Inhaltsverzeichnis

stylo ah online - Handbuch

Ursprung: stylo ah online - stylo ohne Installation

Im Rahmen unsere Lehrtätigkeit und unserer Forschungstätigkeit haben wir das R-Paket **stylo** eingesetzt. Wir haben dieses mittels R und Python weiterentwickelt. Vornehmlich vier Dinge sind an der R Implementierung schwierig:

1. Die Möglichkeit zur Selbstdokumentation der Experimente ist gering.
2. Die Interaktion zwischen Python und R ist problematisch.
3. Die Möglichkeit eines einfachen Einsatzes von Multiprocessing ist fraglich.
4. Die Software ist teilweise schwierig zu installieren.

Daher wollte wir die Funktionalität von **stylo** nachempfinden; ohne Installation und mit erweiterten Dokumentations- und Vergleichsmöglichkeiten ausstatten.

Links zum Thema

1. stylo das Original in R geschrieben.
2. stylo ah die R / Python Kopie mit zusätzlichen Funktionen für klassische Texte.
3. stylo ah online die originale Kopie in JavaScript implementiert.

Implementierte Analyse-Pipeline(s)

Die Implementierte Textanalyse verläuft in sieben Schnitten. Diese stellen die Abfolge bzw. die Pipeline der Verarbeitung dar:

1. Auswahl der Files (Corpus) (jedes File repräsentiert einen Text)
2. Normalisierung (Formatanpassung, Zeichenvereinheitlichung, Löschung von Struktur und Metadaten, Maskierung von Wortformen)
3. Zerlegung in Token (Wortformen, Silben, grame, Zeichen)
4. Zählung / Vektorbildung (01-Kodierung, abs. / rel. Häufigkeit, TF-IDF)
5. Auswahl aus den Vektoren (Beschränkung der Häufigkeitsliste, Häufigkeitsfenster, Culling)
6. Anwendung eines Maßes
7. Anwendung einer Cluster-Methode

Funktionsüberblick

In stylo ah online sind folgende Funktionen verfügbar:

Normalisierung:

Normalform Wortmaskierung (Stopworte) Vereinheitichen bestimmter Zeichen UV-Angleich JI-Angleich Markup-Löschen Interpunktion löschen Zeilenumbrüche löschen Elision auflösen Alpha

privativum behandeln Entfernen der Nummerierung Worttrennungen zusammenführen Vereinheitlichung des Iota subscriptum und Iota adscriptum Veränderung des Abschluss-Sigmas Diakritische Zeichen löschen Ligaturen auflösen Kleinschreibung Klammern entfernen

Normalisierung

Zerlegung (Token):

Zusätzlich ohne Konsonanten Zusätzlich ohne Vokale Zusätzlich lediglich kleine Wörter Zusätzlich lediglich große Wörter Zerlegung in Wortformen Zerlegung in Buchstaben n-grame Zerlegung in Buchstaben n-grame der Worformen Zerlegung in Wort n-grame Zerlegung in n-grame mit Lücken Zerlegung in Pseudo-Silben Zerlegung in Kopf Körper und Coda Zerlegungen in alle Permutationen von Kopf Körper und Coda

Zerlegung

Zählung:

absolute Häufigkeit relative Häufigkeit 0-1-Codierung TF-IDF Beschränkung/Spannen der Frequenzlisten (nach Rang, min-max-Angabe) Culling

Zählung

Maße:

euclidean, chebyshev, minkowski, manhatten, canberra, soerensen, gower, soergel, lorentzian, intersection, wavehedges, motyka, ruzicka, tanimoto, innerproduct, harmonicmean, cosine, kumar hasse brook, dice, fidelity, bhattacarya 1, bhattacarya 2, hellinger, jensen, jensen shannon, topsoee, kullback divergence, jeffreys, kullback leibler, squared euclidean, pearson chi squared, neyman chi squared, squared chi squared, divergence, clarck, additive symmetric chi squared, eder simple, burrows delta, argamon linear delta, eders delta, argamons quadratic delta, wasserstein 1d

Maße

Gruppierung:

hierarchische Clusterung, multidimensional scaling (MDS), tSNR (t-distributed stochastic neigbor embedding)

Gruppierung

Benutzung

GUI Konzept

Jeder Verarbeitungsschritt wird in einem Konfigurationsschritt eingestellt. Jeder Konfigurationsschritt besitzt ein zusätzliches Kommentarfeld und eine der Anzeige der Zwischenergebnissen (rechte Spalte). Die GUI empfindet die Schritte der Textanalyse nach. Einzige Ausnahme bildet die Auswahl des Corpus, die ist zwar an oberer Stelle angezeigt, aber von der Benutzung her ist das der letzte Schritt! Jedes GUI-Element ist mit einer Beschriftung versehen. Sofern die Funktion einer weiteren Erklärung bedarf, dann ist zusätzlich eine kursive Erklärung angefügt.

Überblick zur Benutzung

1. Konfiguration des Browsers
2. Konfigurieren der Analyseschritte
3. Corpus-Auswahl
4. Erneute Berechnung und Serien

1. Konfiguration des Browsers

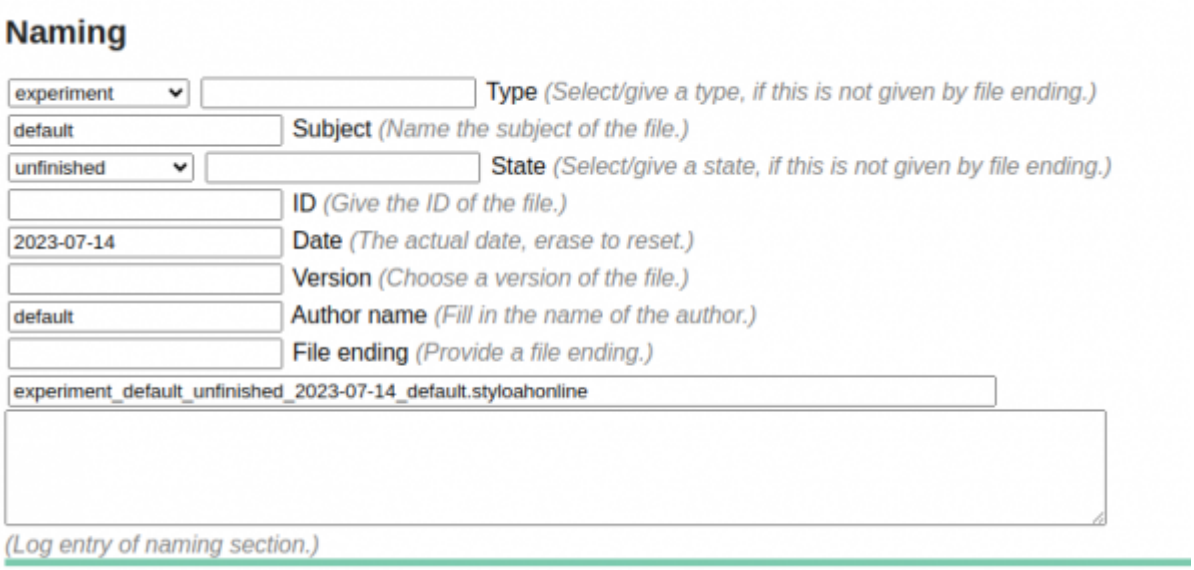
Speicherort: Legen sie einen Ordner an in welchem die Ergebnisse von stylo-ah-online gespeichert werden können. Konfigurieren Sie ihren Browser so, dass er die Dowloads in diesem Ordner ablegt. die Anleitung für Firefox:
https://support.mozilla.org/de/kb/suchen-und-verwalten-heruntergeladener-dateien#w_ziel-ordner-der-heruntergeladenen-dateien-andern

Datenbanken: Teilen sie dem Browser mit, dass er Datenbanken anlegen und nicht löschen soll. Dazu stellen sie sicher, dass die Chronik angelegt wird:
https://support.mozilla.org/de/kb/firefox-chronik-zeigt-ihre-besuchten-webseiten

Web-Konsole: stylo-ah-online ist in JavaScript geschrieben. Die Web-Konsole bietet eine Darstellung von Meldungen zum Programmablauf und Hinweise auf Fehler. Nach einem Fehler können sie die Seite neu laden und den Ablauf erneut starten. Es bietet sich an die Web-Konsole anzuzeigen. Für Firefox aktiviert man diese so:
https://firefox-source-docs.mozilla.org/devtools-user/tools_toolbox/index.html

2. Konfigurieren der Analyseschritte

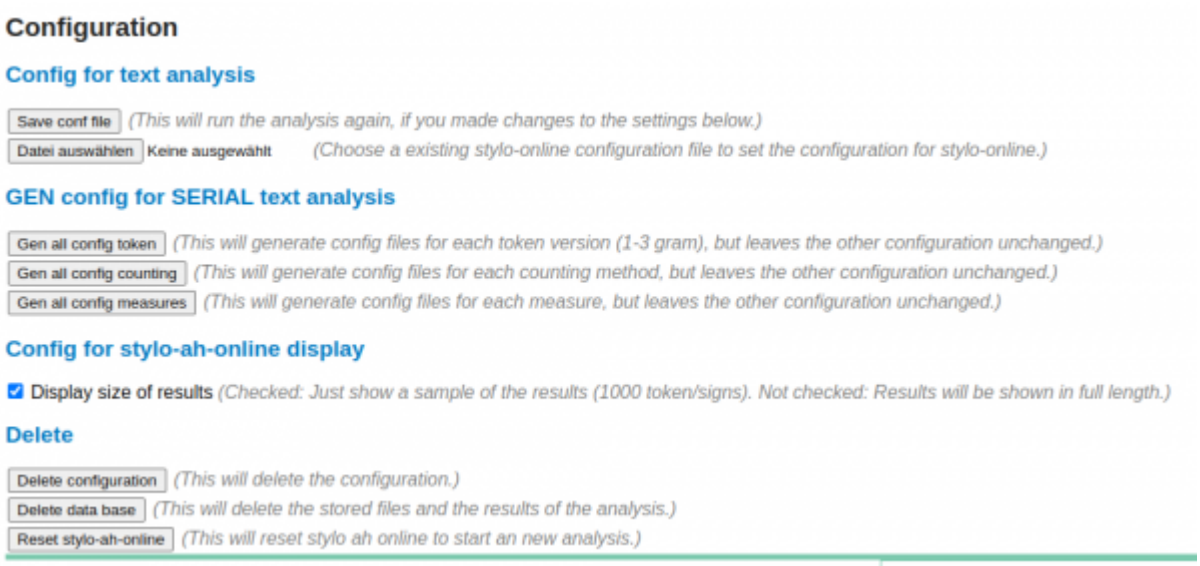
2.1 Arbeitsschritt: Benennung



Das Benennungsmodel beinhaltet zusammengesetzte Namen für die Experimente. Dabei werden die einzelnen Teile der Benennung durch Unterstriche abgetrennt. Leerzeichen in Einträgen werden durch Bindestriche ersetzt. Die einzelnen Einträge der Benennung erlauben verschiedene

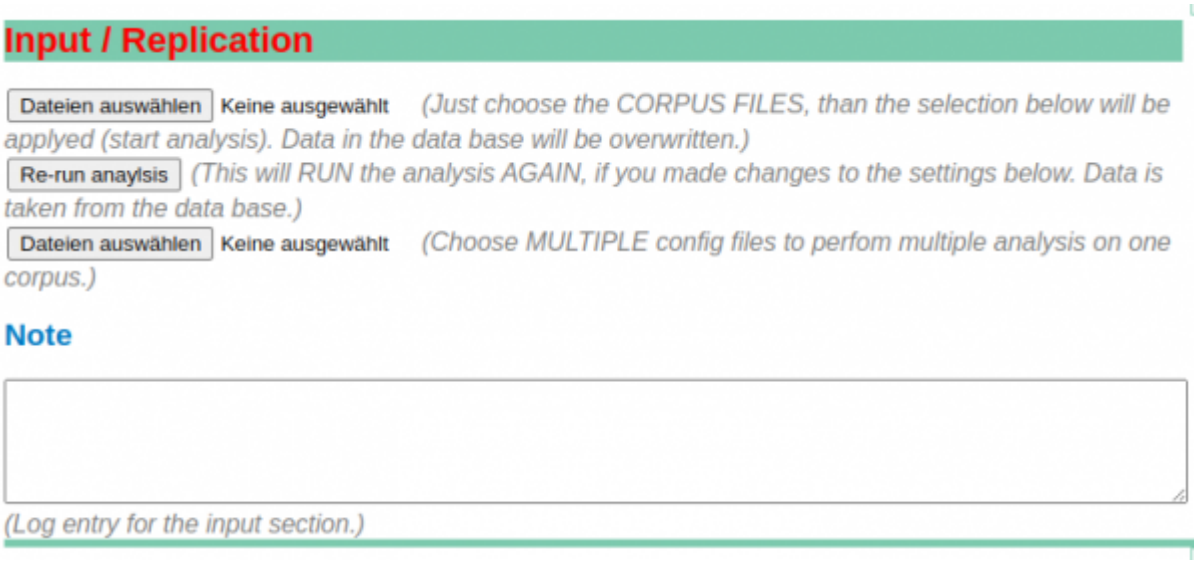
Organisationsmuster für Experimente. Das Datum wird beim ersten Programmlauf automatisch ausgefüllt. Der einmal erstellte Name charakterisiert den Programmablauf und die Einstellungen genau. Änderungen in der Benennung führt auf ein neues Experiment, welches auch separat gespeichert wird. Die Benennung geht in die Config-Datei ein und ein erneuter Aufruf einer Config-Datei stellt alles unter dieser Bezeichnung wieder her. Auf diese Weise können mehrer Konfigurationen unterschieden werden.

2.2 Arbeitsschritt: Configurations-Datei / Datenbank



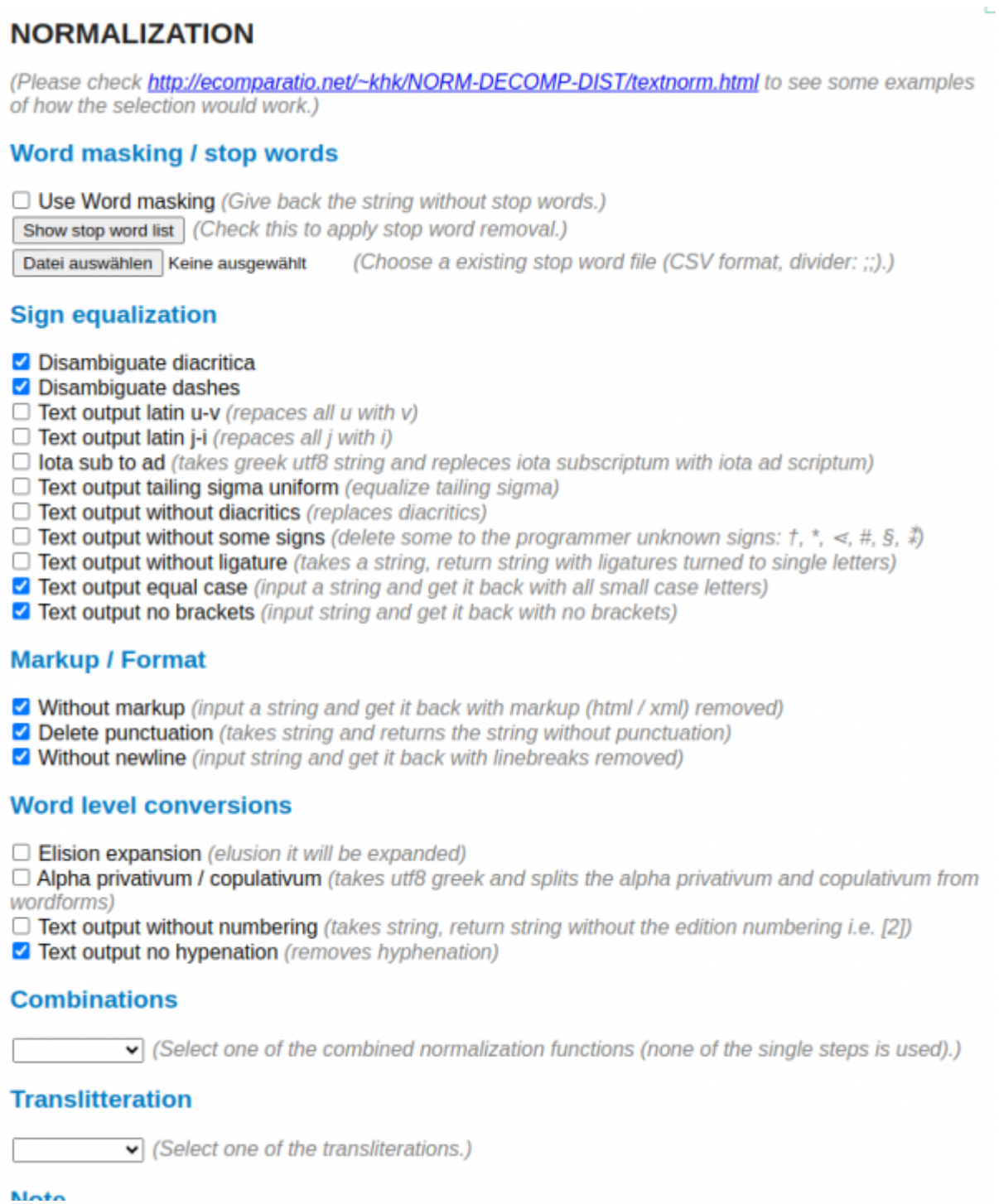
Im Abschnitt „Configuration“ können alle Einstellungen vorgenommen werden, die sich auf die Config-Datei beziehen. stylo-ah-online erlaubt die Speicherung der Konfiguration und genauso ihren erneuten Aufruf durch Eingabe einer Config-Datei. Für die Erstellung von Serien von Experimenten erlaubt es stylo-ah-online mehrer Config-Dateien zuerstellen und als Serie wieder zu öffnen. Die Ergebnisse werden für jede so gespeicherte Einstellung unter einem extra Namen abgespeichert. Für umfangreichere Corpora ist es nötig die Darstellung von Zwischenergebnissen zu beschränken. Das kann an dieser Stelle mit dem Haken bei „Display size of results“ erzielt werden. stylo-ah-online verfügt über eine Auto-save-Funktion, außerdem werden die eingegebenen Daten gespeichert, um diese bei einem erneuten Analysedruchlauf aus der Browser eigenen Datenbank zu holen. Man kann hier die Einstellungen und die Datenbanken oder beides zurücksetzen.

2.3 Arbeitsschritt: Input / Replication

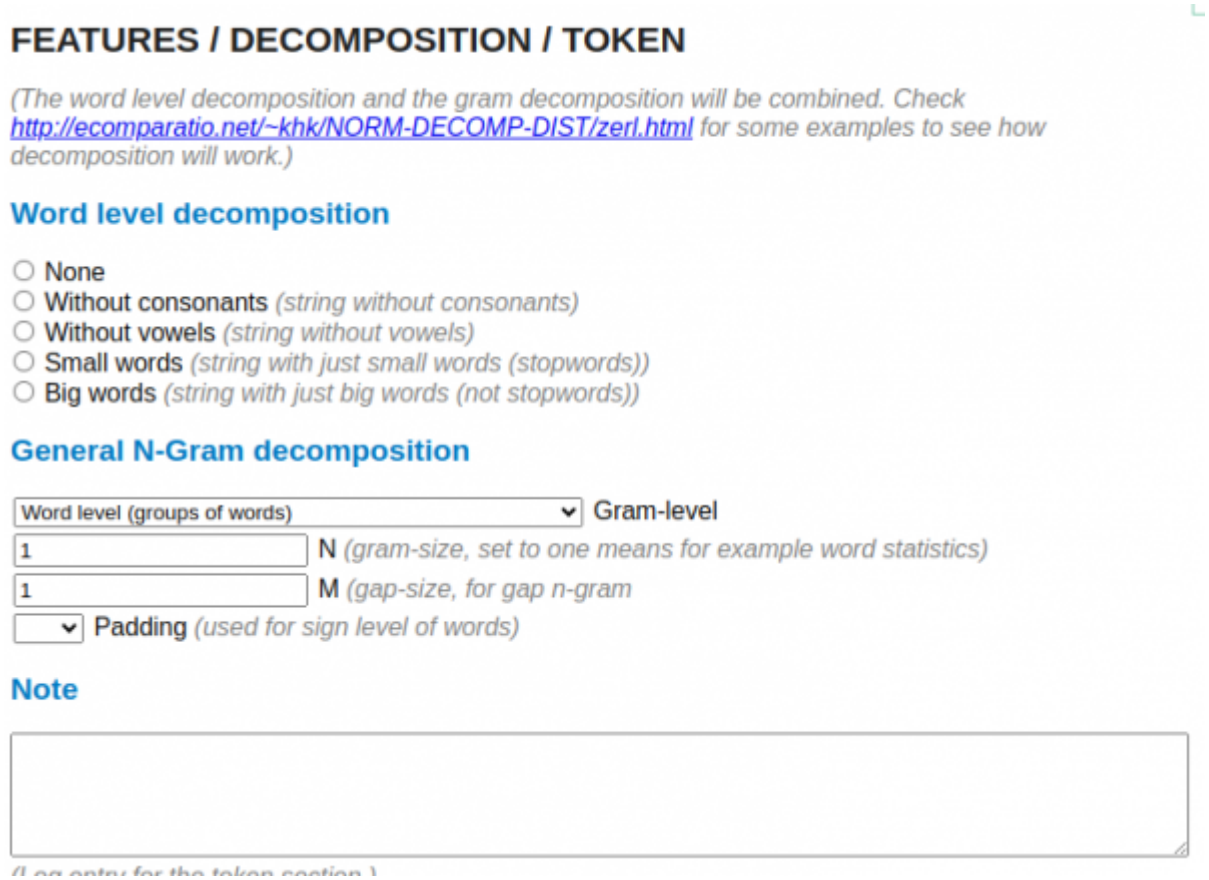


Im Abschnitt „input / Replication“ geht es um den Aufruf der Textdateien, die analysiert werden sollen. Diese Handlung löst die erste Analyse aus und speichert Ergebnisse in der Datenbank. Mit „rerun“ kann man eine veränderte Konfiguration auf die vorhandenen Daten anwenden. Sollten Serienexperimente vorbereitet wurden sein, dann kann man die dazugehörigen Config-Files hier eingeben und so die Serienverarbeitung auslösen. Dazu werden die Daten aus der Datenbank verwendet. Der Arbeitsschritt der Dateieingabe muss der Letze sein, nachdem alle anderen Konfigurationen vorgenommen wurden.

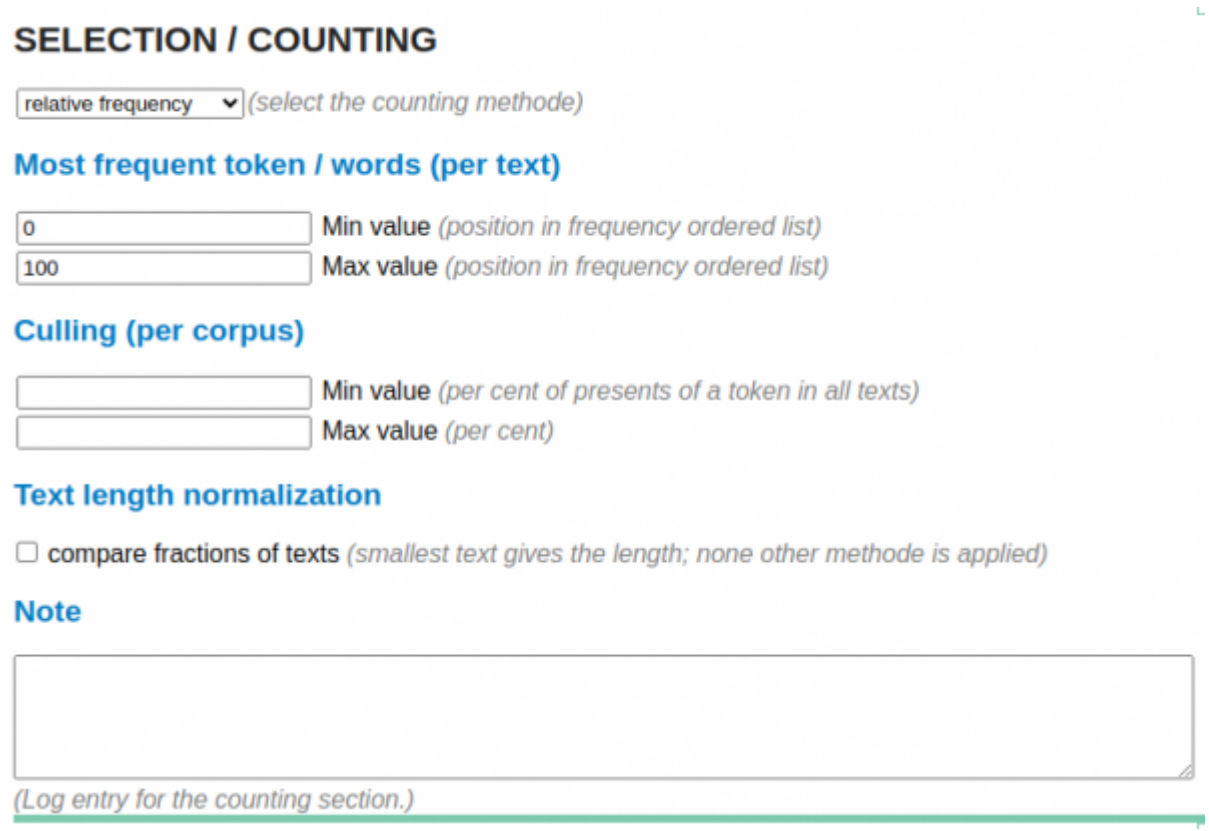
2.4 Arbeitsschritt: Normalisierung (Normalization)



2.5 Arbeitsschritt: Zerlegung (features / decomposition / token)



2.6 Arbeitsschritt: Zählung (Selection / Counting)



Arbeitsschritt: Maßanwendung

Arbeitsschritt: Gruppierung

Arbeitsschritt: Export

3. Corpus-Auswahl

1. Aufruf von

4. Erneute Berechnung und Serien