

Inhaltsverzeichnis

| | |
|---|----|
| stylo-ah-online Handbuch | 2 |
| <i>Ursprung: stylo-ah-online - stylo ohne Installation</i> | 2 |
| <i>Links zum Thema</i> | 2 |
| <i>Implementierte Analyse-Pipeline(s)</i> | 2 |
| <i>Funktionsüberblick</i> | 2 |
| <i>Benutzung</i> | 3 |
| GUI Konzept | 3 |
| Überblick zur Benutzung | 4 |
| 1. Konfiguration des Browsers | 4 |
| 2. Konfigurieren der Analyseschritte | 4 |
| 2.1 Arbeitsschritt: Benennung | 4 |
| 2.1.1 Auswahl und Anzahl der Zeichen | 5 |
| 2.1.2 Mögliche Bestandteile im Dateinamen | 5 |
| Typ des Dokumentes: | 5 |
| Subjekt: | 5 |
| Status: | 5 |
| ID: | 6 |
| Datum: | 6 |
| Versionsnummer: | 6 |
| Autorname: | 6 |
| Dateiendung: | 6 |
| 2.2 Arbeitsschritt: Configurations-Datei / Datenbank | 6 |
| Input eines Config files | 7 |
| Mehrere Config files | 7 |
| Serien von config files | 7 |
| Reset des Tools | 8 |
| 2.3 Arbeitsschritt: Input / Replication | 8 |
| 2.4 Arbeitsschritt: Normalisierung (Normalization) | 9 |
| 2.4.1 Word masking / stop words | 9 |
| 2.4.2 Sign equalization | 9 |
| 2.4.3 Markup / Format | 10 |
| 2.4.4 Word level conversions | 10 |
| 2.4.5 Combinations | 10 |
| 2.4.6 Transliteration | 10 |
| 2.5 Arbeitsschritt: Zerlegung (features / decomposition / token) | 10 |
| 2.5.1 Word level decomposition | 11 |
| 2.5.2 General N-Gram decomposition | 11 |
| 2.6 Arbeitsschritt: Zählung (Selection / Counting) | 11 |
| 2.6.1 Most frequent token / words (per text) | 12 |
| 2.6.2 Culling | 12 |
| 2.6.3 Textlängen Normalisierung | 12 |
| 2.7 Arbeitsschritt: Maßanwendung (Measure selection) | 12 |
| 2.8 Arbeitsschritt: Gruppierung (Clustering) | 13 |
| 2.9 Arbeitsschritt: Export | 14 |
| 3. Erneute Berechnung und Serien | 14 |

stylo-ah-online Handbuch

Ursprung: stylo-ah-online - stylo ohne Installation

Im Rahmen unserer Lehr- und 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 und erscheint langsam.

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

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 Schritte 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, N-Gramme, Zeichen)
4. Zählung / Vektorbildung (01-Kodierung, absolute und relative 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 (Stand 14.11.2023):

Normalisierung:

[Normalform](#) [Wortmaskierung \(Stoppwörter\)](#) [Vereinheitlichen bestimmter Zeichen](#) [UV-Angleich](#) [Jl-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-Gramme Zerlegung in Buchstaben-N-Gramme der Wortformen Zerlegung in Wort-N-Gramme Zerlegung in n-Gramme 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, manhattan, canberra, soerensen, gower, soergel, lorentzian, intersection, wavehedges, motyka, ruzicka, tanimoto, innerproduct, harmonicmean, cosine, kumar, hasse brook, dice, fidelity, bhattacharya 1, bhattacharya 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 Clustering, multidimensional scaling (MDS), tSNE (t-distributed stochastic neighbor 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, ist zusätzlich eine kursive Erklärung angefügt.

Überblick zur Benutzung

1. Konfiguration des Browsers
2. Konfigurieren der Analyseschritte
3. 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 Downloads 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

Naming

| | | |
|--|--|---|
| experiment | | Type (Select/give a type, if this is not given by file ending.) |
| default | | Subject (Name the subject of the file.) |
| unfinished | | State (Select/give a state, if this is not given by file ending.) |
| | | ID (Give the ID of the file.) |
| 2023-07-14 | | Date (The actual date, erase to reset.) |
| | | Version (Choose a version of the file.) |
| default | | Author name (Fill in the name of the author.) |
| | | File ending (Provide a file ending.) |
| experiment_default_unfinished_2023-07-14_default.styloahonline | | |
| <div></div> | | |
| (Log entry of naming section.) | | |

Die Benennung 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 mehrere Konfigurationen unterschieden werden.

2.1.1 Auswahl und Anzahl der Zeichen

- Dateinamen sollten so lang wie nötig und so kurz wie möglich gehalten werden.
- Es werden nur **Kleinbuchstaben [a-z]** und **Zahlen [0-9]** sowie der **Unterstrich [_]** und **Bindestrich [-]** verwendet. Für von der Software automatisch benannte Dateien wird als Ausnahme die Zeichenklasse [A-Z] zugelassen.
- Es dürfen nicht benutzt werden Sonderzeichen, Umlaute, ß.
- Altgriechische Bezeichnungen zu Werken oder Autoren werden ins lateinische transliteriert, falls sie nicht über die entsprechende Corpus-ID kenntlich gemacht werden können. Die Corpora-IDs sind immer den Namen vorzuziehen, weil sie gegebenenfalls einen Verweis auf die benutzte Edition liefern und meist kürzer sind.
- Aus Kompatibilitätsgründen sollte die MAX_PATH Einstellung von 260 bei Windows nicht überschritten werden, das bedeutet, Ordnerstruktur + Dateinamen sollten maximal aus 256 Zeichen bestehen, da die Laufwerksbezeichnung und das unsichtbar endende Nullzeichen mitgezählt werden. Der maximale Pfad auf einem Windows-Laufwerk D wäre demzufolge „D:\some_256-zeichen-path_string<NUL>“. Bei Unicode-Pfaden sind es gegebenenfalls 255 Zeichen oder weniger. Deswegen wäre eine **Obergrenze von 250 Zeichen** für die gesamte Pfadangabe (Dateiname + Ordnerstruktur) festzulegen.

2.1.2 Mögliche Bestandteile im Dateinamen

Typ des Dokumentes:

beschluss / protokoll / rechnung / notiz / skizze etc. Ein Typ ist immer dann anzugeben, wenn er sich nicht zwangsläufig aus der Dateiendung ergibt. Bei den Endungen jpg oder png beispielsweise den Typ „foto“ voranzustellen, ist überflüssig und widerspricht einer Regel aus Punkt 1 (so kurz wie möglich).

Subjekt:

als Subjekt wird jene Entität verstanden, über die eine Aussage getroffen wird. Eine Rechnung über den bestellten Medion-Rechner bei Saturn würde demzufolge als Subjekt saturn und medion bekommen, die Datei mit rechnung_saturn_medion beginnen.

Status:

der Bearbeitungszustand des Dokumentes, wenn es der stetigen Veränderung unterworfen ist. Bei Listen böte sich zum Beispiel voll oder unvoll an, um zu erklären, ob sie komplett sind. Bei Office-Dokumenten könnte auf final oder unfinal zurückgegriffen werden.

ID:

bei automatisch vergebenen Identifikationsnummern ist es mitunter hilfreich, die ID im Namen zu behalten, um die Datei der Originalversion zuordnen zu können. Zum Beispiel wäre es sinnvoll eine Statistik zu Plutarchs *De tranquillitate animi* aus dem TLG mit tlg0007-096 im Dateinamen zu kennzeichnen.

Datum:

Datumsangaben im Dateinamen sind insbesondere dann sinnvoll, wenn ein Dokument unveränderlich zu Archivierung abgelegt werden soll (protokoll_projektmeeting_replikation_2023-04-28.pdf) oder es zu sukzessiven Änderungen kommt, um die Versionierung übersichtlicher zu gestalten (replikation_stylesheed_2023_v01).

Versionsnummer:

Um die Bedeutung als Versionsnummer klar zu machen wird ein kleines v vorangestellt. Die Anzahl der führenden Nullen ist abhängig von der zu erwartbaren Menge.

Auturname:

Auf die Benennung des (letzten) Autoren im Dateinamen sollte verzichtet werden außer eine Dateiversion wird parallel und unabhängig voneinander von mehreren Personen bearbeitet und es entstehen dadurch horizontale (anstatt vertikale) Dateiversionen, die später von Hand zusammengeführt werden müssen.

Dateiendung:

Die Dateiendung bezeichnet den Teil, der nach dem abschließenden Punkt folgt. Er verweist in aller Regel auf den Dateityp, also die Art und Weise, wie die Informationen in der Datei kodiert sind.

2.2 Arbeitsschritt: Configurations-Datei / Datenbank

Configuration

Config for text analysis

(This will run the analysis again, if you made changes to the settings below.)
 Keine ausgewählt (Choose a existing stylo-online configuration file to set the configuration for stylo-online.)

GEN config for SERIAL text analysis

(This will generate config files for each token version (1-3 gram), but leaves the other configuration unchanged.)
 (This will generate config files for each counting method, but leaves the other configuration unchanged.)
 (This will generate config files for each measure, but leaves the other configuration unchanged.)

Config for stylo-ah-online display

☒ Display size of results (Checked: Just show a sample of the results (1000 token/signs). Not checked: Results will be shown in full length.)

Delete

(This will delete the configuration.)
 (This will delete the stored files and the results of the analysis.)
 (This will reset stylo ah online to start an new analysis.)

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 das Erzeugen von Serien aus Experimenten erlaubt es stylo-ah-online mehrer Config-Dateien zu erstellen und diese 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 Analysedurchlauf aus der Browser eigenen Datenbank zu holen. Die letzten Buttons des Abschnitts erlauben es, die Einstellungen und die Datenbanken oder beides zurücksetzen.

Input eines Config files

Wenn config files über die „Save config file“- oder „Gen all...“-Buttons erzeugt wurden, so können sie mittels dieses Buttons wieder geladen werden. Das Tool stellt die Konfiguration, dann so ein, wie diese im config file angegeben ist. Alte config files sind mit neueren kompatibel. Das Konfiguration wird stets auf das aktuelle Korpus angewendet.

Mehrer Config files

Werden mehrer config files geladen, dann werden diese nacheinander verarbeitet und die Ergebnisse werden, wie angegeben, abgespeichert. Mehrer Konfigurationen dieser Art werden auf dem aktuellen Korpus ausgeführt.

Serien von config files

Wenn man die „Gen all...“-Buttons verwendet, dann werden alle Einstellungen die zur Zeit gemacht wurden, übernommen und lediglich die Einstellungen eines Arbeitsschritts werden variiert. Auf diese Weise kann man zu diesem Arbeitsschritt eine Serie von config files erstellen und somit eine Serie von Ergebnissen. Diese Funktion dient der systematischen Untersuchung von Arbeitsschritten in Bezug

auf ein gegebenes Corpus. Die „Gen all“-Funktion steht für die Arbeitsschritte Token-Bildung, Token-Zählung und Anwendung der Maße zur Verfügung.

Reset des Tools

Sowohl das Tool, als auch der Browser betreiben ein Caching. Caching bezeichnet die Speicherung von Einstellungen, Daten und Zuständen. Wenn der Button „Delet configuration“ benutzt wird, dann werden alle Einstellungen gelöscht und es werden auch keine default-Einstellungen vorgenommen. Im Anschluss können alle Einstellungen neu vorgenommen werden. Wenn der Button „Delete data base“ benutzt wird, so bleiben die Einstellungen erhalten, allerdings werden alle Eingabedaten und alle Zwischenergebnisse gelöscht. Die Datenbank der Eingabedaten und Zwischenergebnisse wird nicht automatisch gelöscht. Wenn man den Button „Reset stylo-ah-online“ auswählt, so werden die Einstellungen und die Datenbanken gelöscht. Will man die Seite und somit den Zustand (Programmversion) des Tools nicht aus dem Browser-Cach laden, sondern den neusten Code vom Server, dann kann man den Button „Refresh page/tool“ verwenden.

2.3 Arbeitsschritt: Input / Replication

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

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.

Mit „Select some data from database“ können bereits geladene Dateien aus der Datenbank

ausgewählt und so kann ein neues Corpus zusammengestellt werden.

Zudem wird nun ein Histogramm der absoluten Häufigkeiten der Zeichen der Eingabe angezeigt.

2.4 Arbeitsschritt: Normalisierung (Normalization)

NORMALIZATION

(Please check <http://ecomparato.net/~khk/NORM-DECOMP-DIST/textnorm.html> to see some examples of how the selection would work.)

Word masking / stop words

- ☐ None
- ☐ Use Word masking (Give back the string without stop words.)
- ☐ Use positiv stop word list (Give back the string of only stop words.)
- ☐ Show stop word list (Check this to apply stop word removal.)
- Datei auswählen Keine ausgewählt (Choose a existing stop word file (CSV format, divider: ;))

Sign equalization

- ☒ Disambiguate diacritics
- ☒ Disambiguate dashes
- ☐ Text output latin u-v (replace all v with u)
- ☐ Text output latin i-j (replace all j with i)
- ☐ Iota sub to ad (takes greek iota string and replaces join subscriptum with join ad scriptum)
- ☒ Text output taling sigma uniform (equalize taling sigma)
- ☒ Text output without diacritics (replace diacritics)
- ☒ Without modern diacritics (replace diacritics of modern languages: performance issues!)
- ☒ Text output without some signs (delete some to the programmer unknown signs: f, *, <, >, &, &)
- ☒ Text output without ligature (takes a string, return string with ligatures turned to single letters)
- ☒ Text output equal case (input a string and get it back with all small case letters)
- ☒ Text output no brackets (input string and get it back with no brackets)

Markup / Format

- ☒ Without markup (input a string and get it back with markup (html / xml) removed)
- ☒ Delete punctuation (takes string and returns the string without punctuation)
- ☒ Without newline (input string and get it back with linebreaks removed)

Word level conversions

- ☐ Elision expansion (elision it will be expanded)
- ☐ Alpha privatium / copulativum (takes u28 greek and splits the alpha privatium and copulativum from wordforms)
- ☐ Text output without numbering (takes string, return string without the edition numbering i.e. [2])
- ☐ Text output no hyphenation (removes hyphenation)

Combinations

(Select one of the combined normalization functions (none of the single steps is used).)

Text input:

Text output:

Sign count: 12253

Sign count: 38172

Sign count: 34465

Sign count: 14597

Sign count: 103641

Sign count: 127459

Sign count: 51350

Sign count: 621821

Histogram of character frequencies:

| Character | Frequency |
|-----------|-----------|
| α | 532 |
| β | 1966 |
| γ | 1034 |
| δ | 10179 |
| ε | 8870 |
| ζ | 13917 |
| η | 3592 |
| θ | 13114 |
| ι | 17594 |
| κ | 2221 |
| λ | 16536 |
| μ | 17911 |
| ν | 17056 |
| ξ | 1795 |
| ο | 3402 |
| π | 18958 |
| ρ | 13006 |
| σ | 13006 |
| τ | 13006 |
| υ | 13006 |
| φ | 13006 |
| χ | 13006 |
| ψ | 13006 |
| ω | 13006 |

Auf der rechten Seite werden die Zwischenergebnisse des Normalisierungsschritts angezeigt. Man kann somit Einsicht in die Güte der Normalisierung nehmen und gegebenen falls Änderungen vornehmen. Unter dem Anzeigetext befindet sich die Liste der Eingabetexte, die als Menü funktioniert. Es wird außerdem ein Histogramm der absoluten Häufigkeiten der Zeichen in den normalisierten Texten angezeigt.

2.4.1 Word masking / stop words

Setzt man den Haken in diesem Abschnitt, dann werden die Wortformen auf der Stopp-Wortliste aus den Strings entfernt. Den Vorgang nennt man ebenfalls Maskierung von Wortformen. Man kann sich mit dem Button die aktuelle Stopp-Wortliste anzeigen lassen. Genauso ist es möglich, eine andere Stopp-Wortliste durch die Eingabe einer Datei, die jedes Stopp-Wort durch „;“ vom nächsten getrennt enthält, zu 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 vereinheitlicht. 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).

2.5.1 Word level decomposition

Die Einstellungen setzen 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-Gramm-Einstellung geschehen.

2.5.2 General N-Gram decomposition

Mittels dieser Einstellung kann der String in N-Gramme (Aufteilungen der Länge N, „N“ und „n“ sind Synonyme und stehen für eine gewählte natürliche Zahl) 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-Gramme spielt 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“ teilt man die Zeichen der Wortformen auf. Durch die Auffüllung können Wortendungen und -anfänge kodiert werden. Die Wortübergänge werden nicht kodiert. Mit der Auswahl „signs of whole string“ wird die Aufteilung kontinuierlich auf dem String vorgenommen. Dabei werden Wortformübergänge (ab $n = 3$) berücksichtigt. „Gap-ngram“ ist die Skip-Gramm-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 teilt zu gleichen Teilen auf, die zweite teilt jede Wortform so auf, dass alle Partitionen des Strings entstehen.

2.6 Arbeitsschritt: Zählung (Selection / Counting)

SELECTION / COUNTING

relative frequency (select the counting method)

Most frequent token / words

☐ MFV per text (get the most frequent word from each input text)

0 Min value (position in frequency ordered list)

1000 Max value (position in frequency ordered list)

Culling (per corpus)

☐ Min value (per cent of presents of a token in all texts)

☐ Max value (per cent)

Text length normalization

☐ Compare fractions of texts (smallest text gives the length; none other method is applied, only on first run on corpus)

☐ Fraction length (if a number is provided, then every text will be split into fractions of that size on sign level)

☐ Count of fractions (if a number is provided, then only that number or less text fractions will be used)

Note

(Log entry for the counting section)

0.015274541305428107

1.010559342831274052

2.011910218246768657

3.010524664540646072

4.004173240122929663

5.009356420437458489

6.01280115274702124

7.0054597861177056665

8.007146774222080702

9.0042568657438901035

10.007921919400983244

11.007225575287985732

12.004736104879384165

13.0046412219633044345

14.008401158536487307

15.004374263250238413

16.004795607725077555

17.005261981380433856

18.00651797387865893

19.004626748298138204

20.002354382867040087

21.005874699872470928

22.003498493934730205

23.004666952923599956

24.003681135507277841

25.004395169655478524

26.006308909827457829

27.007468411225774703

0.00332005apologia.replicatio.v03.txt

1.00070969anagallitae.replicatio.v03.txt

2.00070969ara.replicatio.v03.txt

3.00059038aviochus.replicatio.v03.txt

4.00059008politius.replicatio.v03.txt

5.00059004phaedo.replicatio.v03.txt

6.00059002apologia.replicatio.v03.txt

7.000590034reges.replicatio.v03.txt

0.00070969anagallitae.replicatio.v03.txt

1.00070969ara.replicatio.v03.txt

2.00059038aviochus.replicatio.v03.txt

3.00059008politius.replicatio.v03.txt

4.00059004phaedo.replicatio.v03.txt

5.00059002apologia.replicatio.v03.txt

6.000590034reges.replicatio.v03.txt

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 ab. Allerdings überdeckt dies das Problem im Corpus sehr kurze Texte mit sehr langen Texten zu vergleichen. Die Kodierung des Auftretens oder

Nichtauftretens in einem Text, 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 Corpus mindert die Funktion der relativen Häufigkeit ab und sollte Werte ergeben, die sich zwischen der absoluten Häufigkeit und der relativen Häufigkeit bewegen.

Um dem Umstand verschieden langer Texte noch auf eine andere Weise zu begegnen, existiert die Einstellung „Text length normalization“. Setzt man hier den Haken, dann wird jeder Text in Teile der Länge des kürzesten Textes im Corpus zerlegt. Will man diesen Vorgang rückgängig machen, so muss man die Texte erneut laden und den Haken zuvor entfernen. Bei umfangreicheren Corpora und sehr unterschiedlich bemessenen Textlängen ist Vorsicht angebracht. Vergleicht man ein Fragment weniger hundert Wortformen mit einem Buch, so wird das Buch in viele tausend Teile zerlegt. Die Funktion steht nur beim Laden eines Corpus zur Verfügung und kann nicht mit dem „Re-run“ Button ausgelöst werden.

Es wird zusätzlich eine Diagramm angegeben, was die Zählung pro Token und Text als Barplot angibt.

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, und den Wert 100, als höchsten zu berücksichtigenden Rang, angeben.

2.6.2 Culling

Mit den beiden Werten kann man angeben, in wie viel Prozent der Texte ein Token mindestens 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.

2.6.3 Textlängen Normalisierung

Eine Normalisierung, die verschiedene Vergleichs- und Rechnungsprobleme mildert ist die Textlängennormalisierung. Dabei kann auf folgende Weise vorgegangen werden, um immer nur gleich lange Texte zu vergleichen. Für ein Corpus gemischter Textlängen, kann der kürzeste Text die Länge angeben in die alle anderen Texte zerlegt werden. Es kann allerdings auch eine Anzahl von Zeichen als Textlänge angegeben werden. Beim Vergleich sehr unterschiedlich langer Text kann es nötig werden eine Anzahl von Textteilen anzugeben, die überhaupt verglichen werden. Die Textteile stellen eine randomisierte Auswahl aus allen möglichen Teilen dar.

2.7 Arbeitsschritt: Maanwendung (Measure selection)

MEASURE SELECTION

(Please check <http://comparatio.net/~hhk/measuresdisplay> to see a description and comparison of the measures usable. See <http://comparatio.net/~hhk/NORM-DECOMP-DIST/index.php> for some examples.)

Measure selection
 Measure order (the order of the measure, additional to minikowski, burrows delta, argenson linear delta, eders delta, argenson quadratic delta, wasserst.1st, power)

Note

(Log entry for the comparing section.)

hg0059034leges.replicatio.v03.txt verglichen mit:
 hg0032005apologia.replicatio.v03.txt: 0.37230273250486523 (T1vT2 = 50.7665912986695)
 hg0007098tranquillate.replicatio.v03.txt: 0.3081231093718942 (T1vT2 = 16.29077809796271)
 hg0007095ira.replicatio.v03.txt: 0.3383273588652781 (T1vT2 = 18.043089690392596)
 hg0059038basiochus.replicatio.v03.txt: 0.324071104937003 (T1vT2 = 42.6049331963001)
 hg0059008politicus.replicatio.v03.txt: 0.20272365289302127 (T1vT2 = 5.9998526839703)
 hg0059004phaedo.replicatio.v03.txt: 0.3108296538058028 (T1vT2 = 4.878657115733149)
 hg0059002apologia.replicatio.v03.txt: 0.36845161271625515 (T1vT2 = 12.109897172236504)
 hg0059034leges.replicatio.v03.txt: 0 (T1vT2 = 1)

hg0032005apologia.replicatio.v03.txt
 hg0007098tranquillate.replicatio.v03.txt
 hg0007095ira.replicatio.v03.txt
 hg0059038basiochus.replicatio.v03.txt
 hg0059008politicus.replicatio.v03.txt
 hg0059004phaedo.replicatio.v03.txt
 hg0059002apologia.replicatio.v03.txt
 hg0059034leges.replicatio.v03.txt

Aus dem Pulldown-Menü können verschiedene Distanzmaße ausgewählt werden, die zum Vergleich zwischen der Token-Profilen benutzt werden sollen. Die Anwendung der Maße auf die Profil-Vektoren ergibt die Distanzmatrix. Die Distanzmatrix ist Grundlage der Gruppierung der Texte untereinander. Manche Maße erfordern die Angabe eines Zahlenwerts. Was dieser bedeutet, ist von Maß zu Maß verschieden. Man sollte die Dokumentation der Maß, wie sie in stylo-ah-online verlinkt ist, konsultieren.

Es wird außerdem eine Darstellung des Distanzmatrix (Distanzen aller Textpaare) als Heatmap angegeben.

2.8 Arbeitsschritt: Gruppierung (Clustering)

CLUSTERING

Cluster method
 Hierarchical cluster linkage method

Display options

Offset pixels (set the pixel distance for the labels in the visualization; used in distance heatmap and cluster visualization)
 Width of diagram (set the width (pixel) of the diagram, the space for the labels is not included)
 Height of diagram (set the height in pixels of the visualization, space for labels not included)

Note

(Log entry for the cluster section.)

Zur Zeit stehen noch nicht alle Cluster-Verfahren, die **stylo** verwendet zur Verfügung. **stylo** nutzt

vornehmlich die Implementierungen Dritter, um die Funktionalität anzubieten, diesen Vorteil haben wir für die Implementierung von stylo-ah-online nicht genutzt. Zur Zeit steht die hierarchische Clusterung in zwei verschiedenen Darstellungen zur Verfügung. Ebenso die tSNE, von der, als optimale Einbettung, allerdings als Cluster-Methode für den gesamten Gegenstand der Textanalyse abzurufen ist. Auch steht eine alternative Variante des MDS (multi dimensional scaling, Anwendung einer radialen Funktion) zur Verfügung. Die Implementierung wird noch komplettiert.

Wichtig für die graphische Darstellung sind die Offset-Werte (Bereich für Beschriftungen) und die Größenangabe der Grafik. Die Grafiken können als SVG gespeichert werden und in Veröffentlichungen Einsatz finden. Die Werte beziehen sich auf alle erstellten Grafiken.

2.9 Arbeitsschritt: Export

EXPORT

Export Configuration / Presets

- ☐ Export config as text file
- ☐ Export stop words as CSV file

Multi file export


- ☐ Export raw text input (as text file, renamed)
- ☐ Export normed string (as text file)
- ☐ Export decomposition (as text file)
- ☐ Export frequency of token (as CSV file)

Single file result export

- ☐ Export distance matrix (as text file; usable as gephi import)
- ☒ Export cluster analysis (as nodes and edges file; for example as gephi import)
- ☐ Export cluster visualization (as SVG)

Note

(Log entry for the export section.)

[University Trier](#) /
 [Ancient History Trier](#) /
 [eAQUA digital resources](#) /
  VolkswagenStiftung

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: Dem 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. Erneute Berechnung und Serien

stylo-ah-online speichert das Corpus, welches aktuell bearbeitet wird. Es speichert auch

Zwischenergebnisse. Werden Änderungen der Konfiguration vorgenommen, dann müssen nicht unbedingt alle Arbeitsschritte neu ausgeführt werden. Verwenden Sie den „Re-run“ Button, um eine 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 mehrere Config-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.

From:

<http://replicatio.science/dokuwiki/> - **documentatio replicationis**

Permanent link:

<http://replicatio.science/dokuwiki/doku.php/de/styloahonline/handbuch?rev=1726052441>

Last update: **2024-09-11**

