2. Log-likelihood-ratio test
3. ratio of relative frequencies
4. Welsh’s t test
5. Wilcoxon Rank-Sum test
6. Burrows’ Zeta
7. Logarithmic Zeta
8. tf-idf weighted absolute frequencies based measure
9. Eta

Die Ergebnisse variieren von Maß zu Maß und es lohnt sich die Maße genau zu vergleichen und zu evaluieren!



Deshalb ➔ Projekt ***Zeta and Company (https://zeta-project.eu)***

Pydistinto ist doch nicht einfach!

Im **Beginner-Modus** können auch weniger erfahrene Nutzende mit geringen Programmier- und Statistikkenntnissen Textkorpora vergleichen.

Wer sich für die statistischen Eigenschaften der unterschiedlichen Maße interessiert und diese vergleichen möchte, kann den **Profi-Modus** verwenden. Die Nutzenden können dann selbst darüber bestimmen, welche Maße und statistischen Eigenschaften der Features (z.B. absolute Häufigkeit, relative Häufigkeit, Dispersion) für die Berechnung der Distinktivität kombiniert werden sollen.

Universität Trier

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften

Computational LiteraryStudies

Keli Du, Julia Dudar, Christof Schöch
(‘Zeta and Company’ | https://zeta-project.eu | Universität Trier)

DHD2022, Poster Session
7 - 11 März 2022