Bibliometric Review – eine Anleitung zur Erschließung von Forschungsthemen

Arbeitsschritt 2 von 3: Datenbereinigung

Zitiervorschlag:

Schneijderberg, Christian und Steinhardt, Isabel (2020). Bibliometric Review – eine Anleitung zur systematischen Erschließung von Forschungsthemen. Arbeitsschritt 2: Datenbereinigung. Kassel. https://doi.org/10.5281/zenodo.3919721

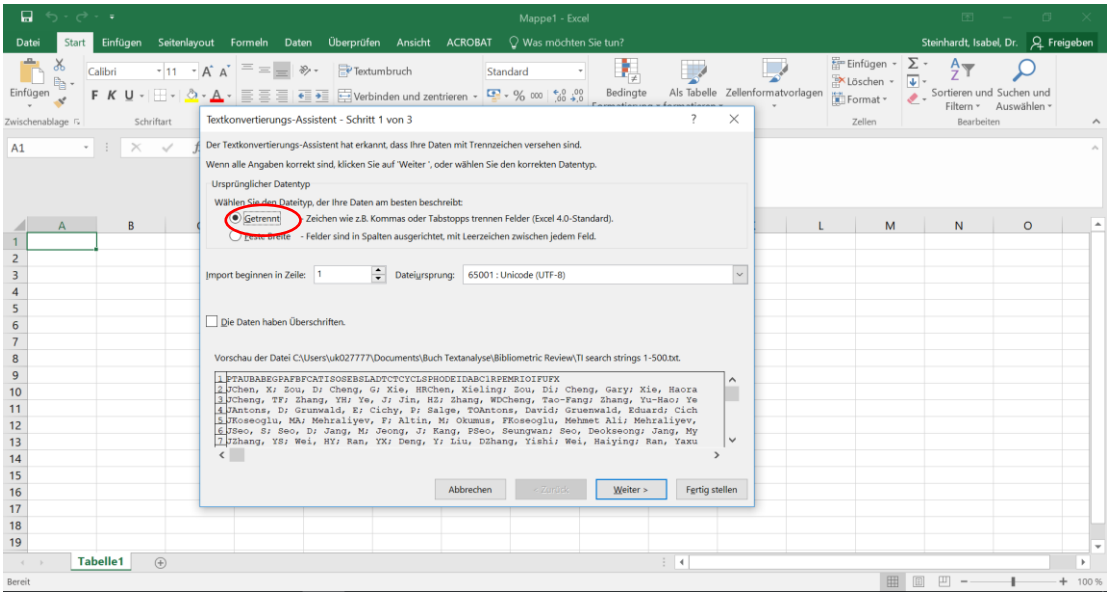
Nachdem die Datenerhebung (Arbeitsschritt 1) geschafft ist, kann nun die Datenbereinigung folgen.

Benötigte Software: Excel und VosViewer (FreeWare)

Schritt 2.1

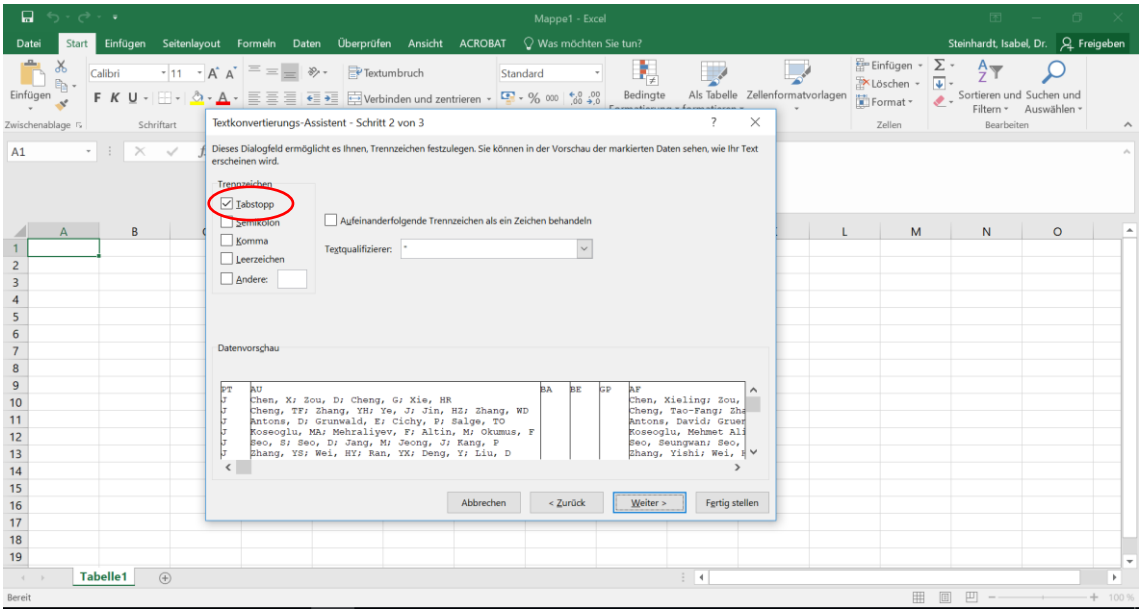
Zur Datenbereinigung ist es sinnvoll zunächst zu prüfen, ob im Datensatz doppelte Werte vorhanden sind. Dazu das txt Dokument in Excel öffnen, das Ihr im Arbeitsschritt 1 Datenerhebung zusammengefügt habt. **ACHTUNG**: Die txt-Datei wird Euch nur angezeigt, wenn ihr beim Öffnen auf „alle Dateien“ klickt.

Wenn ihr die txt-Datei öffnet, erscheint dieses Dialogfenster. Angegeben muss sein „getrennt“. Dann „weiter“ klicken.



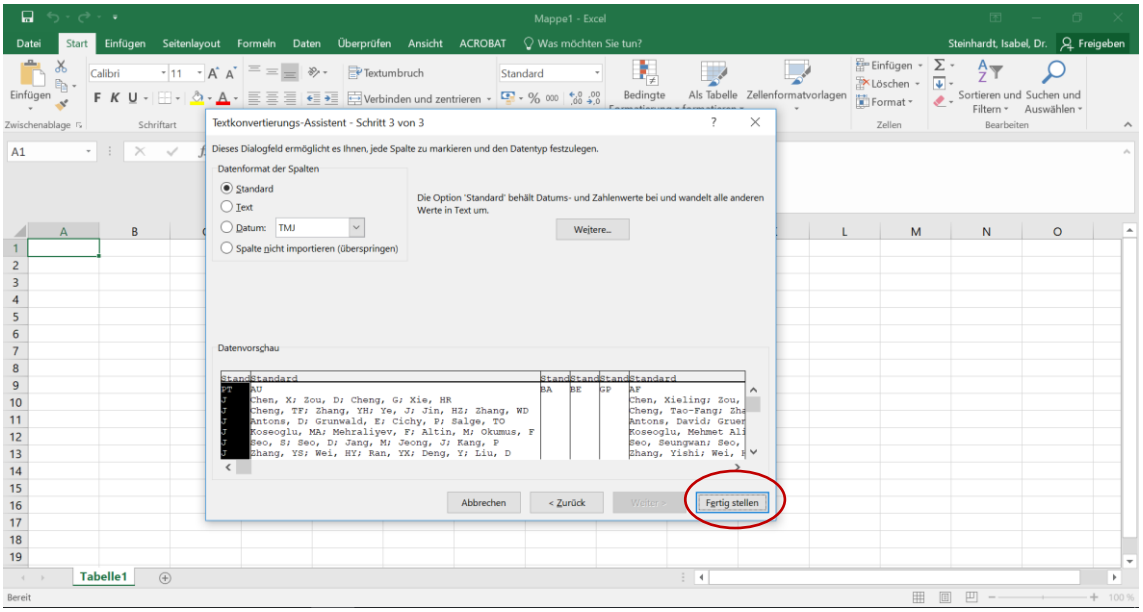
Schritt 2.2

Im nächsten Fenster „Tabstopp“ auswählen.



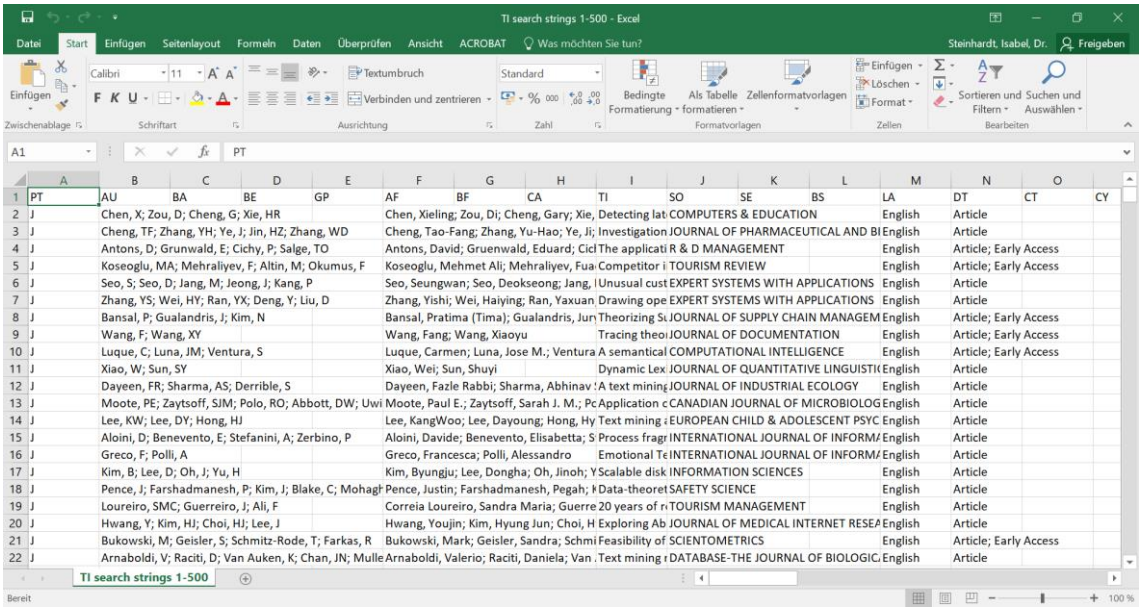
Schritt 2.3

Im nächsten Dialogfeld alles so lassen, wie es ist und auf „fertig stellen“ klicken



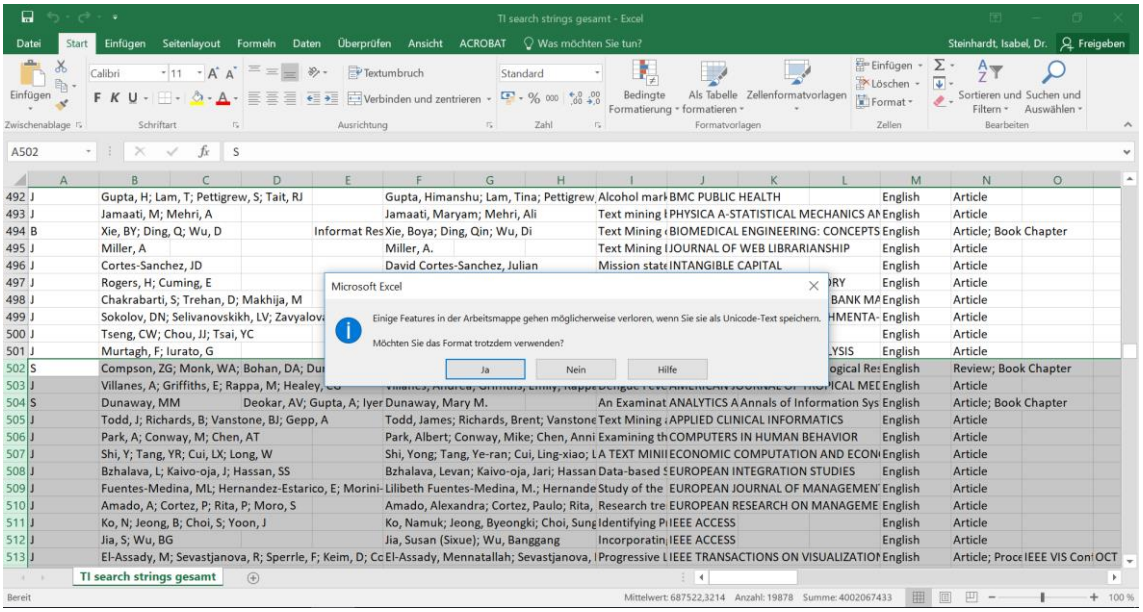
Schritt 2.4

Dann erhaltet ihr eine Liste mit allen Literaturangaben des Korpus.



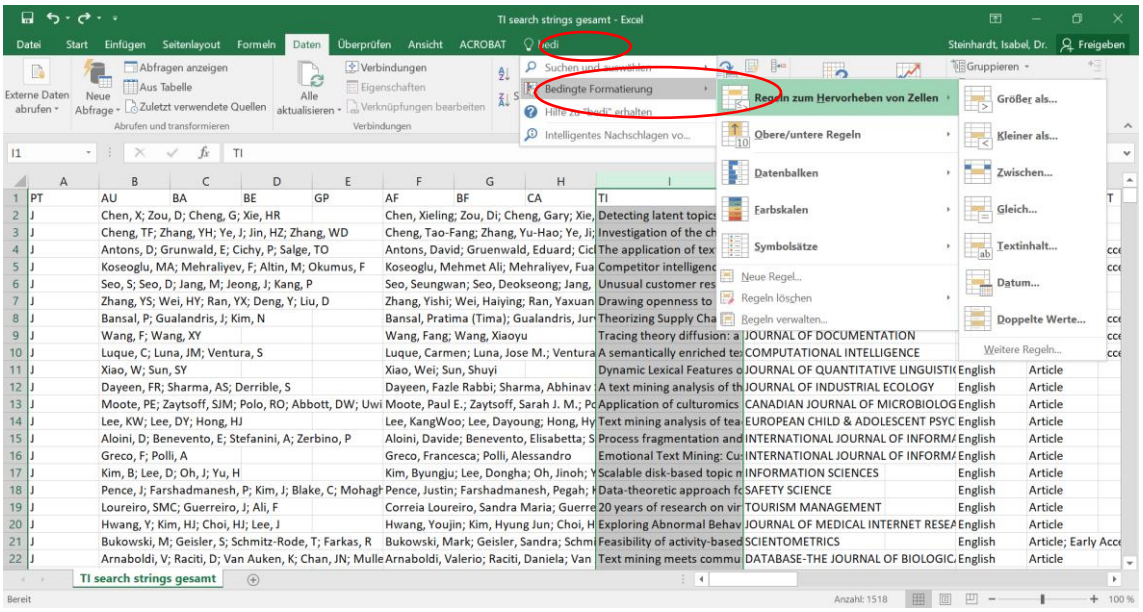
Schritt 2.5

Das Dokument unter einem neuen Namen speichern. **WICHTIG**: Datei immer als unicode-Dokument speichern (nie als Excel-Dokument), sonst kann VosViewer das Dokument nicht mehr lesen! Bei dem Dialogfeld „Einige Features in der Arbeitsmappe…“ auf „Ja“ klicken. Die hier verwendete Datei wurde benannt in „TI search strings gesamt“.



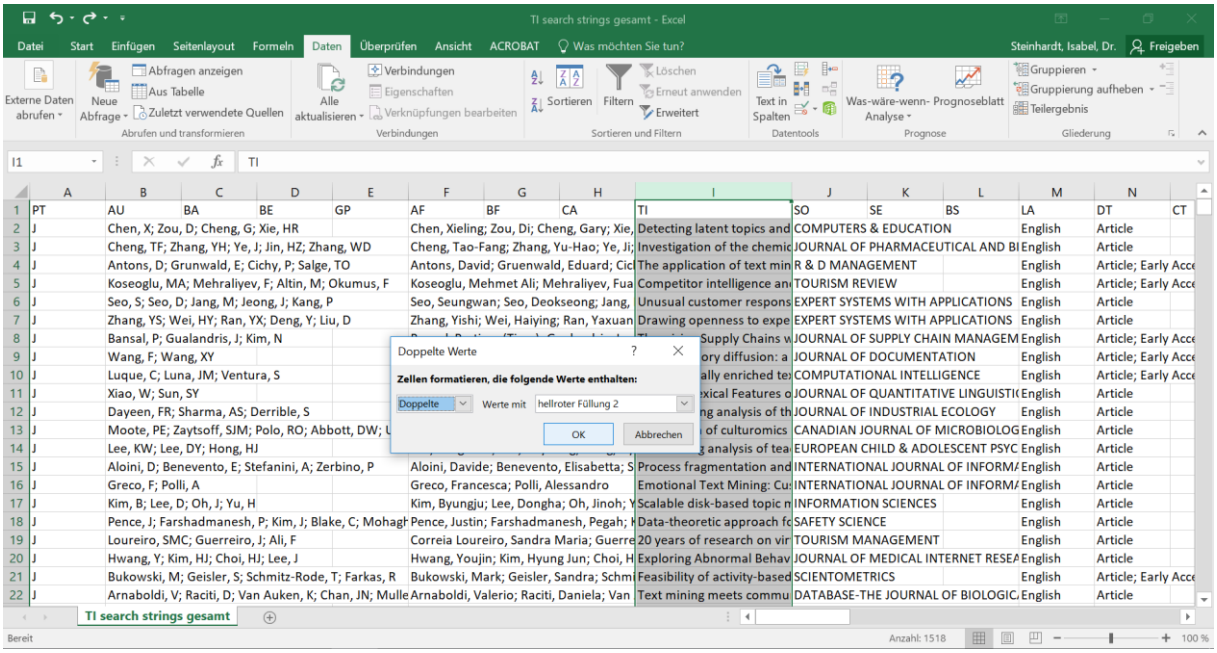
Schritt 2.6

Als nächstes könnt Ihr nach Duplikaten im Dokument suchen. Dazu müsste Ihr zunächst die Spalte TI (=Titel) markieren. Es ist am einfachsten über die Titel zu suchen. Dazu einfach die Suchoption verwenden, das ist die Glühbirne hinter der steht: „Was möchten Sie tun?“. In die Suchmaske „bedingte Formatierung“ eingeben, „Regeln zum Hervorheben von Zellen“ auswählen und auf „Doppelte Werte“ klicken. **ACHTUNG**: es ist wichtig, dass die Spalte TI ausgewählt ist, sonst sucht ihr im gesamten Dokument und werdet nichts finden.



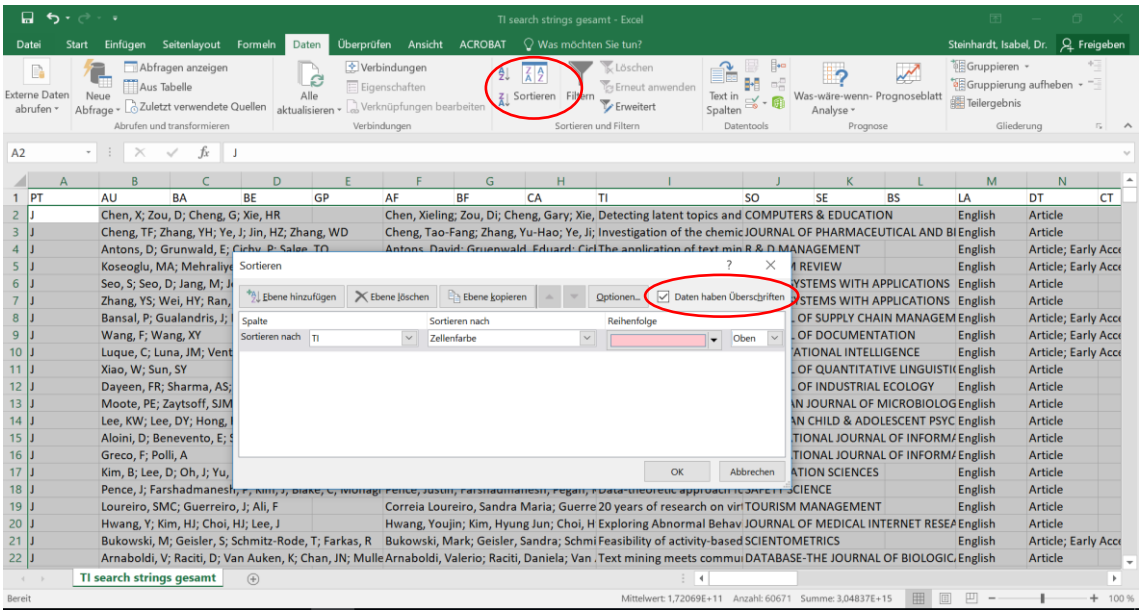
Schritt 2.7

Beim nächsten Fenster auf „OK“ klicken damit die doppelten Werte rot markiert werden.



Schritt 2.8

Nun nach den doppelten Werten sortieren. Dazu das gesamte Tabellenblatt markieren. Auf „Daten“ und „Sortieren“ gehen und dann „Spalte“ = TI, „Sortieren nach“ = Zellenfarbe und „Reihenfolge“ = rot oben auswählen. **ACHTUNG** es muss „Daten haben Überschriften“ ausgewählt sein.



Schritt 2.9

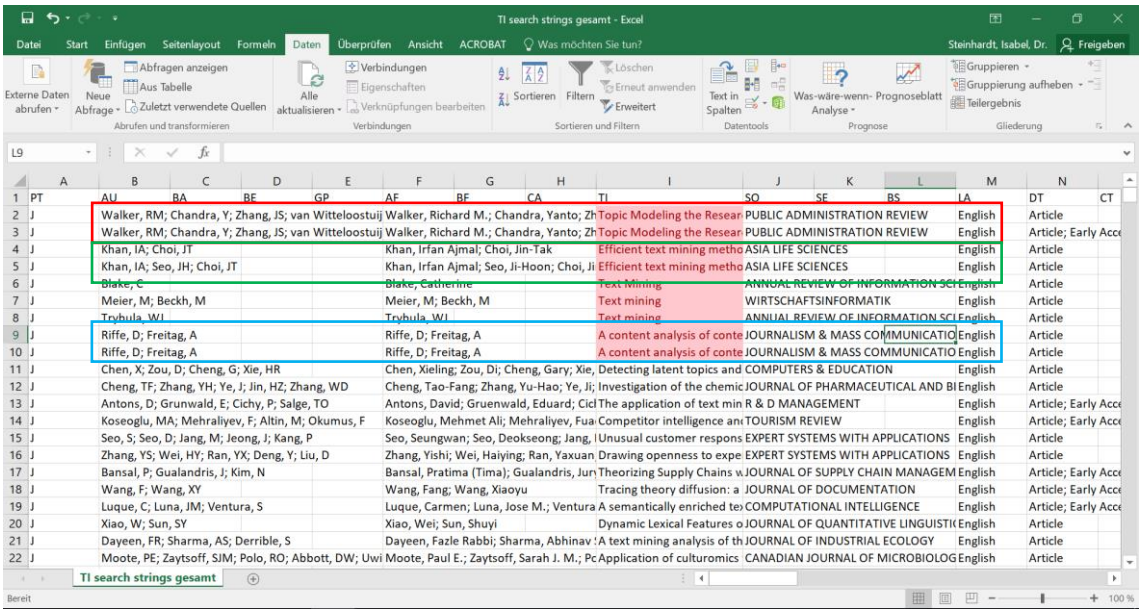
Jetzt sind die Treffer mit doppelten Werten oben gelistet. In diesem Beispiel sind es 9 Angaben. Die ersten beiden Treffer (= rot umrandet) sind die gleichen Texte, der Unterschied besteht im Erscheinungsjahr. Einmal ist kein Erscheinungsjahr angegeben, einmal als *Early Access*. Hier die Angabe mit mehr Information auswählen, also die mit *Early Access*.

Beim den nächsten Angaben (= grün umrandet) ist es ebenfalls der gleiche Text, aber einmal mit zwei Autor*innen und einmal mit drei. Hier muss geprüft werden, welche Angabe korrekt ist, am besten auf der Homepage der Zeitschrift.

Die nächsten drei Angaben haben zwar den gleichen Titel, sind aber unterschiedliche Artikel.

Bei den letzten beiden Angaben (blau umrandet) handelt es sich um den gleichen Artikel, mit unterschiedlichen Angaben bzgl. der Heftnummer der Zeitschrift. Hier muss geprüft werden, welche Angabe korrekt ist, am besten auf der Homepage der Zeitschrift.

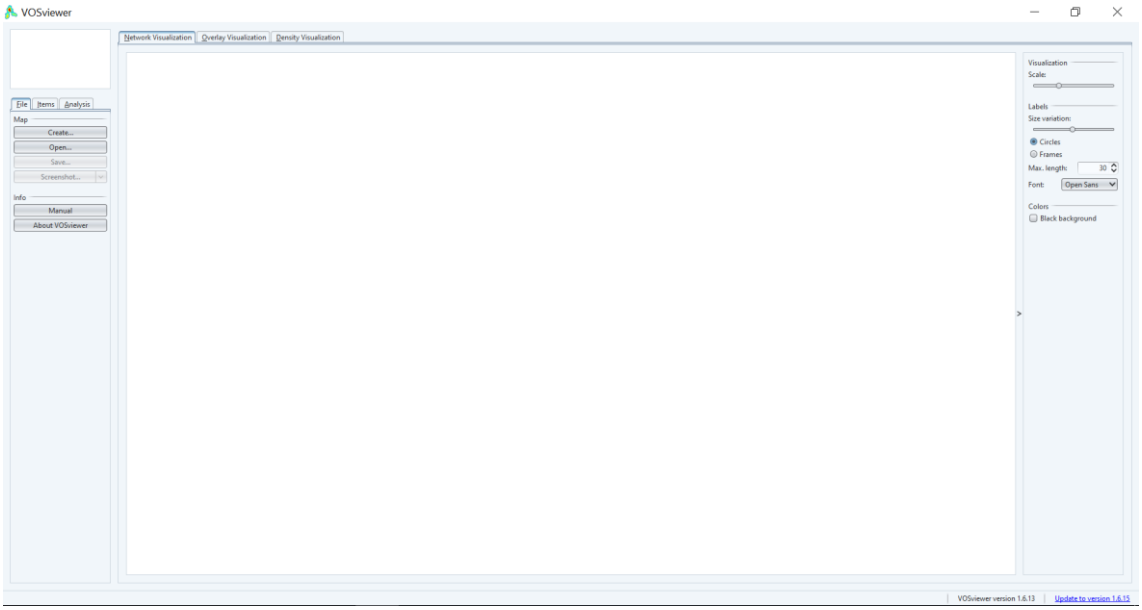
Nachdem die Duplikate entfernt wurden die Excel-Tabelle als unicode speichern.



Schritt 2.10

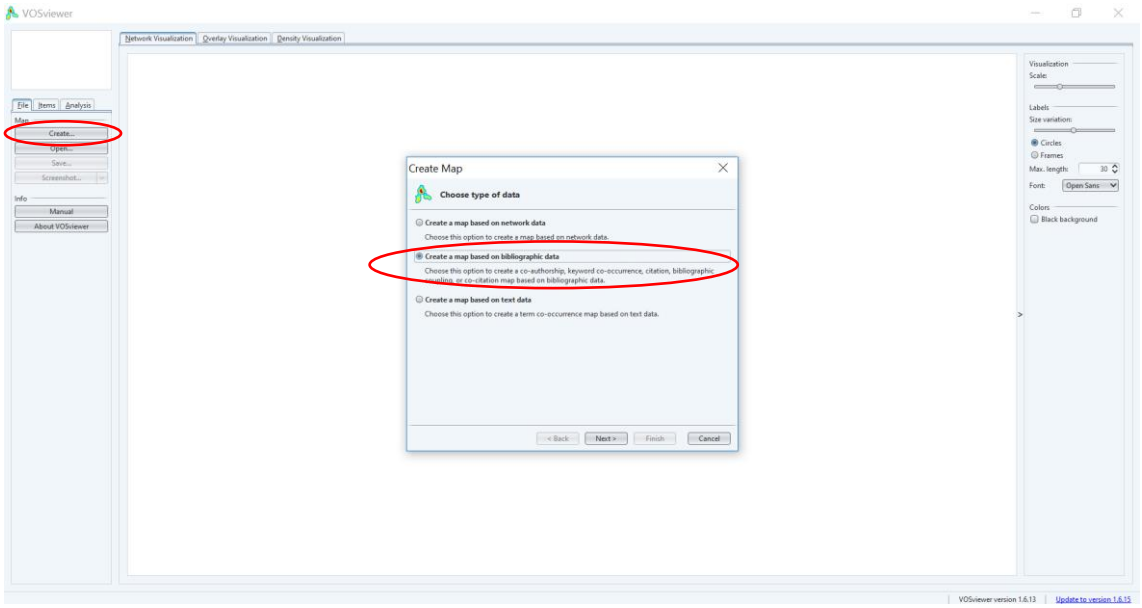
Um nun die Bereinigung der Zitationen vorzunehmen kann VosViewer genutzt werden. Das Programm ist Open Source und kann hier heruntergeladen werden: https://www.vosviewer.com/. **ACHTUNG**: Damit das Programm funktioniert muss Java (https://www.java.com/de/download/) installiert sein.

Für die Bereinigung wird eine *Co-Citation* Analyse genutzt. Unten seht Ihr das Startfenster.



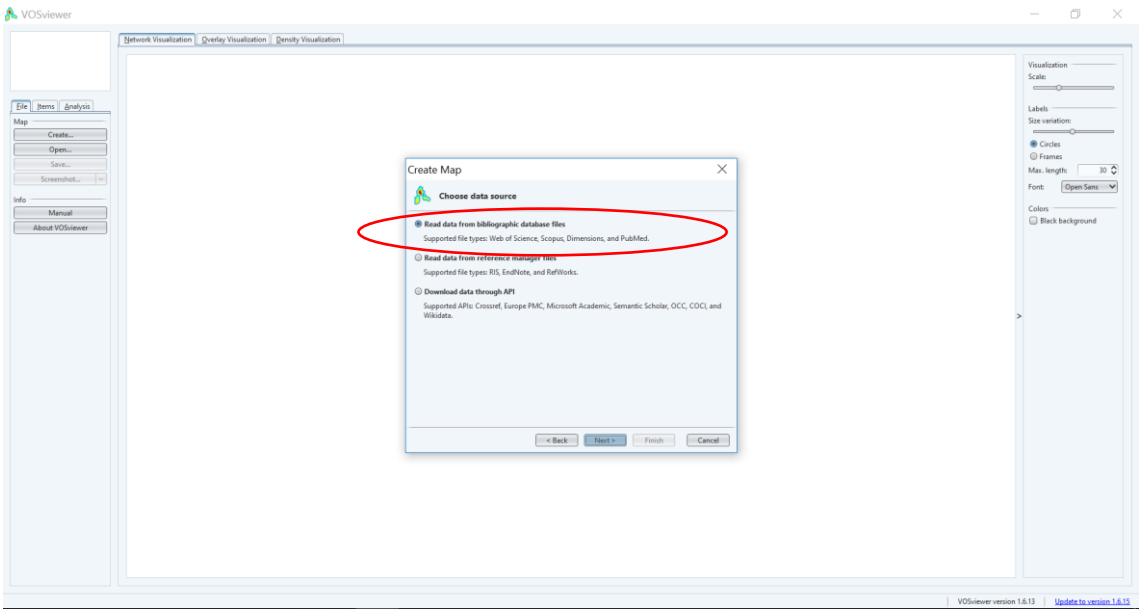
Schritt 2.11

Als erster Schritt muss die *Co-Citation* Analyse gestartet werden. Dazu geht man auf „Create“ und wählt im aufscheinenden Fenster „Create a map based on bibliometric data“ aus und klickt auf „Next“.



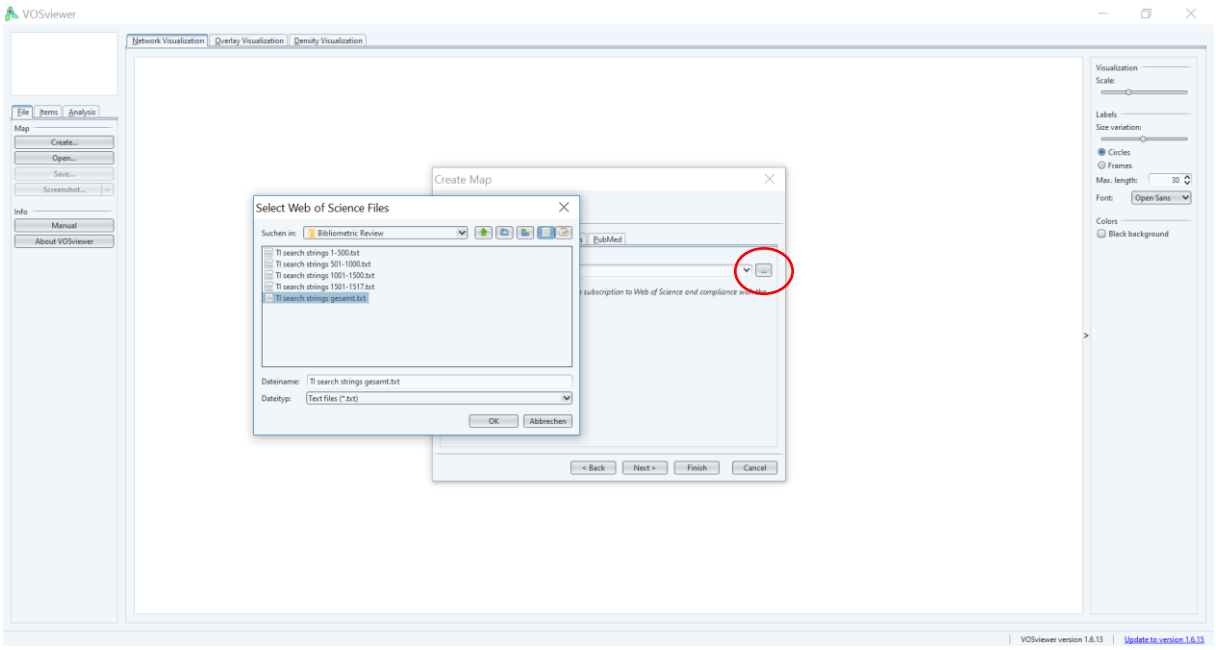
Schritt 2.12

Im nächsten Fenster „Real data from bibliographic database file“ auswählen und „Next“ klicken.



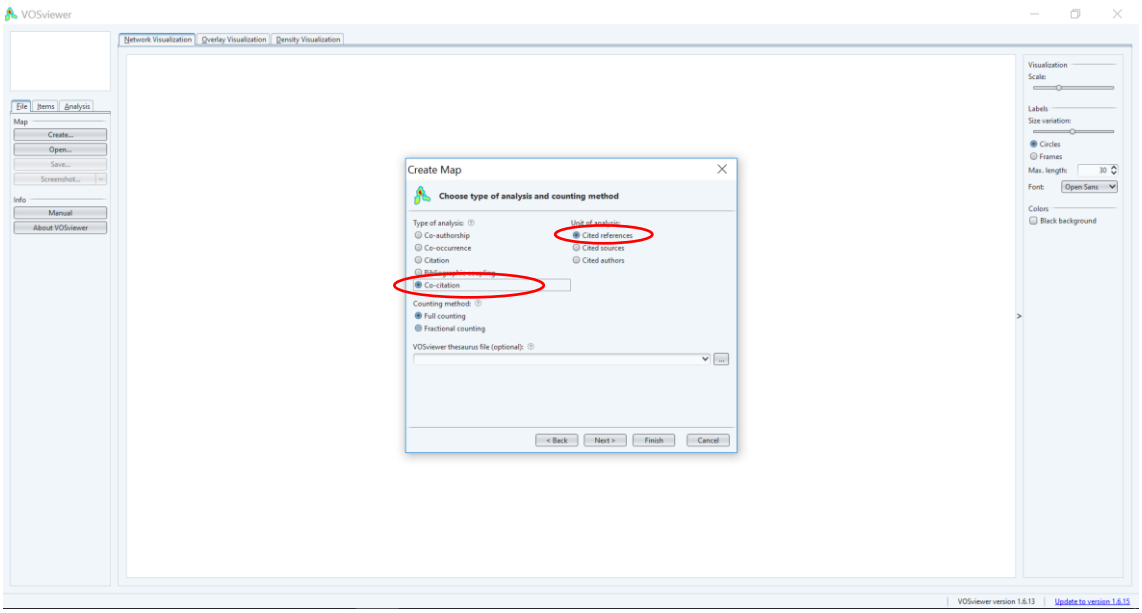
Schritt 2.13

Im nächsten Fenster auf das Viereck mit den drei Punkten klicken und die bereinigte txt-Datei mit dem erstellten Korpus auswählen. Dann klickt auf „Next“.



Schritt 2.14

Im nächsten Fenster kann die Art der Analyse ausgewählt werden. Für die Datenbereinigung der Zitationen wählt ihr „Co-citation“ und „Cited references“. Und klickt auf „Next“.



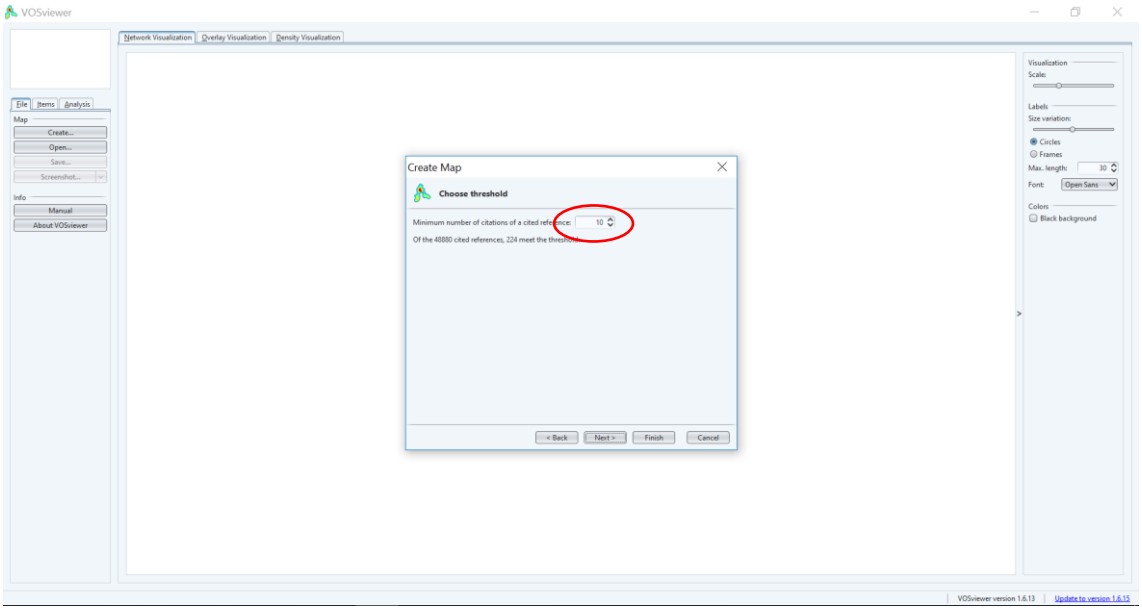
Schritt 2.15

Im nächsten Fenster kann nun die Anzahl der Zitationen ausgewählt werden, die ein Text mindestens haben muss.

Welche Parameter hier ausgewählt werden hängt davon ab, welche Analyse man mit den Daten machen will. Soll es eine sehr feine Analyse werden, in der viele kleine *Cluster* auftauchen und z. B. die Verzweigungen eines Forschungsfeldes wichtig sind, dann muss auch eine feine Datenbereinigung vorgenommen werden, also hier eine möglichst kleine Anzahl an Zitationen gewählt werden.

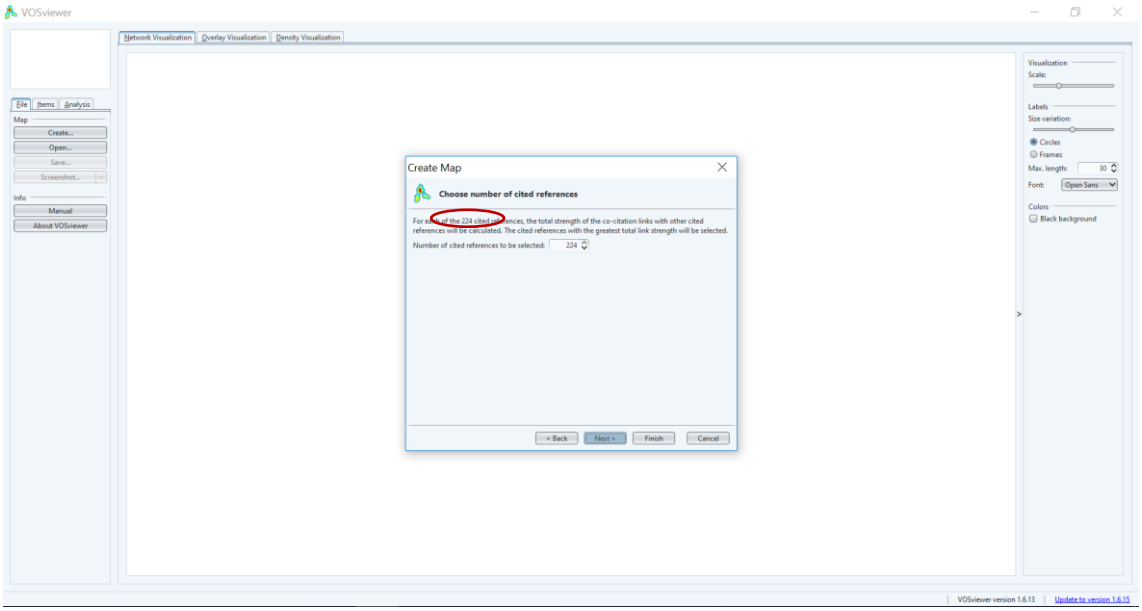
Was eine kleine Anzahl ist, hängt stark von der Größe des Korpus ab. Ist es ein großer Korpus, dann ist eine kleine Anzahl z. B. 10. Ist es ein kleiner Korpus dann eher 2 oder 3.

Will man hingegen vor allem die zentralsten Artikel eines Korpus ermitteln und nicht in die Tiefe gehen, dann muss auch die Datenbereinigung nicht so fein sein. Denn Artikel die nur ein oder zweimal zitiert wurden, tauchen dann eh nicht mehr in der Analyse auf. Wenn die Analyse nicht so fein sein soll, dann würde man je nach Größe des Korpus eher eine Mindestanzahl an Zitationen („Minimum number of citations“) von 10 bis 20 wählen. Nach der Auswahl wieder auf „Next“ klicken.



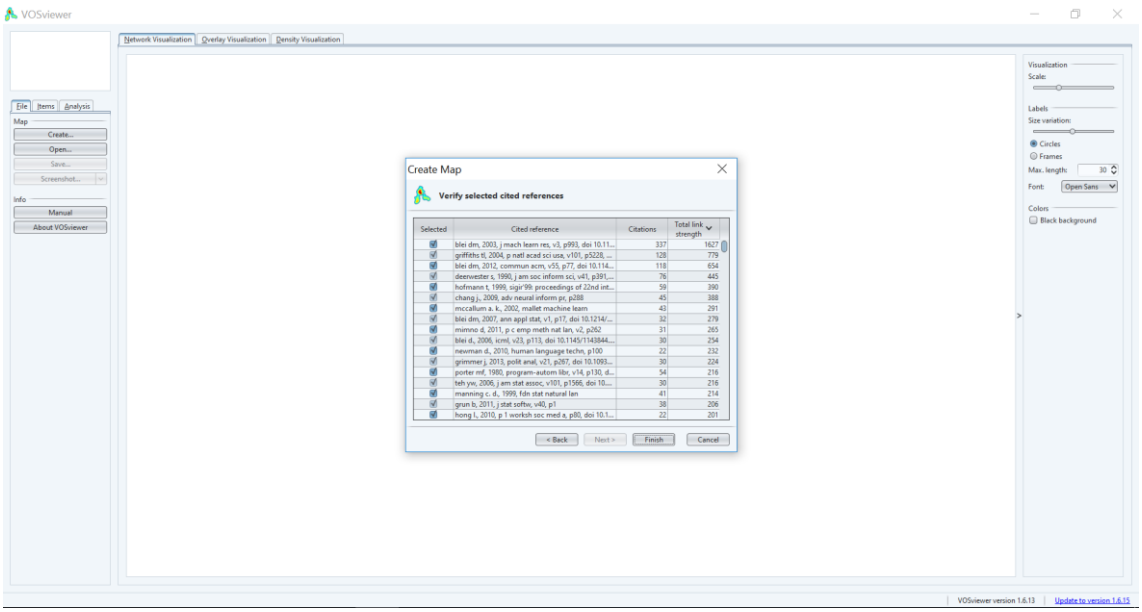
Schritt 2.16

Das nächste Fenster zeigt an, wie viele Artikel nun in die Analyse einbezogen werden sollen. Hier die Anzahl auswählen („Number of cited references“), die im Text angezeigt wird und auf „Next“ klicken.



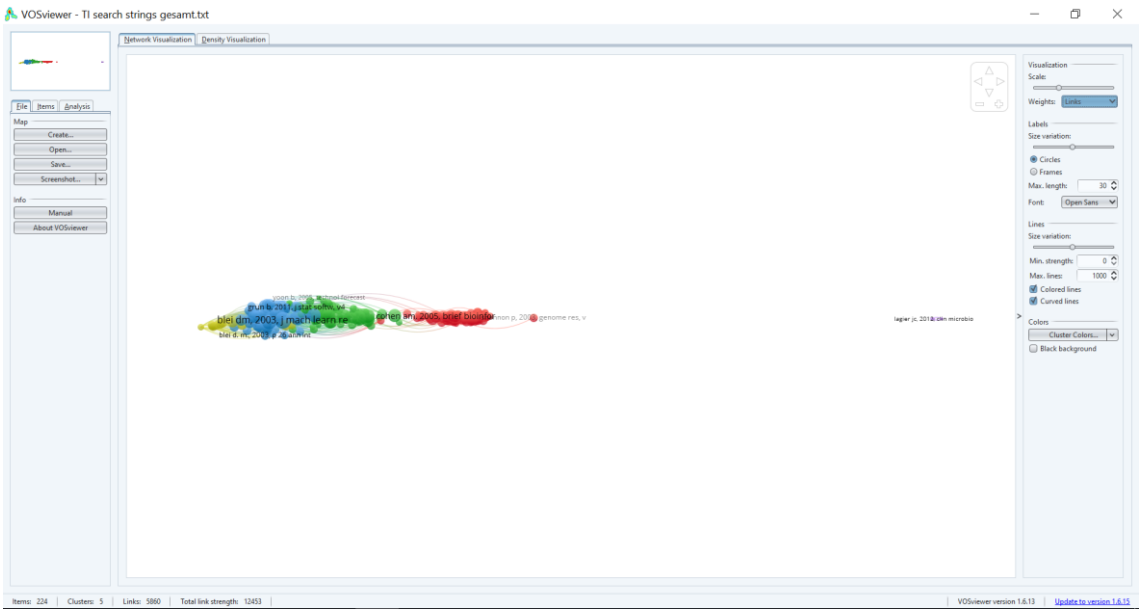
Schritt 2.17

Das nächste Fenster zeigt die Liste der in die Berechnung involvierten Artikel an. Einfach „Finish“ klicken.



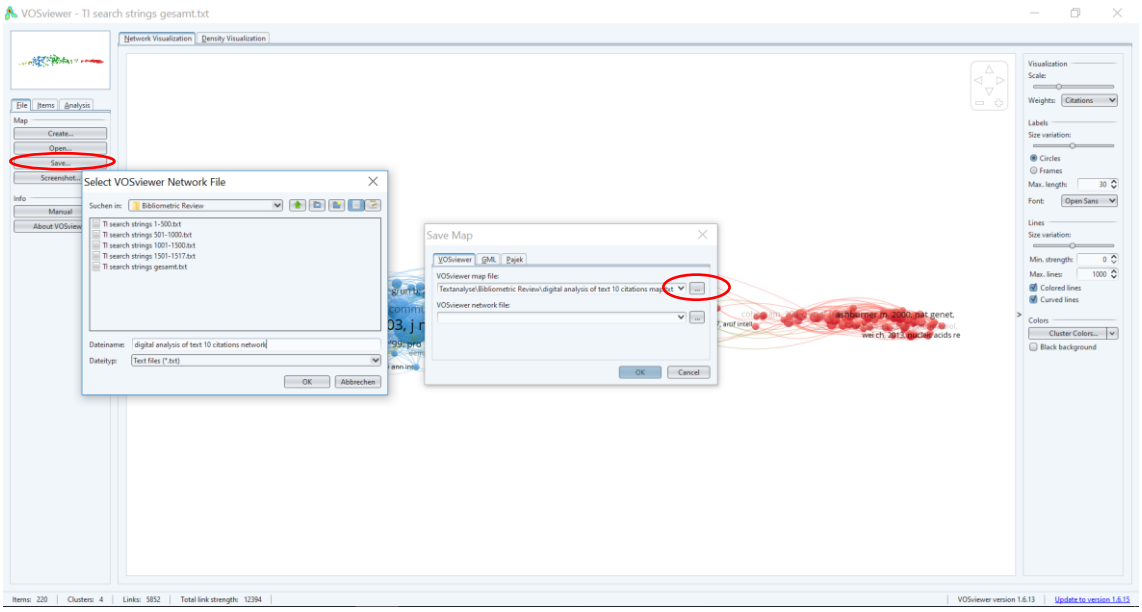
Schritt 2.18

Nun hat man die erste *Co-Citation* Analyse gemacht. Wie die *Map* aussieht ist im Moment egal, da es hier zunächst nur um die Datenbereinigung geht.



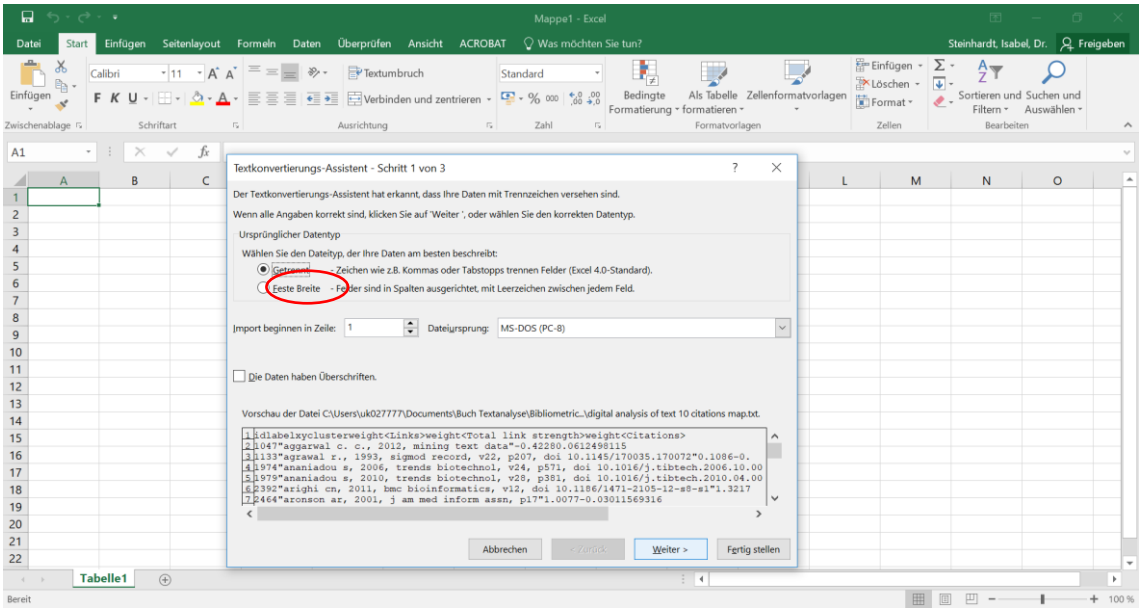
Schritt 2.19

Um die Datenbereinigung der Zitationen durchführen zu können muss als nächstes die Datei mit den zitierten Artikeln gespeichert werden. Dazu auf „Save“ klicken und „Save Map“ auswählen. Im darauf erscheinenden Fenster wieder auf das Viereck mit den drei Punkten klicken und den Speicherort auswählen. Dabei ist wichtig sowohl die *Map* als auch das Network zu speichern. ACHTUNG die Dateien gut benennen, so dass sie einfach auseinandergehalten werden können. Z. B. „Projekt xy MAP“ und „Projekt xy NETWORK“.



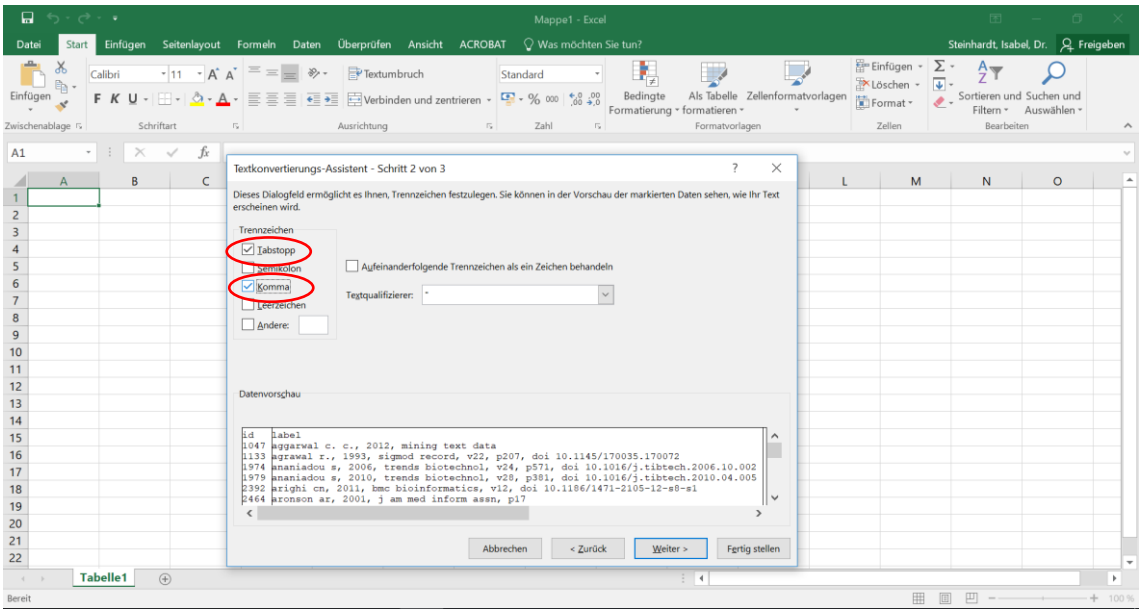
Schritt 2.20

Als nächster Schritt kann die Datei „Projekt xy MAP“ in Excel geöffnet werden. Es handelt sich dabei wieder um eine txt-Datei, weshalb sie wieder konvertiert werden muss. Dazu wieder als ersten Schritt „Getrennt“ auswählen.



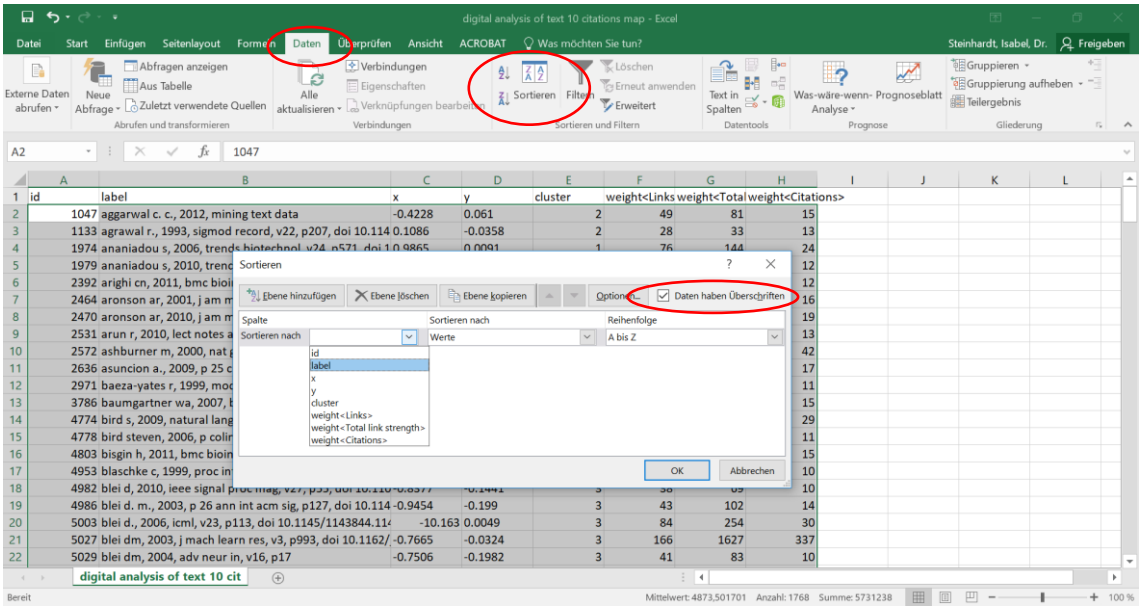
Schritt 2.21

Diesmal bei Schritt 2 von 3 der Textkonvertierung bitte „Tabstopp“ und „Komma“ auswählen.



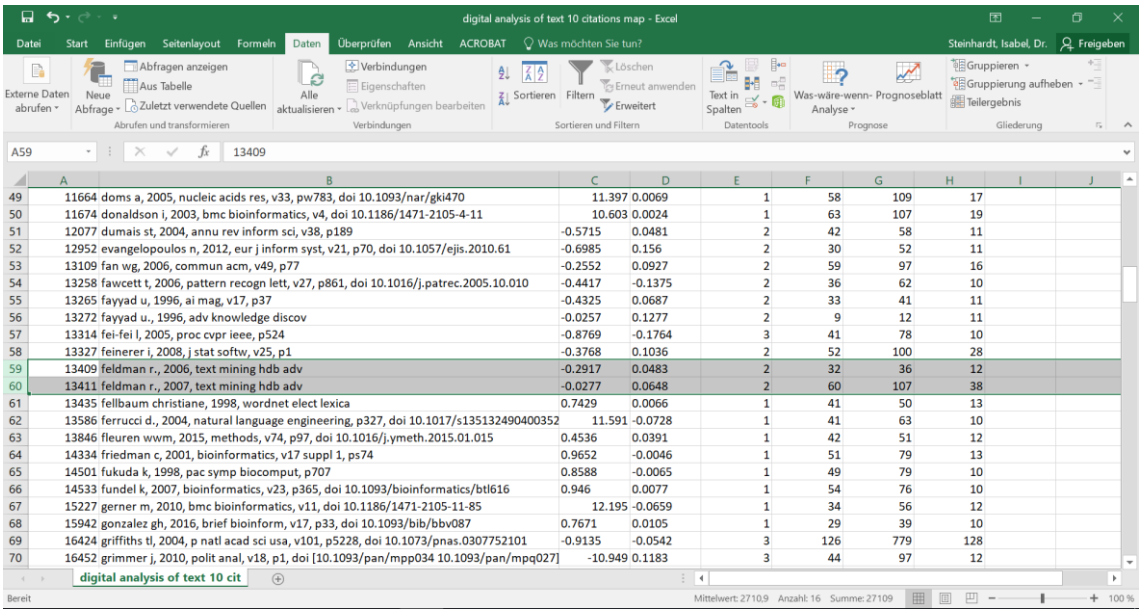
Schritt 2.22

Als nächstes das Tabellenblatt markieren und dann unter „Daten“ „Sortieren“ wählen. Unter „Sortieren nach“ „Label“ auswählen. **ACHTUNG** „Daten haben Überschriften“ muss markiert sein.



Schritt 2.23

Dann muss der Datensatz systematisch auf der Suche nach Doppelungen durchgegangen werden.



Schritt 2.24

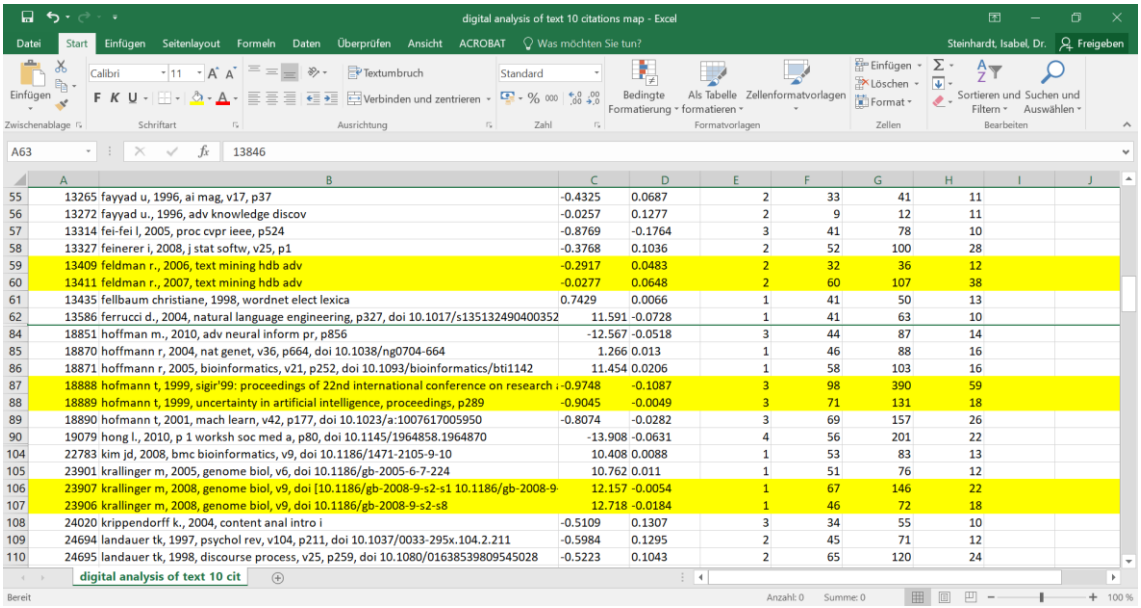
In dem vorliegenden Datensatz war das Ziel eine grobe Datenbereinigung, da es um die zentralsten *Cluster* und Texte ging. Deshalb wurde bei der Mindestanzahl an Zitationen 10 gewählt, was zu einer Liste mit 241 Artikeln führte. In dieser sind drei Übereinstimmungen aufgefallen, die mögliche Datenfehler sein könnten.

Das sind sehr wenige Fehler. Das hat zwei Gründe: Erstens sind die meisten Zitationen Zeitschriftenartikel, die bessere Metadaten haben als z. B. Bücher (bei denen z. B. Erst- und Zweitauflagen existieren, die nicht als ein Buch erkannt werden). Schlechte Metadaten zeichnen sich auch dadurch aus, dass z. B. Vornamen manchmal abgekürzt werden, manchmal aber ausgeschrieben sind, was nicht erkannt werden kann. Zudem haben vor allem ältere Texte keine DOI, wodurch eine klare Zuordnung schwieriger ist.

Der zweite Grund, warum wenige Fehler in diesem Datensatz sind liegt darin begründet, dass es sich bei dem Beispiel um ein junges Forschungsfeld handelt und deshalb v.a. auch neuere Literatur zitiert wird, die wie oben beschrieben, bessere Metadaten hat.

In diesem Beispiel wurden also drei mögliche Fehler gefunden (alle gelb markiert):

1. Feldman: hier handelt es sich um ein Buch, das mit zwei unterschiedlichen Erscheinungsdaten vorhanden ist. Hier muss recherchiert werden, welches Datum stimmt. Gibt es zwei Auflagen muss geprüft werden, ob zwischen den Auflagen der Inhalt des Buchs verändert wurde. Ist das nicht der Fall, dann wird die erste Auflage gewählt und die anderen Einträge entsprechend angepasst –> das passiert im nächsten Schritt.
2. Hoffmann: hier handelt es sich um unterschiedliche Texte.
3. Krallinger: hier handelt es sich um die gleichen Texte. Es gab einen Fehler bei der DOI. Auch hier muss geprüft werden, welche Angabe stimmt und die andere Angabe entsprechend ersetzt werden -> nächster Schritt.

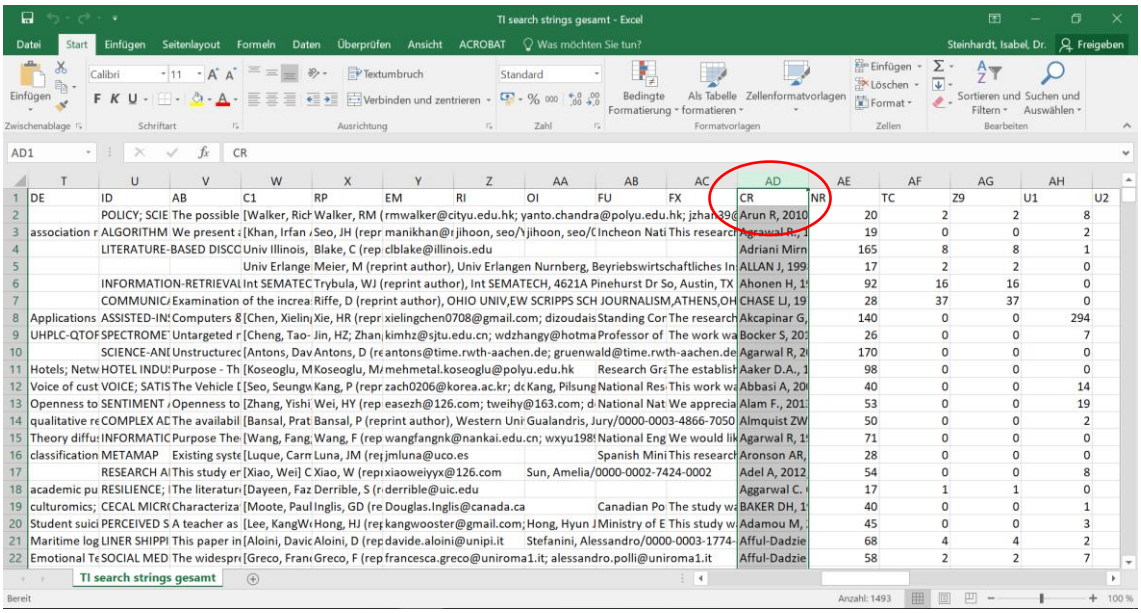


55	13265	fayyad u, 1996, ai mag, v17, p37	-0.4325	0.0687	2	33	41	11	
56	13272	fayyad u., 1996, adv knowledge discov	-0.0257	0.1277	2	9	12	11	
57	13314	fei-fei l, 2005, proc cvpr ieee, p524	-0.8769	-0.1764	3	41	78	10	
58	13327	feinerer i, 2008, j stat softw, v25, p1	-0.3768	0.1036	2	52	100	28	
59	13409	feldman r., 2006, text mining hdb adv	-0.2917	0.0483	2	32	36	12	
60	13411	feldman r., 2007, text mining hdb adv	-0.0277	0.0648	2	60	107	38	
61	13435	fellbaum christiane, 1998, wordnet elect lexica	0.7429	0.0066	1	41	50	13	
62	13586	ferrucci d., 2004, natural language engineering, p327, doi 10.1017/s135132490400352	11.591	-0.0728	1	41	63	10	
84	18851	hoffman m., 2010, adv neural inform pr, p856	-12.567	-0.0518	3	44	87	14	
85	18870	hoffmann r, 2004, nat genet, v36, p664, doi 10.1038/ng0704-664	1.266	0.013	1	46	88	16	
86	18871	hoffmann r, 2005, bioinformatics, v21, p252, doi 10.1093/bioinformatics/bti1142	11.454	0.0206	1	58	103	16	
87	18888	hofmann t, 1999, sigir'99: proceedings of 22nd international conference on research	-0.9748	-0.1087	3	98	390	59	
88	18889	hofmann t, 1999, uncertainty in artificial intelligence, proceedings, p289	-0.9045	-0.0049	3	71	131	18	
89	18890	hofmann t, 2001, mach learn, v42, p177, doi 10.1023/a:1007617005950	-0.8074	-0.0282	3	69	157	26	
90	19079	hong l., 2010, p 1 worksh soc med a, p80, doi 10.1145/1964858.1964870	-13.908	-0.0631	4	56	201	22	
104	22783	kim jd, 2008, bmc bioinformatics, v9, doi 10.1186/1471-2105-9-10	10.408	0.0088	1	53	83	13	
105	23901	krallinger m, 2005, genome biol, v6, doi 10.1186/gb-2005-6-7-224	10.762	0.011	1	51	76	12	
106	23907	krallinger m, 2008, genome biol, v9, doi [10.1186/gb-2008-9-s2-s1 10.1186/gb-2008-9	12.157	-0.0054	1	67	146	22	
107	23906	krallinger m, 2008, genome biol, v9, doi 10.1186/gb-2008-9-s2-s8	12.718	-0.0184	1	46	72	18	
108	24020	krippendorff k., 2004, content anal intro i	-0.5109	0.1307	3	34	55	10	
109	24694	landauer tk, 1997, psychol rev, v104, p211, doi 10.1037/0033-295x.104.2.211	-0.5984	0.1295	2	45	71	12	
110	24695	landauer tk, 1998, discourse process, v25, p259, doi 10.1080/01638539809545028	-0.5223	0.1043	2	65	120	24	

Die Datei am besten zwischenspeichern aber für den nächsten Schritt offen lassen.

Schritt 2.25

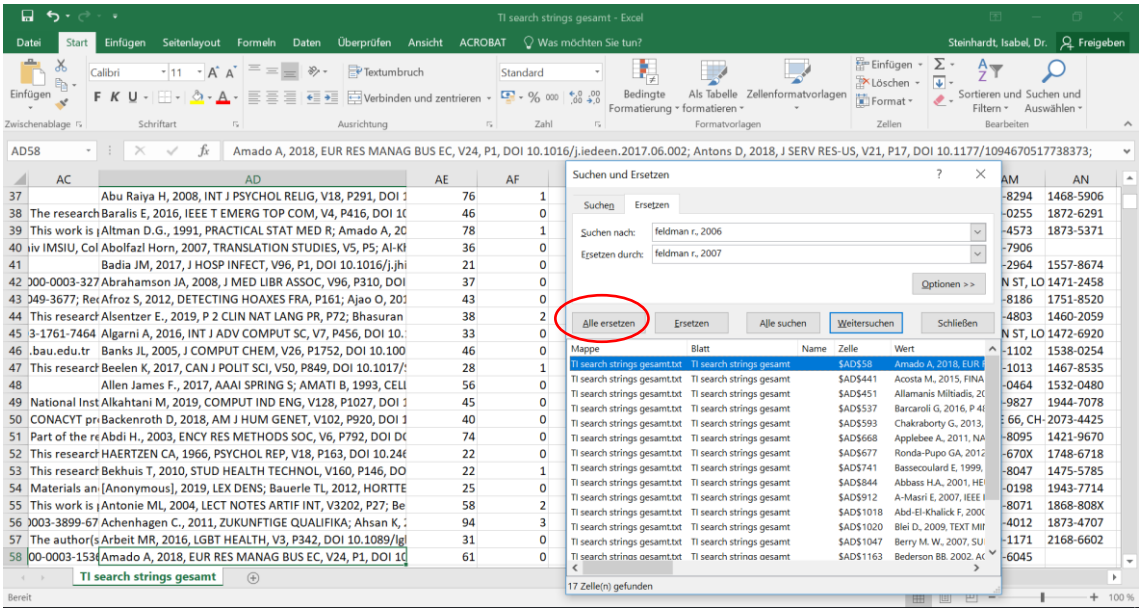
Um die fehlerhaften Angaben durch die korrekten Angaben zu ersetzten muss zunächst die txt-Datei wieder in Excel geöffnet werden. Im vorliegenden Fall also die Datei „TI search strings gesamt“. Bitte die Schritte 2.1-2.4 wiederholen. Liegt Euch die Excel-Tabelle wieder vor, dann die Spalte CR (*Cited References*) markieren (das ist in der Excel-Spalten-Bezeichnung AD).



Schritt 2.26

Als nächstes können dann mit dem Befehl „Suchen und Ersetzen“ die fehlerhaften Angaben ersetzt werden. Dazu einfach durch STRG F den Befehl „Suchen und Ersetzen“ aufrufen. In das Feld „Suche nach“ trägt man je eine fehlerhafte Angabe ein, die bei Schritt 2.24 ermittelt wurden. Bei „Ersetzen durch“ kopiert man die jeweilige korrekte Angabe aus Schritt 2.24 hinein. Dann kann man sich unter „Alle suchen“ zunächst die Stellen anschauen, die fehlerhaft sind, bevor auf „Alle ersetzen“ geklickt wird, um die Bereinigung durchzuführen.

Dieser Vorgang muss für alle Fehler, die gefunden wurden, wiederholt werden.



Nun ist der Datensatz bereinigt und die Analysen können durchführt werden, z. B. unter Verwendung des VosViewer.

ACHTUNG: Die Datei wieder als Unicode abspeichern, sonst kann VosViewer die Daten nicht lesen. WICHTIG ist auch, dass keine Spalten eingefügt oder Spalten umbenannt werden!