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

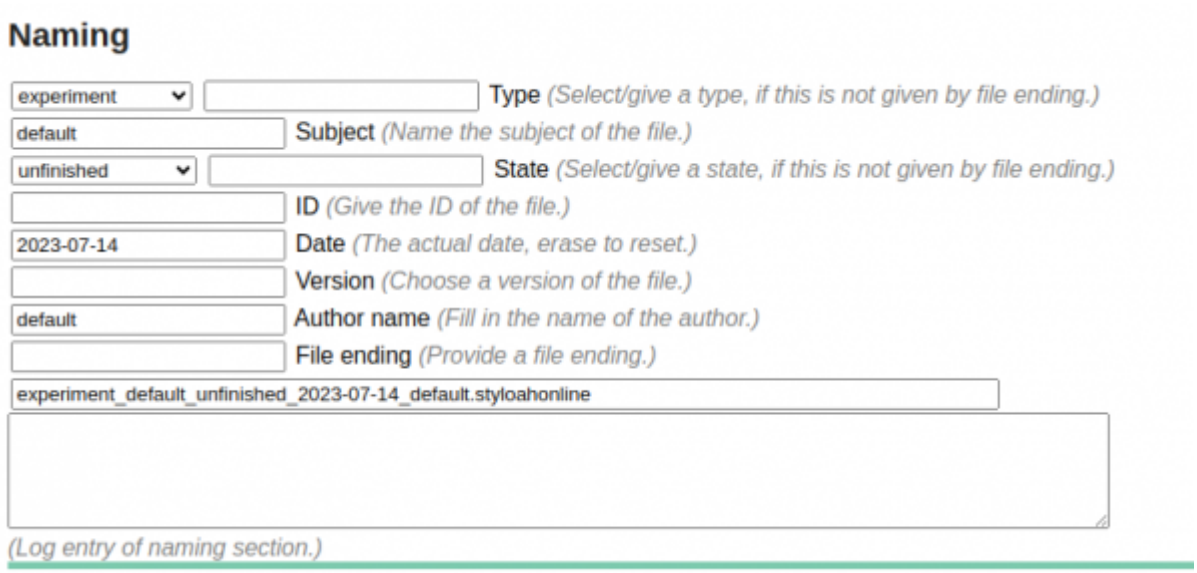
Speicherort: Legen sie einen Ordner an in welchem die Ergebnisse von stylo-ah-online gespeichert werden können. Konfigurieren Sie ihren Browser so, dass er die Dowloads in diesem Ordner ablegt. die Anleitung für Firefox:
https://support.mozilla.org/de/kb/suchen-und-verwalten-heruntergeladener-dateien#w_ziel-ordner-der-heruntergeladenen-dateien-andern

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

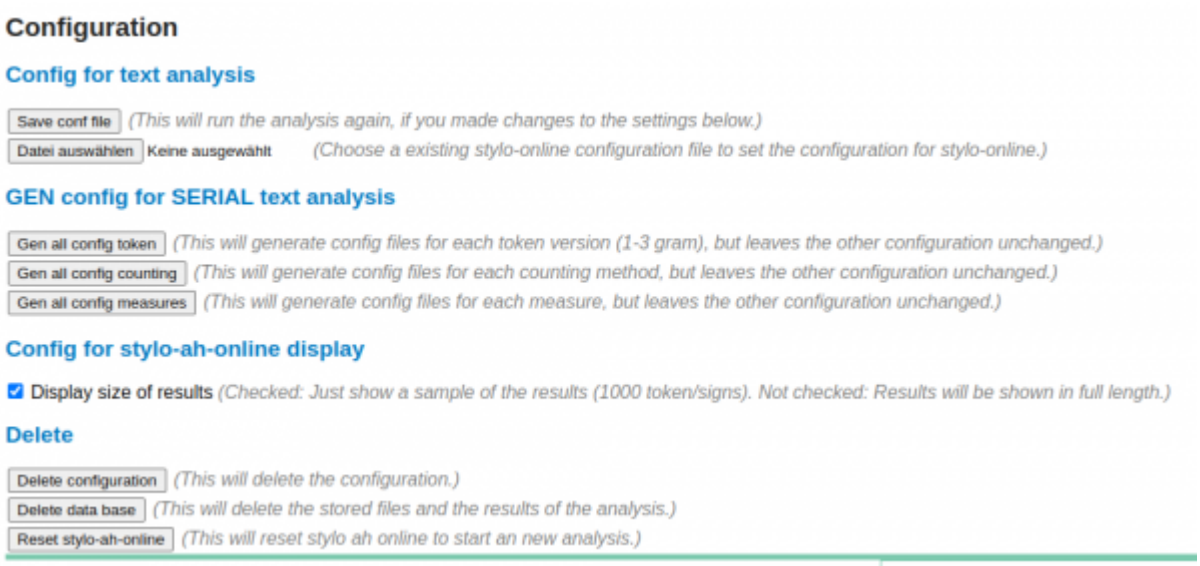
2.1 Arbeitsschritt: Benennung



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

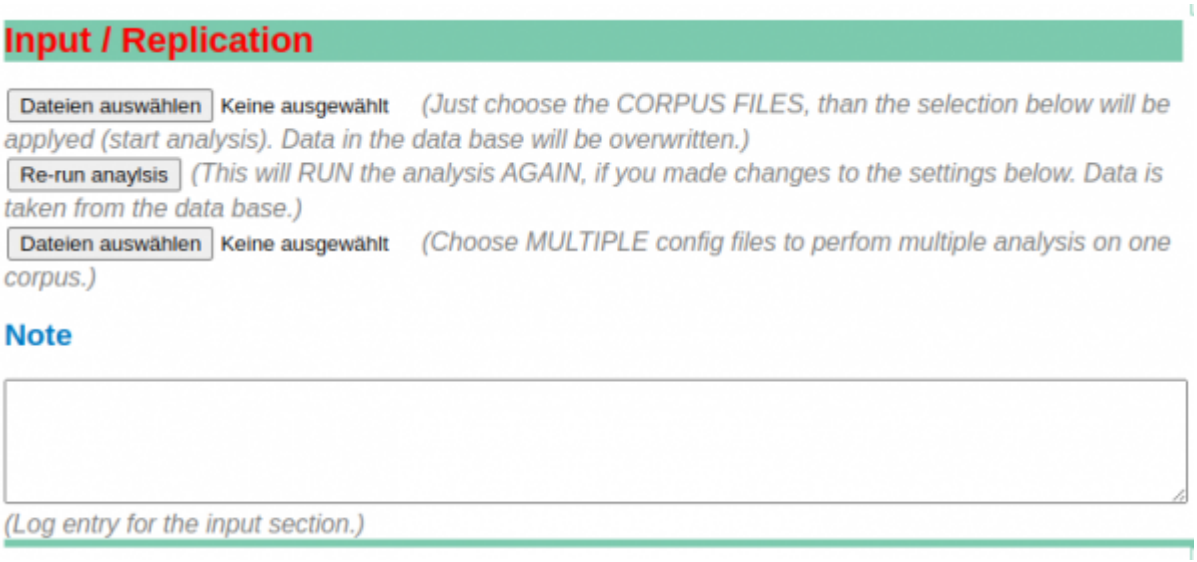
Organisationsmuster für Experimente. Das Datum wird beim ersten Programmlauf automatisch ausgefüllt. Der einmal erstellte Name charakterisiert den Programmablauf und die Einstellungen genau. Änderungen in der Benennung führt auf ein neues Experiment, welches auch separat gespeichert wird. Die Benennung 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



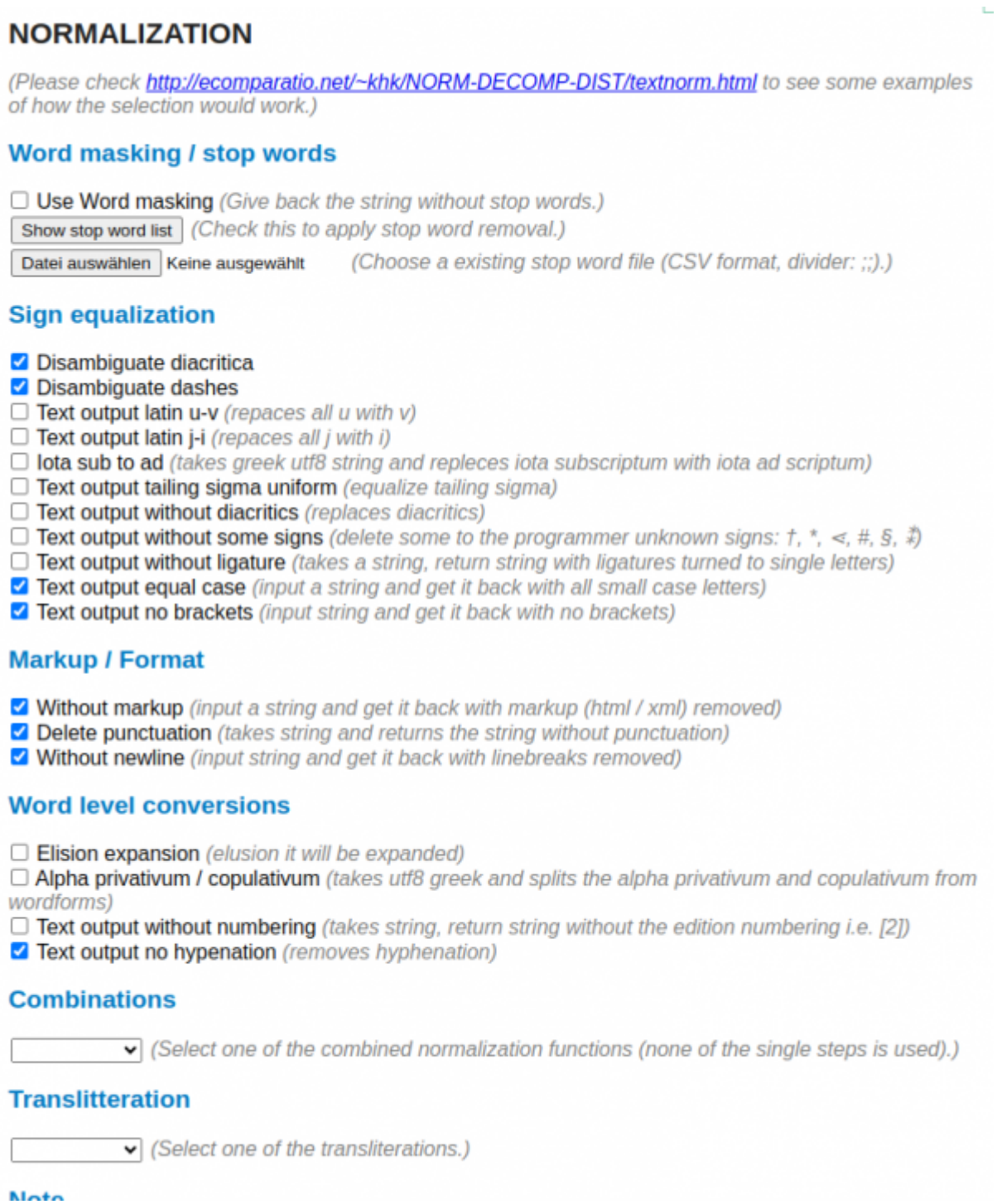
Im Abschnitt „Configuration“ können alle Einstellungen vorgenommen werden, die sich auf die Config-Datei beziehen. stylo-ah-online erlaubt die Speicherung der Konfiguration und genauso ihren erneuten Aufruf durch Eingabe einer Config-Datei. Für die Erstellung von Serien von Experimenten erlaubt es stylo-ah-online 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)



2.4.1 Word masking / stop words

Setzt man den Haken in diesem abschnitt, dann werden die Wortformen auf der Stop-Wortliste aus den Strings entfernt. Den Vorgang nennt man ebenfalls Maskierung von Wortformen. Mann kann sich mit dem Button die aktuelle Stop-Wortliste anzeigen lassen. Eine andere Stop-Wortliste kann man durch die Eingabe einer Datei, die jedes Stop-Wort durch „;;;“ vom nächsten getrennt enthält, nutzen.

2.4.2 Sign equalization

In diesem Abschnitt der Einstellungen geht es um die Behandlung einzelner Zeichen und

Zeichengruppen. Die ersten beiden Haken dienen der Vereinheitlichung von diakritischen Zeichen und Strichen. Dadurch wird die unterschiedliche Kodierung visuell gleicher Zeichen bewirkt. Die weiteren Normalisierungen erklären sich (Ersetzung aller u und v durch v, Ersetzung aller i und j durch i, Vereinheitlichung von Schreibweisen Iota und Sigma, Auflösung von Ligaturen, die Lösung einiger Zeichen).

2.4.3 Markup / Format

Vornehmlich geht es um die Lösung zusätzlicher Formatierungsangaben und der Metadaten/Struktur. Sollte die Eingabe in wohlgeformtem XML geschehen, dann löscht der erste Haken dieses unter Verwendung eines XML Parsers. Sollte dem nicht so sein, dann werden die XML Tags mittels eines regulären Ausdrucks gelöscht. Hier können zudem Interpunktion und Zeilenumbrüche gelöscht werden.

2.4.4 Word level conversions

Unter die Veränderungen auf Wortform ebene zählt alles, was die Wortform als logische, organisatorische Einheit berücksichtigt. Die betrifft Wortformtrennungen, Nummerierungen, Elisionen und Alpha privativum.

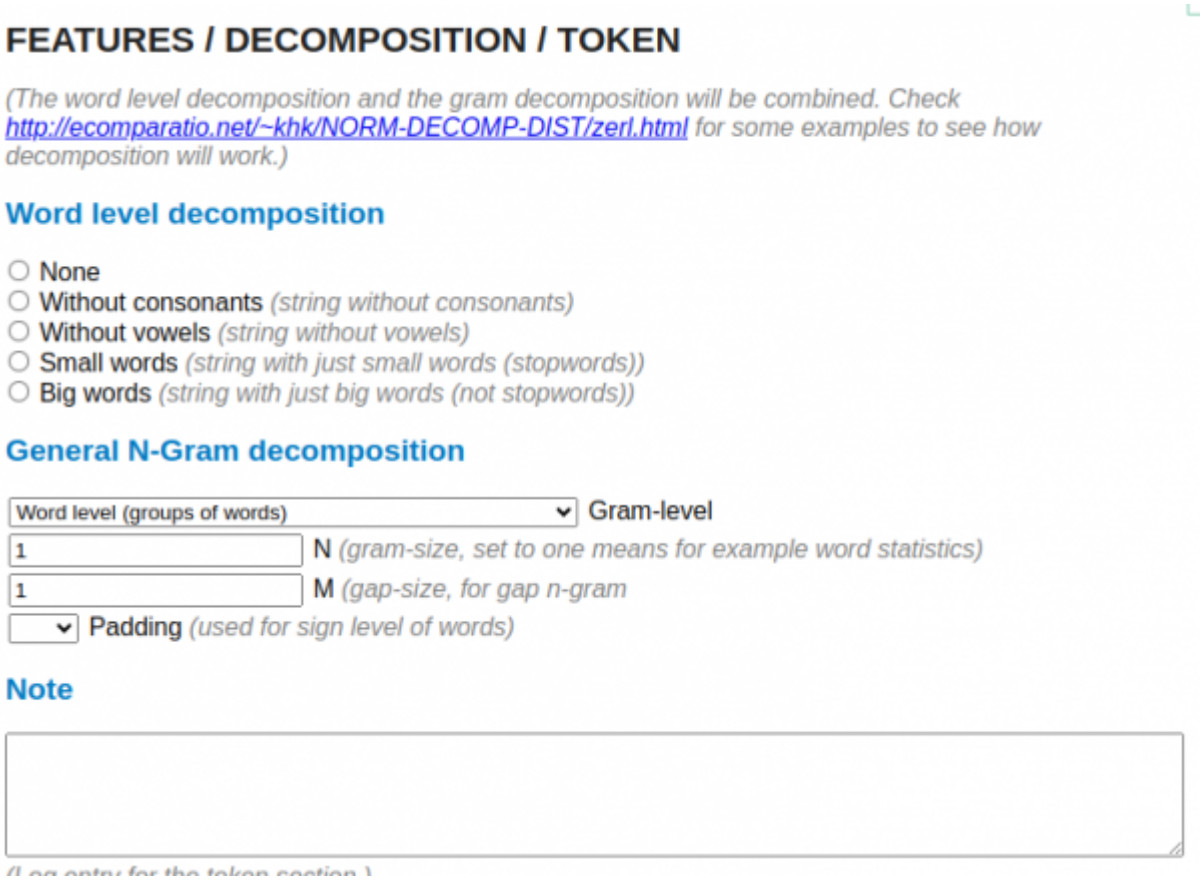
2.4.5 Combinations

Hier können Kombinationen von Normalisierungsschritten gewählt werden. Andere Einstellungen werden dann ignoriert.

2.4.6 Translitteration

Sollte es nötig sein Texte ihrem Zeichenbestand nach zu vereinheitlichen, dann kann diese durch die Transliteration (Griechisch / Latein) geschehen.

2.5 Arbeitsschritt: Zerlegung (features / decomposition / token)



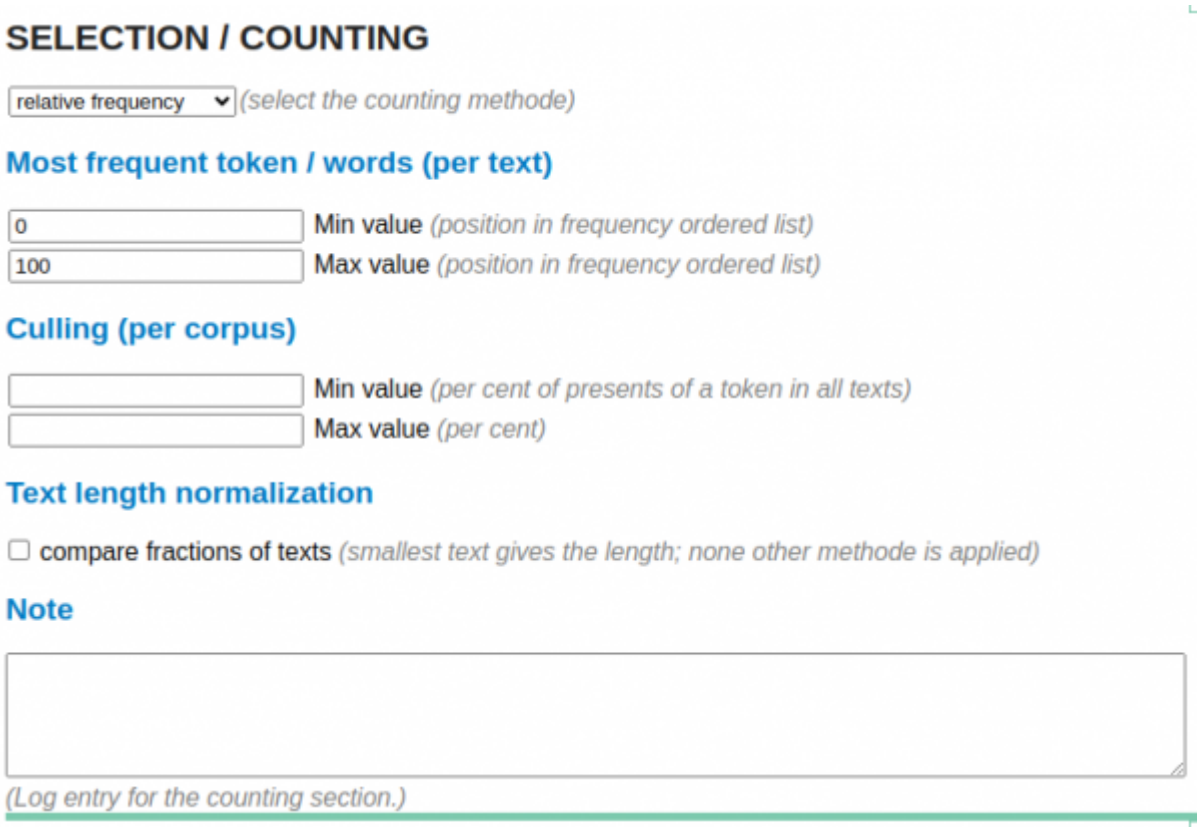
2.5.1 Word level decomposition

Die Einstellungen haben detzt die Vorstellung von Wortformen um, aus denen ein Text besteht. Jede Auswahl tut etwas mit der Wortform. Es kann nur ein Verfahren gewählt werden. Sollen lediglich Wortformen als Token behandelt werden, dann muss dies über die n-gram Einstellung geschehen.

2.5.2 General N-Gram decomposition

Mittels dieser Einstellung kann der String in n-gram Aufteilungen zerlegt werden. Dabei gibt die Auswahl im Pulldown-Menü die Zerlegungsebene an. Die zusätzlichen Zahlen müssen ausgefüllt werden, um das „n“ der Zerlegung oder der Lücken angeben zu können. Für Skip-gram Zerlegungen spiel nicht nur das „n“ (Länge) der Teilung, sondern auch die Länge der Auslassung eine Rolle. Mit der Auswahl „Word level“ zerlegt man den String in „n“ lange Gruppen von Wortformen, die dann als Token ausgezählt werden. Mit der Einstellung „sign level of words“ zerlegt man die Zeichen der Wortformen in n-gram Aufteilungen, durch die Auffüllung können Wortenungen und Anfäng kodiert werden. Die Wortübergänge werden nicht mit Kodiert. Mit der Auswahl „signs of whole string“ wird die n-gram Aufteilung kontinuierlich auf dem String vorgenommen. Dabei werden Wortformübergänge (ab n = 3) berücksichtigt. „Gab-ngram“ ist die Skip-gram Implementierung, hier muss man zusätzlich die Länge der Lücke angeben. Für lateinische und griechische Texte steht die zerlegung in silben zur Verfügung. Abschließend gibt es noch die Zerlegung der Wortformen in drei Abschnitte. Die erste dieser Einstellungen tut diese Aufteilung zu gleichen Teilen, die zweite teilt jede Wortform so auf dass alle Partitionen des Strings entstehen.

2.6 Arbeitsschritt: Zählung (Selection / Counting)



Im oberen Pulldown-Menü des Abschnitts kann man sich für die Zählweise entscheiden, die auf die Menge aller Token (Wortformen) angewendet werden soll. Die absolute Häufigkeit, ist die Anzahl des Auftretens eines Token. Die relative Häufigkeit ist die Auftretenszahl eines Token geteilt durch die Textlänge. Das schwächt den Einfluss der Textlänge ein. Allerdings überdeckt dies das Problem im Korpus sehr kurzer Texte. Die Kodierung des Auftretens in einem Text oder Nichtauftretens kann mittels 0 und 1 geschehen. Dann wird lediglich die Existenz von Token untersucht. Abschließend stehen zwei Zählungen bereit, die eine Beziehung zum Gesamtkorpus herstellen. Die TF-IDF stellt den Bezug zwischen relativer Häufigkeit in einem Text eines Token zur Häufigkeit in Texten vertreten zu sein her. Der Quotient aus relativer Häufigkeit pro Text und relativer Häufigkeit im Korpus mindert mildert die Funktion der relativen Häufigkeit ab und sollte Werte ergeben, die sich zwischen der absoluten Häufigkeit und der relativen Häufigkeit bewegen.

2.6.1 Most frequent token / words (per text)

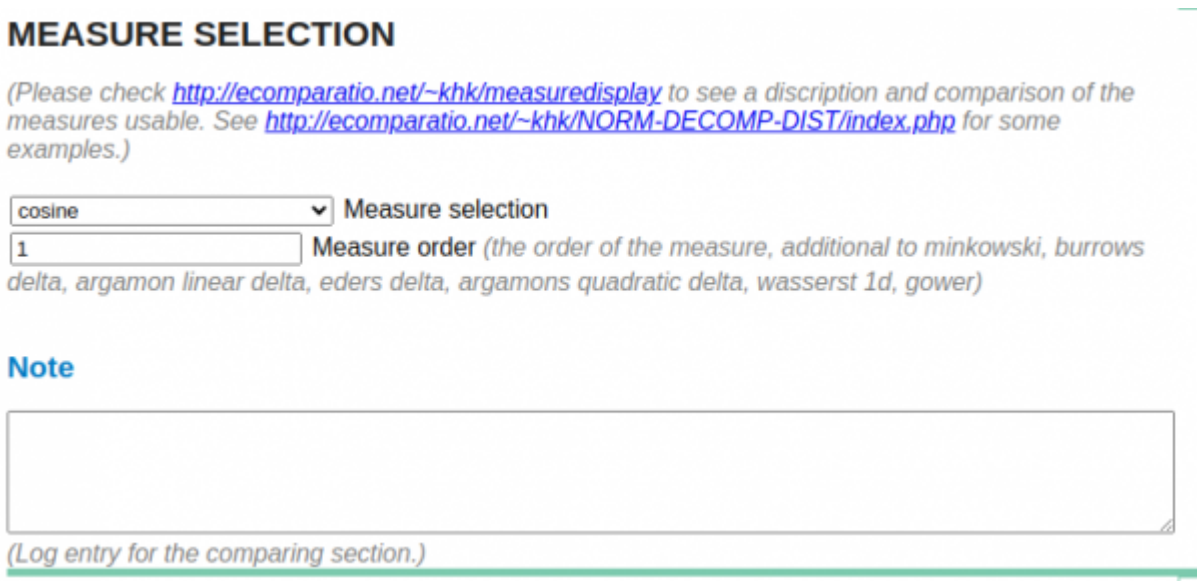
Mit den beiden Zahlenangaben kann der Frequenzrang angegeben werden, der im Profil Berücksichtigung finden soll. Sollen lediglich die Werte der 100 häufigsten Token Betrachtung finden, dann muss man den niedrigsten Rang 0, als minimal zu berücksichtigenden Rang, angeben und den Wert 100, als höchsten zu berücksichtigenden Rang.

2.6.2 Culling

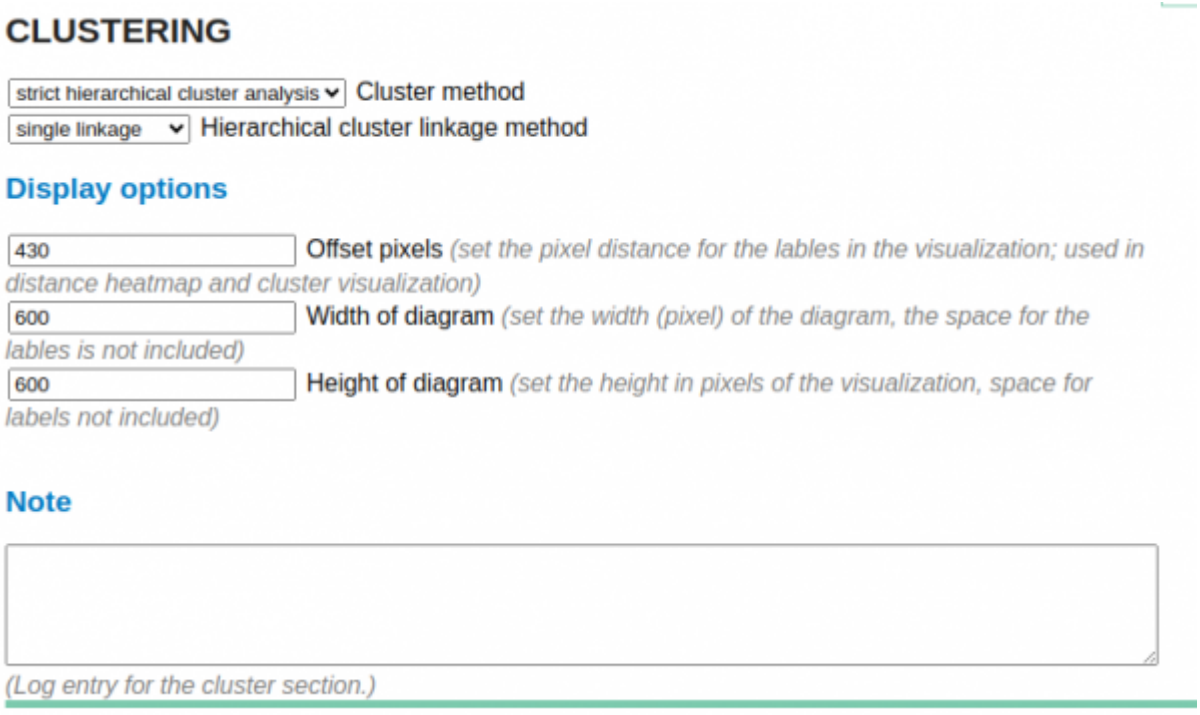
Mit den beiden Werten kann man angeben in wie viel Prozent der Texte ein Token mindestens auftreten soll und maximal auftreten soll, um im Profil Berücksichtigung zu finden. Im Kontrast zu Most frequent words oder TF-IDF kann man dabei nicht absehen, wie die Profile beeinflusst werden. Es

kann durchaus sein, dass damit null-besetzte Bereiche vermindert werden, allerdings kann auch das Gegenteil eintreten. Hier sollte man die Profile genau untersuchen.

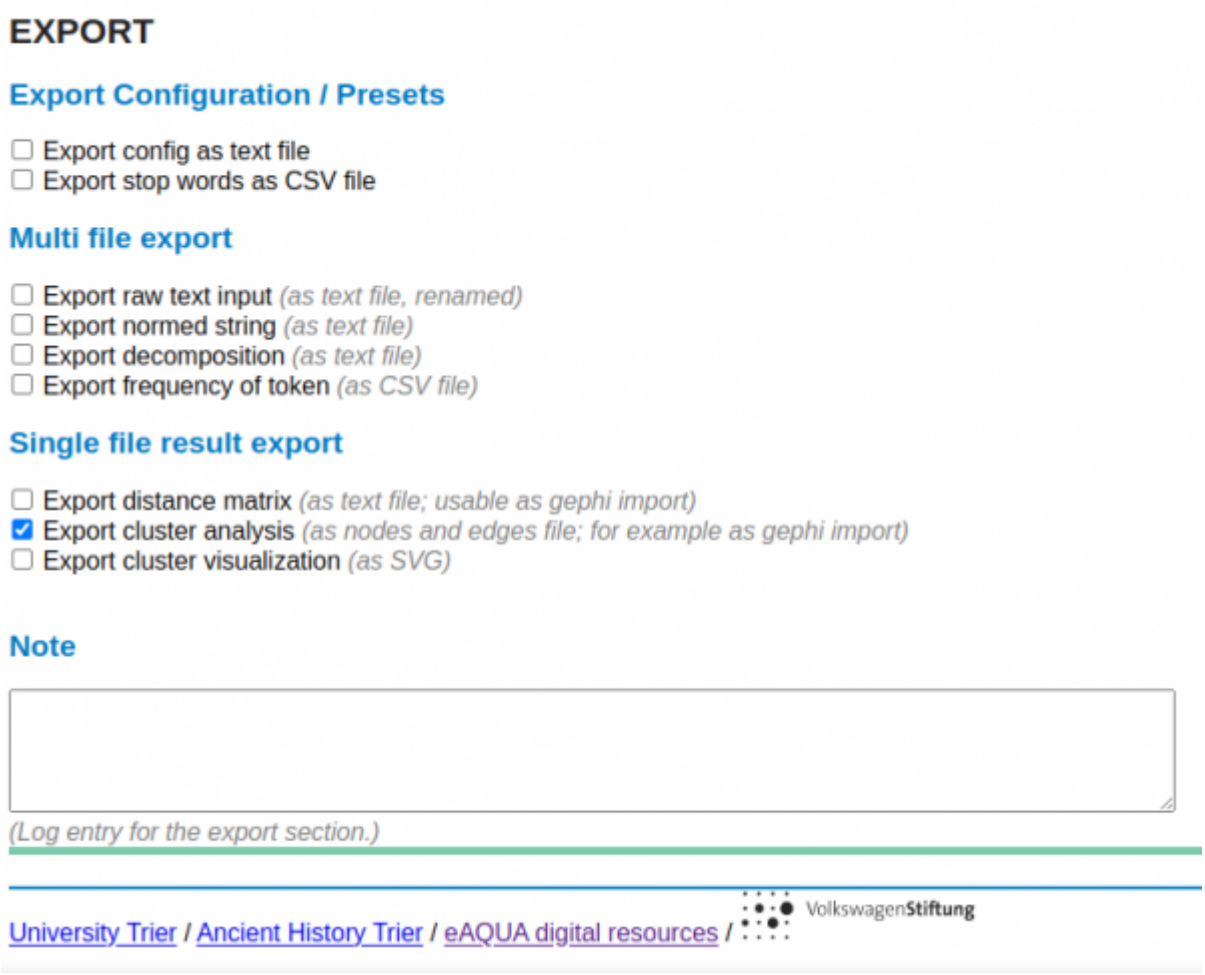
2.7 Arbeitsschritt: Maßanwendung (Measure selection)



2.8 Arbeitsschritt: Gruppierung (Clustering)



2.9 Arbeitsschritt: Export



im Abschnitt „Export“ können, durch das Setzen von Haken, verschiedene Dateien exportiert werden. Diese werden bei jedem Programmdurchlauf geschrieben. Im Abschnitt wird nach drei Typen von Export unterschieden. Der Export grundsätzlicher Dateien, wie der Config-Datei, dem Export von Zwischenergebnissen (die dann auch für jeden eingegebenen Text existieren) und dem Export von Ergebnissen, die für das ganze Corpus gelten. Die Dateien werden im angewählten Download Ordner gespeichert.

3. Corpus-Auswahl

Im Vergleich zu stylo ist der Vorgang anders, wenn es um die Auswahl von Corpora geht: In stylo genügt es einen Ordner anzugeben, welcher den Ordner „corpus“ enthält. Für die Eingabe in stylo-ah-online muss die gesamte Liste der Texte ausgewählt werden, die das Corpus bilden.

4. Erneute Berechnung und Serien

stylo-ah-online speichert das Corpus, welches aktuell bearbeitet wird. Es speichert auch Zwsichenergebnisse. Werden Änderungen der Konfiguration vorgenommen, dann müssen nicht unbedingt alle Arbeitsschritte neu ausgeführt werden. Verwenden Sie den „Rerun“ Button, um ein neue Konfiguration auf das gespeicherte Corpus anzuwenden. Man profitiert damit von Vorberechnungen.

Sollen mehrere verschiedene Konfigurationen ausgeführt und die Ergebnisse exportiert werden, dann legen sie mehrer Konfig-Dateien an und öffnen diese als Datei-Liste. Jede Konfiguration wird dann auf

die gespeicherten Daten des Corpus angewendet. Auch hier kann man von der Wiederverwendung von unveränderten Zwischenergebnissen profitieren.