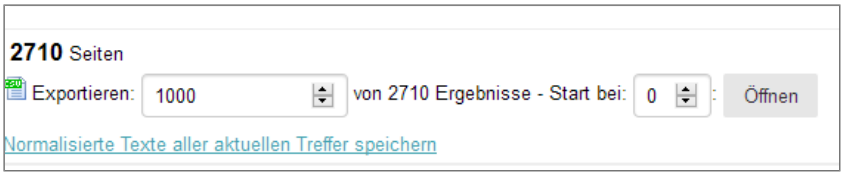
Export von Texten aus WdK-Explorer

Die Suchergebnisse im WdK-Explorer können als eine CSV-Datei exportiert und in anderen Programmen weitergehend analysiert werden. Viele Programme erfordern jedoch als Import eine Sammlung von Textdateien. Der Export und die Umwandlung in Textdateien wird hier mittels *LibreOffice* und einem dafür erstellten Makro vorgenommen. Exemplarisch wird dann der Import und die Analyse in *ConText*, einem Tool für die inhaltsbasierte Analyse und die Erstellung von Netzwerken aus den Texten vorgestellt. Dieses wird genutzt, um Korpusstatistiken und eigene Topic-Modelle zu erstellen.

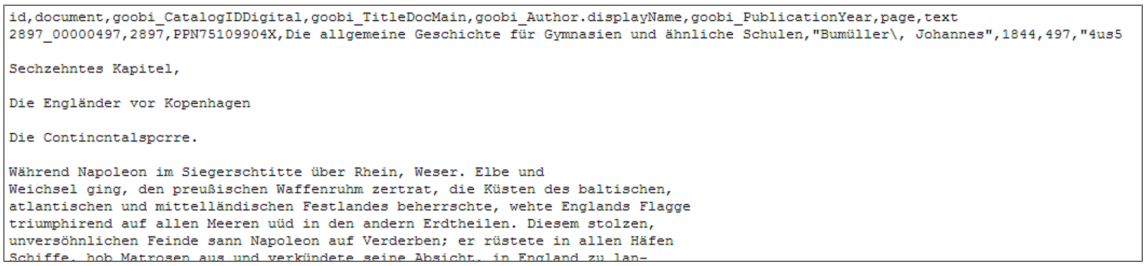
1. Export von Treffern

Für den Export von Treffern stehen in WdK-Explore zwei Möglichkeiten zur Verfügung: Ein Formular mit der Möglichkeit zur Angabe von Untermengen und ein Textlink zum Export größerer Mengen von Seiten im normalisierten Format. Es werden immer die aktuellen Suchtreffer exportiert.



1.a) *Export-Formular:* Exportieren: x von xy Ergebnissen

Für die Analyse von kleineren Treffermengen oder als Testexport wählen Sie in dem Formular, das sich ganz unten auf jeder Ergebnisseite findet, die gewünschte Anzahl an Ergebnissen. Die Ergebnisse sind sortiert nach aktueller Suchanfrage oder dem zur Sortierung ausgewählten Topic. *Hier wird immer der unveränderte Volltext inkl. Stoppwörtern exportiert.*



Weiterhin können die zu exportierenden Felder angepasst werden, indem die URL im Browser manipuliert wird (Paramter `&fl=` listet alle Feldnamen als zu exportierende Liste, vgl. Cheat-Sheet). Auf diese Weise können z.B auch größere Mengen an Metadaten ohne Text exportiert werden.

1.b) *Export-Link*: Normalisierte Texte aller aktuellen Treffer speichern

Alternativ, vor allem für größere Textmengen, kann auch der Link unten verwendet werden. Hier werden immer alle aktuellen Treffer gespeichert. Über linken Mausklick und `Seite speichern unter` kann der Browser auch für den Download von größeren Datenmengen eingesetzt werden. Alternativ kann der Link kopiert und in einem externen Download-Manager wie *curl* verwendet werden.

Die Treffer werden hier mit normalisiertem Text (text_normalized) ohne Stoppwörter exportiert in einem Format wie im Projekt für das Topic-Modeling verwendet.

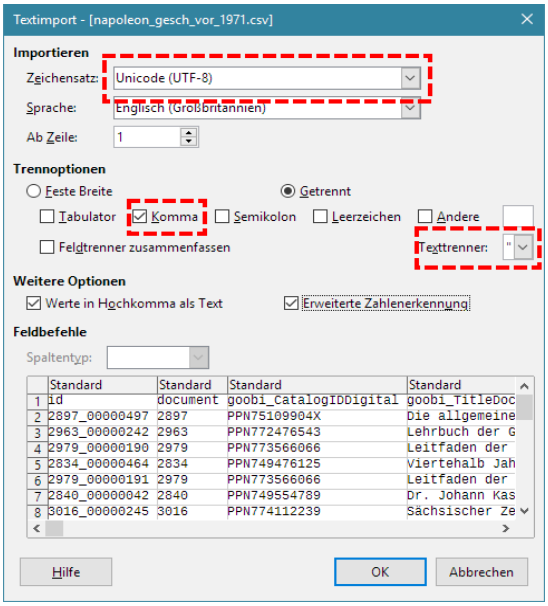
Achtung: Der Export von Volltexten kann zu Fehlern bei der weiteren Verarbeitung führen. Falls ein Text mit einem Schrägstrich endet, maskiert dies das Anführungszeichen, welches das Ende der Tabellenzelle markiert. Funktioniert der Import in ein Tabellenkalkulations-Programm wie Libreoffice-Calc nicht, ersetzen Sie bitte in einem Texteditor wie Editor oder Notepad++ über `Suchen und Ersetzen → Alle ersetzen` alle Zeichenfolgen Schrägstrich \" mit ".

Schritt 1: Bitte wählen Sie für unser Tutorial eine nicht zu große Treffermenge (max. ca. 2000 Seiten) aus und speichern Sie diese mit dem *Export-Formular*. Vergeben Sie am besten direkt beim Download *.csv als Dateiendung.

Öffnen Sie die Datei in einem Texteditor (Editor, Notepad++) und überprüfen Sie das Vorkommen der Zeichenfolge \" und korrigieren Sie diese wenn vorhanden.

2. Import in Tabellenkalkulation

Für eine erste Übersicht ist es hilfreich, die exportierten Daten in *Microsoft Excel* oder *Libre/Open-office Calc* aufzurufen. Hierfür müssen die korrekten Einstellungen für den Textimport getroffen werden. Beim Öffnen der gespeicherten Datei (mit Endung *.csv) wird in LibreOffice Calc direkt der Dialog zum Textimport angezeigt. Die Datei muss als Zeichensatz `UTF-8`, mit *„Komma" als Trennzeichen* und *Hochkomma " als Texttrenner* gespeichert werden wie im Bild unten gezeigt.



Zur besseren Übersicht in der Tabelle kann man alle Zeilen markieren (`Strg-a`) und die Zeilenhöhe reduzieren. Am besten überprüfen Sie gleich, ob die Anzahl der Zeilen der Anzahl der exportierten Seiten entspricht.

Schritt 2: Importieren Sie bitte die gesäuberte .csv-Datei in LibreOffice-Calc und überprüfen Sie den korrekten Import. Speichern Sie die Datei vorübergehend als ODF-Tabellendokument (.ods).

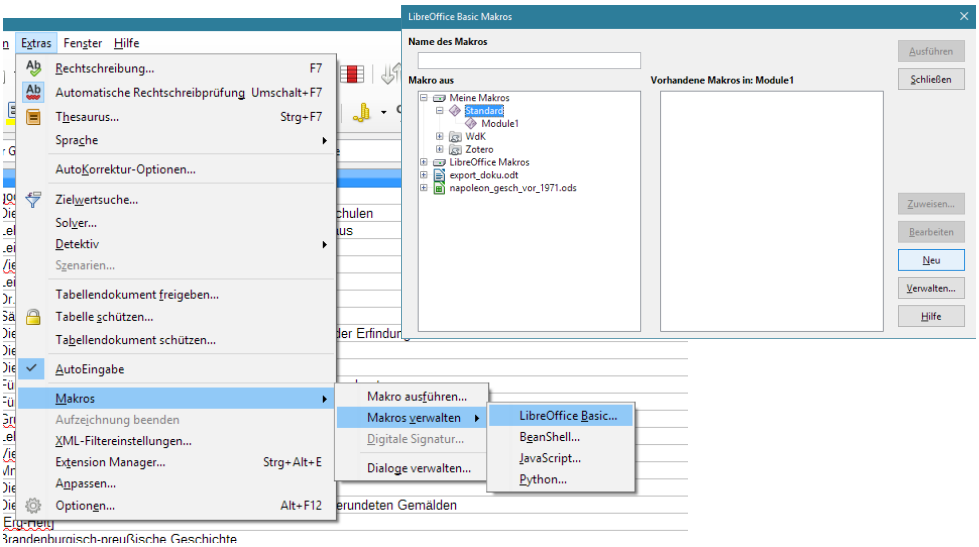
3. Export aus Calc als Sammlung von Textseiten

Für den Export der Texte in der CSV-Datei in der Form einer Sammlung von einfachen Textdateien (eine Datei per Seite) haben wir ein Makro vorbereitet, das den Ablauf vereinfacht. Das Makro verwendet den Text aus der Spalte `text_normalized` oder text und speichert ihn für jede Zeile (=Seite) in einer Datei mit Angaben zum Erscheinungsjahr und einer Id im Dateinamen. Diese Textdateien können in vielen Korpusanalyse-Programmen verwendet werden, von uns bereits in ConText[1] , voyant[2] und AntConc[3].

3.a) Makro importieren

Öffnen Sie die Datei `LibreOfficeBasic_WdK_Export_As_textfiles.bas` in einem Texteditor. Kopieren Sie den gesamten Inhalt in die Zwischenablage.

Öffnen Sie jetzt den Dialog `Extras → Makros → Makros verwalten → Libre Office Basic` und erstellen dort unter `Meine Makros  Standard` ein neues Makro. Es öffnet sich ein Makro-Editor, in dem Sie den bestehenden Text ersetzen und das WdK-Makro einfügen können. Das Makro ist jetzt in LibreOffice in allen Dokumenten auf diesem Rechner verfügbar.



3.b) Textdateien exportieren

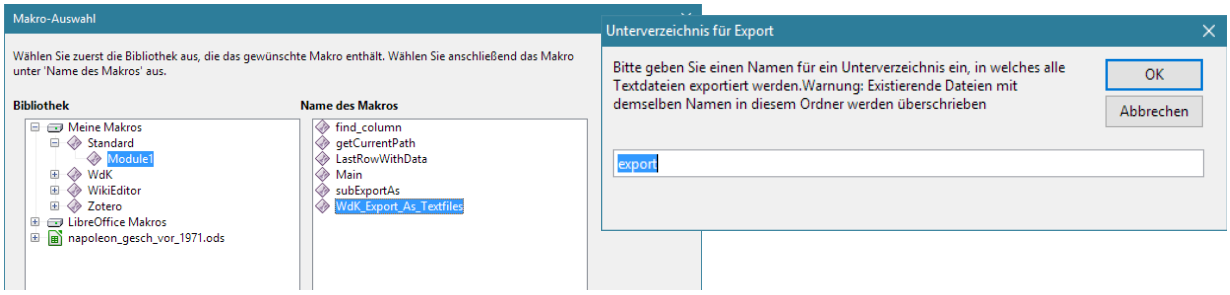
Öffnen Sie jetzt wieder das ODS-Dokument mit den exportierten Daten. Gehen Sie auf `Extras → Makros → Makro ausführen` und führen Sie dort unter `Meine Makros  Standard → Module 1` aus dem neu angelegten Makro die Funktion `WdK_Export_As_Textfiles` aus.

Im folgenden Dialog werden Sie um die Eingabe eines Ordner-Namens gebeten. Hier werden die Textdateien gesammelt gespeichert, *und zwar relativ zum Speicherort des Dokumentes.* Bestätigen Sie dann den Export-Auftrag.

1 http://context.lis.illinois.edu

2 http://voyant-tools.org

3 http://www.laurenceanthony.net/software/antconc



In dem von Ihnen angegebenen Verzeichnis finden Sie dann die exportierten Textdateien. Der Export kann, abhängig von der Dateigröße, einige Sekunden dauern. Es erfolgt eine Meldung über die Anzahl exportierter Zeilen, sobald der Prozess abgeschlossen ist.

Tritt während des Exports ein Fehler auf, werden leider keine sinnvollen Fehlermeldungen ausgegeben (stattdessen wird der Code des Makros geöffnet). Sollte es zu Problemen kommen, überprüfen Sie bitte zunächst die importierte Datei u.a. darauf, ob die Spalten korrekte Namen in der erste Zeile haben. Es muss mindestens eine Spalte mit dem Titel `id` und eine Spalte mit dem Titel `text` bzw. text_normalized vorhanden sein. Überprüfen Sie, ob in allen Zeilen eine `id` in der entsprechenden Spalte vorhanden ist.

Das Format der Dateinamen gibt neben einer laufenden Nummer die id des Dokumentes an. Beispiel:
`000001-id2897_00000497-PPN75109904X-pub1844.txt`

Die Zahl `2897_00000497` hinter `id` ermöglicht das Aufrufen der Seite im WdK-Exporer durch die Eingabe der Anfrage `id:2897_00000497`

Wenn die Spalten `goobi_PublicationYear` oder `goobi_CatalogIDDigital` vorhanden sind, werden entsprechende Angaben (zum Erscheinungsjahr und zur eindeutigen PPN des Werkes unter gei-digital.gei.de) zu den Dateinamen hinzugefügt (beginnend mit pub bzw. PPN).

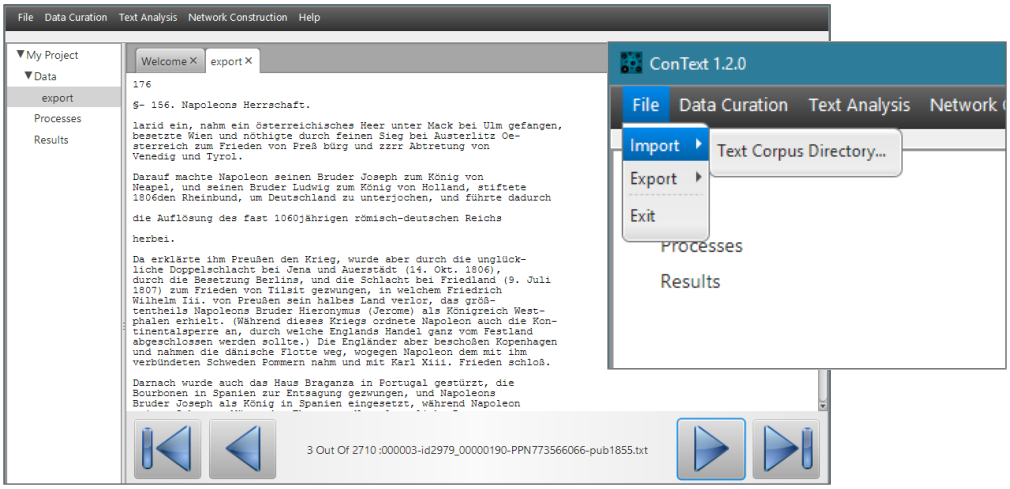
Schritt 3: Exportieren Sie bitte die Seiten aus dem ODS-Dokument als einzelne Textdateien mit Hilfe des WdK-Makros in den Unterordner `export`.

4. Textdateien analysieren mit ConText

ConText ist ein Text-Analyse-Programm zur Erstellung von Netzwerken aus Texten. Es lässt jedoch auch auf einfache Weise die Analyse von Termhäufigkeiten und die Erstellung von Topic-Models mit einer graphischen Benutzeroberfläche zu. Das Programm kann hier frei heruntergeladen werden: http://context.lis.illinois.edu/download.php

4.a) Textdateien importieren

Für den Import wählen Sie File Import Text Corpus Directory. Wählen Sie das Verzeichnis aus, in das Sie die Textdateien exportiert haben (`export`). Der Ordner findet sich dann unter `Data → Export` im Programm. An dieser Stelle sollte überprüft werden, ob der Dokumententext korrekt angezeigt wird, insbesondere Sonderzeichen. Falls dies nicht der Fall ist, liegt ein Problem mit der Zeichenkodierung vor.



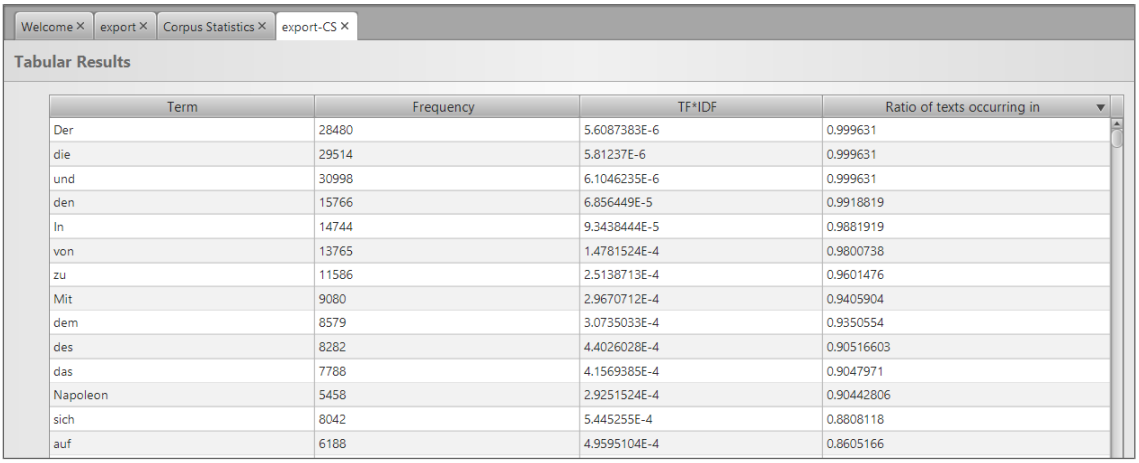
4.b) Analyse häufiger Begriffe

Unter dem Menüpunkt `Text Analysis` → `Summarization` → `Corpus Statistics` finden Sie eine Übersicht über häufige Terme im Korpus. Wählen Sie das importierte Korpusverzeichnis aus und geben Sie ein neues Verzeichnis als Speicherort für die Ergebnisse an (Der Hinweis zu POS ist für uns nicht relevant).

Hinweis: ConText verlangt für die Ausgabe von Ergebnissen immer die Angabe eines Dateiverzeichnisses. Versuchen Sie hier, sinnvolle Namen zu vergeben und vermeiden Sie es, Ergebnisse in Unterverzeichnissen von Korpusverzeichnissen zu speichern! Diese Option ist leider in ConText oft vorausgewählt.

Starten Sie jetzt die Berechnung (`Run`). Die Berechnung kann einige Minuten erfordern. Es kann jedoch auch vorkommen, dass ConText nicht mehr reagiert und neu gestartet werden muss.

Die Ergebnisse werden in dem Ordner gespeichert, können jedoch auch innerhalb von ConText analysiert werden (`Open Tabular Results`).

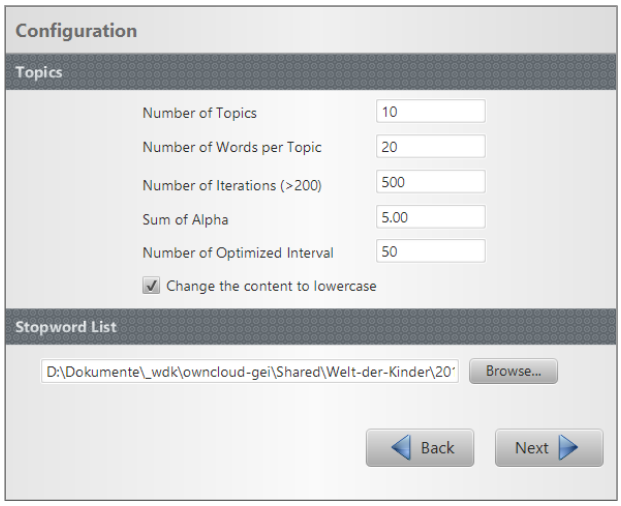


Term	Frequency	TF*IDF	Ratio of texts occurring in
Der	28480	5.6087383E-6	0.999631
die	29514	5.81237E-6	0.999631
und	30998	6.1046235E-6	0.999631
den	15766	6.856449E-5	0.9918819
In	14744	9.3438444E-5	0.9881919
von	13765	1.4781524E-4	0.9800738
zu	11586	2.5138713E-4	0.9601476
Mit	9080	2.9670712E-4	0.9405904
dem	8579	3.0735033E-4	0.9350554
des	8282	4.4026028E-4	0.90516603
das	7788	4.1569385E-4	0.9047971
Napoleon	5458	2.9251524E-4	0.90442806
sich	8042	5.445255E-4	0.8808118
auf	6188	4.9595104E-4	0.8605166

<u>Schritt 4:</u> Importieren Sie die exportierten Dateien in ConText und überprüfen Sie zur Kontrolle die häufigsten Terme.

4.c) Berechnung von Topic Models

Unter dem Menüpunkt `Text Analysis` → `Summarization` → `Topic Modeling` finden Sie die Optionen zur Erstellung von Topic-Models.



Number of Topics: Der wichtigste Parameter ist die Anzahl von Topics, von denen bei der Berechnung ausgegangen wird. Hier hilft es am ehesten, einige Male zu experimentieren - abhängig von der thematischen Vielfalt der Inhalte der Dokumente.

Number of Words per Topic: Die Anzahl von Worten pro Topic wirkt sich dagegen nicht auf die Berechnung sondern nur auf die Anzeige der Ergebnisse aus.

Number of Iterations: Anzahl an Berechnungsdurchläufen für den Optimierungsprozess: Kann die Qualität steigern, wirkt sich jedoch auf die Berechnungsdauer aus. Zum Experimentieren mit den anderen Parametern reduzieren (ca. 200-300). Mehr als 1000 Iterationen bringen nur noch geringe Vorteile.

Sum of Alpha: Dieser Parameter bestimmt, ob stärker auf die Kohärenz der Worte in den Topics oder auf die Kohärenz der Topics in den Dokumenten (wenige Topics je Dokument) optimiert werden soll. Je höher der Alpha-Wert, desto mehr Topics je Dokument werden erwartet.

Number of Optimized Interval: Wirkt sich auf die Berechnungsqualität und die Laufzeit aus.

*Change the content to lowercas*e: Sinnvollerweise aktiviert

Stopword List: Angabe einer einfachen Liste typischerweise häufig auftretender Worte, die nicht inhaltstragend sind und das Ergebnis verfälschen können. Hier ist eine englische Stoppwortliste voreingestellt. Deutsche Stoppwortlisten finden sich online. Wir haben eine um sehr häufig auftretende historische Varianten der Stoppwörter erweiterte deutsche Stoppwortliste verfügbar gemacht (stopwords-mallett-de_wdk.txt) Es kann sinnvoll sein, diese Stoppwortliste um themenspezifisch sehr häufig auftretende, aber nicht inhaltstragende Begriffe zu erweitern oder die für die Auswahl verwendeten Suchbegriffe durch die Aufnahme in die Liste auszuschließen.

<u>Schritt 5:</u> Erstellen Sie aus dem Korpus eigene Topic-Modelle und experimentieren Sie mit dem Parameter *Number of Topics* bis eine für Ihre Fragestellung sinnvolle Auswahl vorliegt.

Vergleichen Sie diese Topics mit denen auf WdK-Explorer zur Verfügung gestellten Topics, die zu Ihrem Untersuchungsthema passen.