READ
Recognition and Enrichment of Archival Documents

Suche in Dokumenten mit Keyword Spotting

Version v1.9.1
Letzte Aktualisierung dieses Guides: 28.11.2019

Dieser Guide erklärt die Verwendung der Keyword Spotting Funktion (KWS) in Transkribus. KWS ist ein neues und effektives Werkzeug, das es Ihnen ermöglicht, Ihre Dokumente nach bestimmten Wörtern zu durchsuchen.

Laden Sie den Transkribus Expert Client herunter oder stellen Sie sicher, dass Sie die neueste Version verwenden:

- https://transkribus.eu/

Besuchen Sie das Transkribus Wiki für mehr Informationen und weitere How to Guides:

- https://transkribus.eu/wiki/

Transkribus und die zugrundeliegende Technologie werden durch die folgenden Projekten und Plattformen verfügbar gemacht:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Kontakt

- Das Transkribus Team: email@transkribus.eu

Inhalt

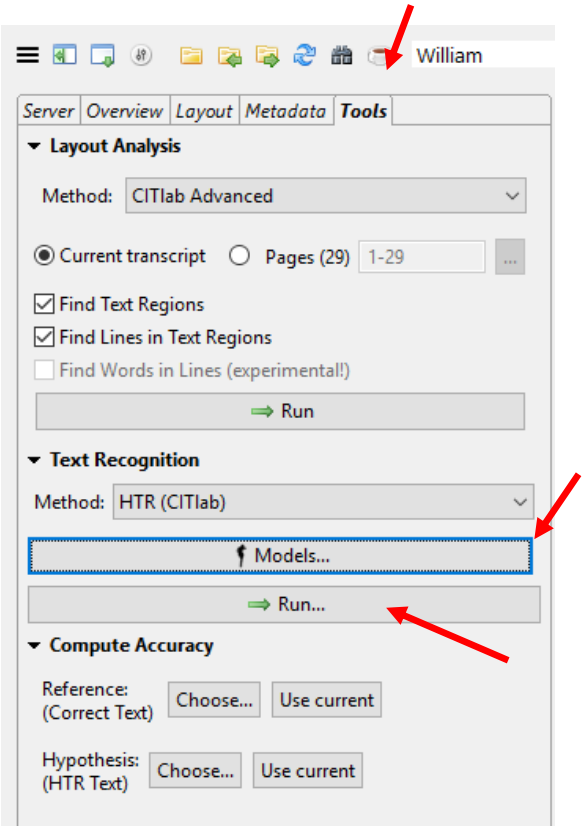
Das READ Projekt wird durch das Horizon 2020 Forschungs- und Innovationsprogramm im Rahmen des Fördervertrags Nr. 674943 finanziert.

Einleitung

- Keyword Spotting (KWS) ermöglicht es, **Ihre Dokumente nach bestimmten Wörtern zu durchsuchen**.
- **Der Hauptvorteil: die Dokumente müssen vor der Suche NICHT transkribiert werden.** Verwenden Sie einfach ein bereits vorhandenes Handwritten Text Recognition (HTR) Modell um ein Transkript zu erstellen und Ihr Dokument dann sofort zu durchsuchen. Selbst wenn das automatisch erstellte Transkript Fehler enthält, findet KWS verlässlich Wörter, Phrasen, und Teile von Wörtern und regulären Ausdrücken in Ihren Dokumenten.
- Das Programm zeig Ihnen, **auf welchen Seiten das Keyword gefunden wurde**. Außerdem erhalten Sie einen Wert zwischen 0 und 1, um die Richtigkeit des Resultats einzuschätzen.
- **Achtung:** KWS ist ein Vorgang, der in Transkribus als "Job" implementiert ist. Das heißt Sie können die Suche starten und andere Arbeiten erledigen, während Sie auf das Suchergebnis warten. Alle Resultate werden auf Transkribus gespeichert und können jederzeit (wieder) geöffnet werden. Wir sind davon überzeugt, dass die Arbeit mit Suchresultaten von KWS in Zukunft Standard für viele Historiker und Philologen sein wird.

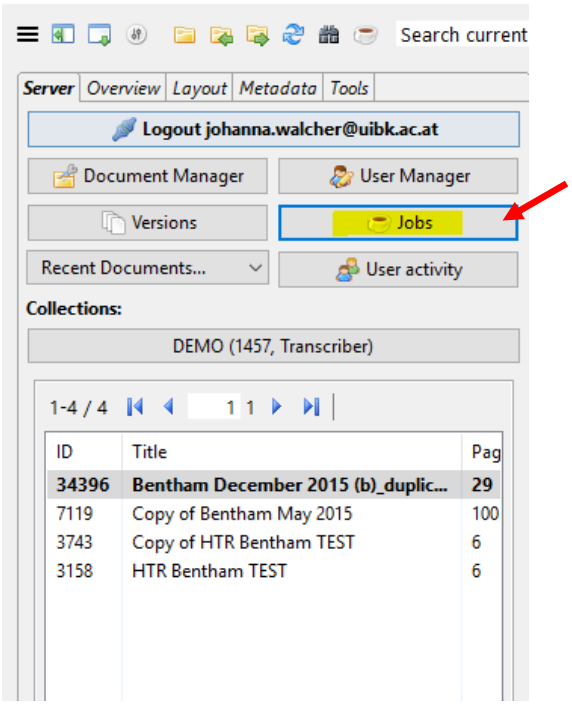
Vorbereitung - Texterkennung

- Bevor Sie KWS verwenden können, müssen Sie ein HTR Modell über Ihre Dokumente laufen lassen, um eine Transkription zu generieren.
- Als erstes laden Sie Ihre Dokumente auf Transkribus hoch.
- Danach segmentieren Sie Ihre Dokumente in Text Regions, Lines und Baselines.
- Für mehr Informationen zum Hochladen und Segmentieren, konsultieren Sie bitte die Anleitung Transkribieren in Transkribus.
- Dann lassen sie ein HTR-Modell über Ihre Dokumente laufen.
- Um das Modell zu verwenden, klicken Sie auf den Reiter "Tools" und gehen Sie zum Bereich "Text Recognition".
- Klicken Sie auf "Run", dann auf "Configure". Wählen Sie ein HTR Modell (linke Bildschirmseite) aus und klicken Sie auf "OK".
- Achtung: KWS funktioniert am besten, wenn Sie ein spezielles HTR Modell verwenden, das auf Basis Ihrer Dokumente trainiert wurde. Brauchbare Resultate sind aber auch mit allgemeinen Modellen erreichbar. Wenn Sie noch kein eigenes HTR Modell haben, können Sie mit den öffentlichen Modellen auf Transkribus experimentieren. Transkribus stellt im Moment öffentliche Modelle auf Basis von englischen und deutschen Handschriften aus dem 18. und 19. Jahrhundert zur Verfügung. Mehr öffentliche Modelle werden bald verfügbar sein.
- Klicken Sie auf "Run" um die Texterkennung zu starten.

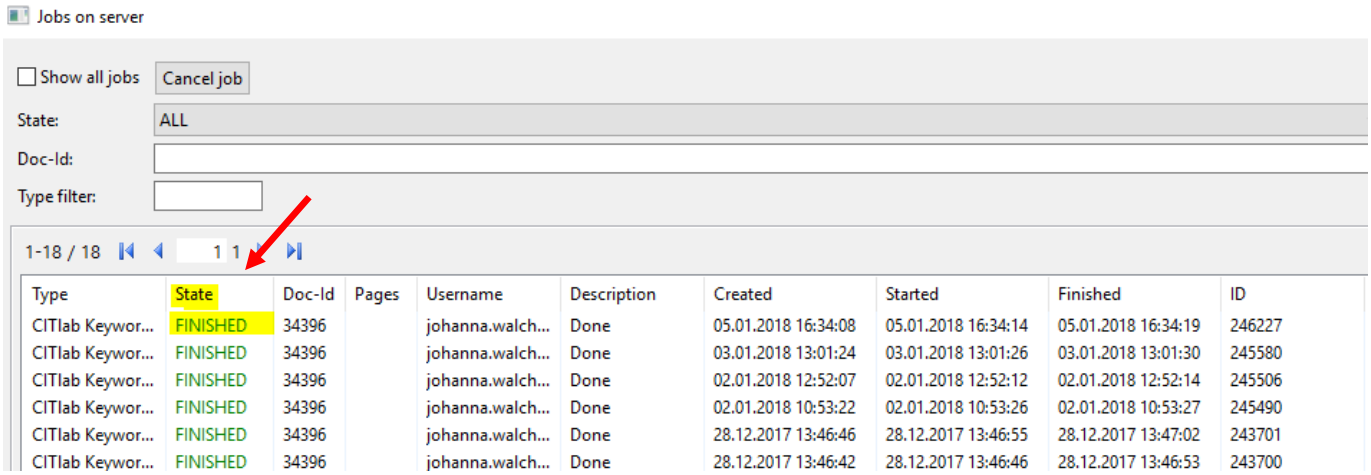


Darstellung 1 HTR Modell starten

- Sie können den Fortschritt der Texterkennung verfolgen, indem Sie auf die Schaltfläche „Jobs“ im „Server“ Tab klicken.

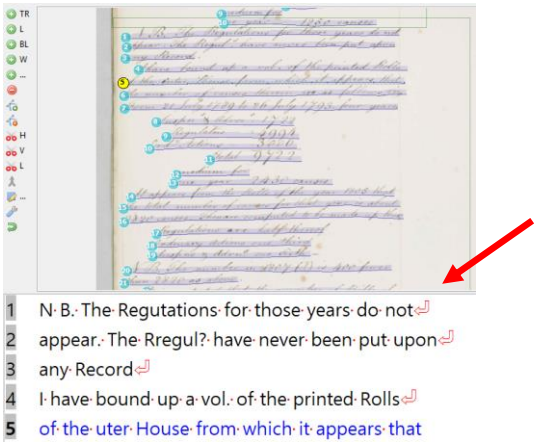


Darstellung 2 Klicken Sie auf “Jobs” um den HTR Fortschritt zu überprüfen.



Darstellung 3 "Jobs on server" Fenster

- Ist die Texterkennung abgeschlossen, erscheint die automatische Transkription im Text Editor.

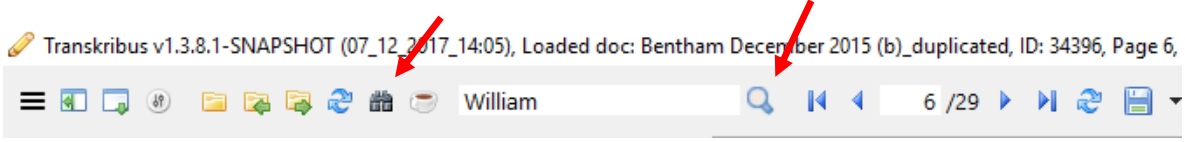


Darstellung 4 Automatisches Transkript erscheint im Text Editor.

Keyword Spotting Funktion

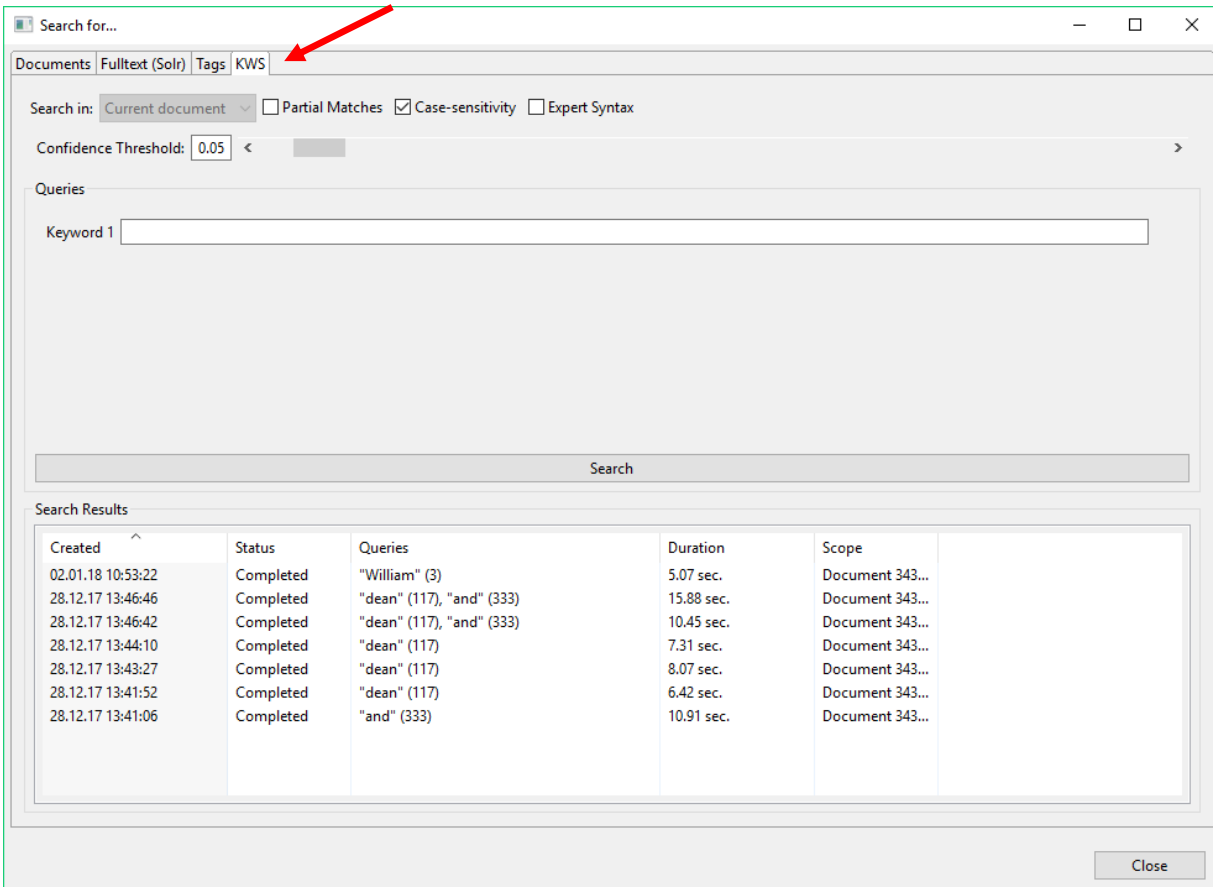
Wo finde ich KWS?

- Sie können das Keyword Spotting Feature öffnen, indem Sie auf die Fernglas-Schaltfläche oder die Lupe im Hauptmenu klicken.



Darstellung 5 Keyword Spotting öffnen

- Klicken Sie auf den Reiter KWS im Suchfenster



Darstellung 6 "KWS" Bereich

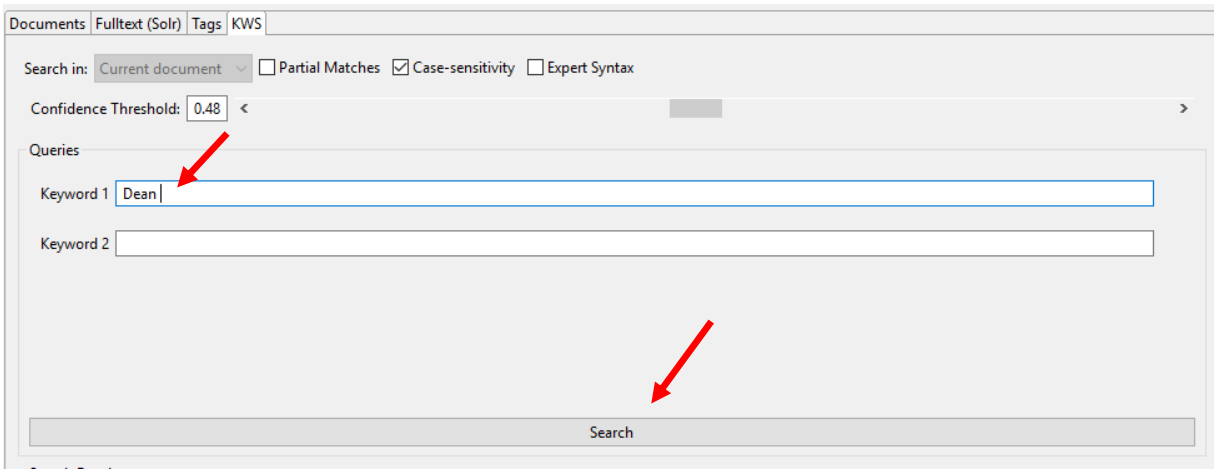
Suchresultate

- Ihre früheren und aktuellen Keyword Spotting Suchanfragen sind unten im „KWS" Tab zu sehen.

Darstellung 7 Aktuelle und frühere Suchresultate

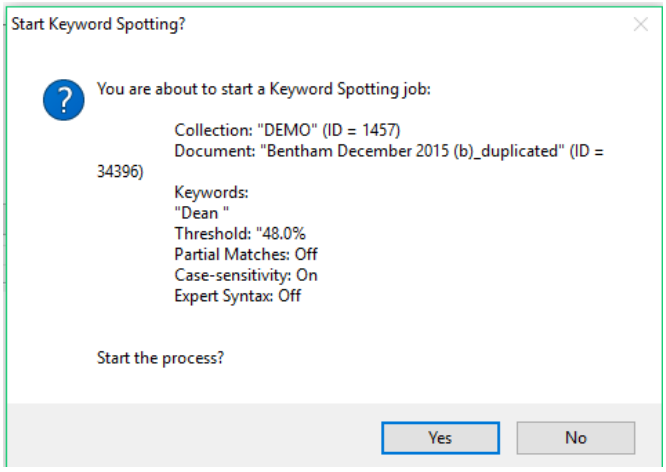
Keyword Spotting verwenden

- Um Keyword Spotting zu verwenden, tippen Sie einfach das Wort, nach dem Sie suchen möchten in das Feld "Keyword 1" und klicken Sie auf „Search".



Darstellung 8 Nach einem Keyword suchen

- Ein Bestätigungsfenster öffnet sich. Klicken Sie “Yes” um Ihre Keyword Spotting Suche zu starten.

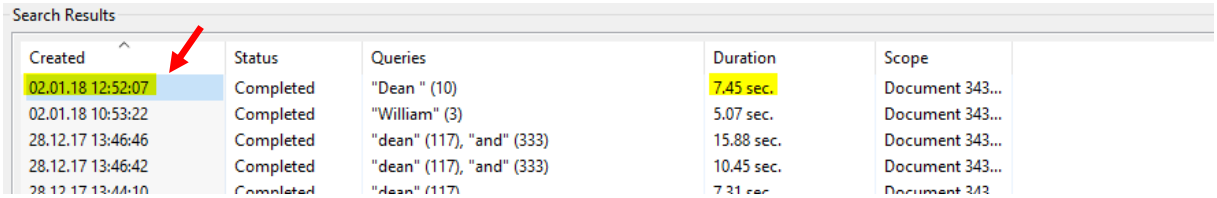


Darstellung 9 Bestätigungsfenster

- Keyword Spotting Anfragen benötigen mindestens einige Sekunden.
- Wenn im“KWS” Tab in der Spalte “Duration” “N/A” angezeigt wird, sucht das Programm noch.
- Wenn der Prozess abgeschlossen ist, wird unter “Duration” die Zeit angezeigt, die die Anfrage benötigt hat.

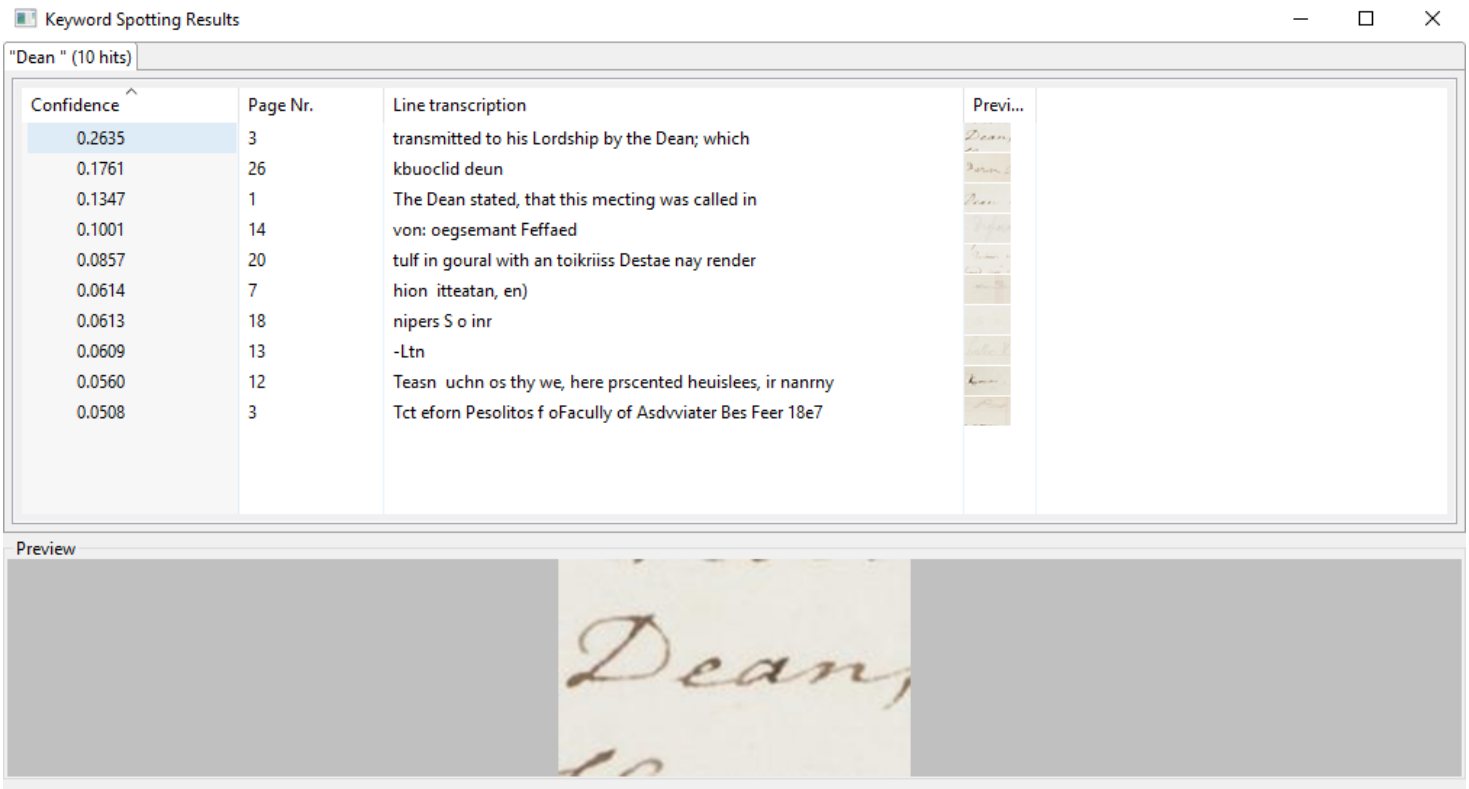
Darstellung 10 Laufendes Keyword Spotting

- Doppelklicken Sie auf Datum und Uhrzeit in der Spalte „Created“ um zu Ihren Suchergebnissen zu gelangen.



Darstellung 11 Keyword Spotting Ergebnisse

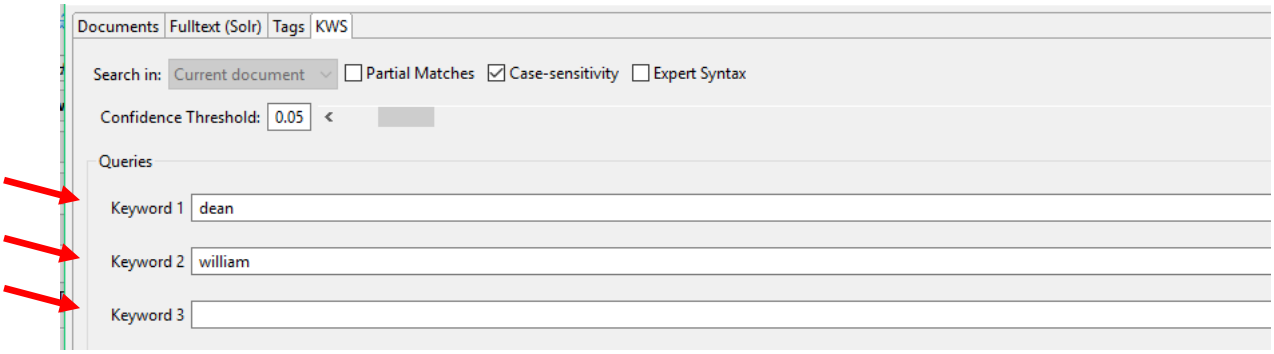
- Das “Keyword Spotting Results” Fenster zeigt Ihnen eine Liste von Orten, an denen das Keyword gefunden wurde, gemeinsam mit folgenden Informationen:
  - Die „Confidence“ (Genauigkeit) des jeweiligen Resultats (zwischen 0 und 1).
  - Die Seitenzahl im Dokument.
  - Die Transkription des Wortes.
  - Ein Bildausschnitt der Seite. Wenn Sie mit Ihrem Cursor über das Bild streichen, erscheint ein größeres Vorschaubild am unteren Rand des Fensters.
  - Doppelklicken Sie auf das Bild in der Spalte “Preview” um direkt zur Seite mit dem Keyword zu gelangen.



Darstellung 12 Informationen zu Ihren Keyword Spotting Ergebnissen

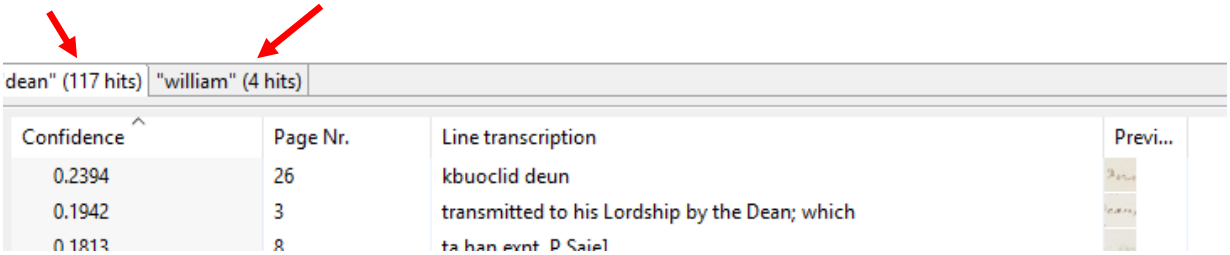
Zwei Keywords gleichzeitig suchen

- Es ist auch möglich gleichzeitig nach zwei oder mehr Keywords zu suchen. Fügen Sie einfach die zusätzlichen Keywords in die Felder im „KWS“ Tab ein.



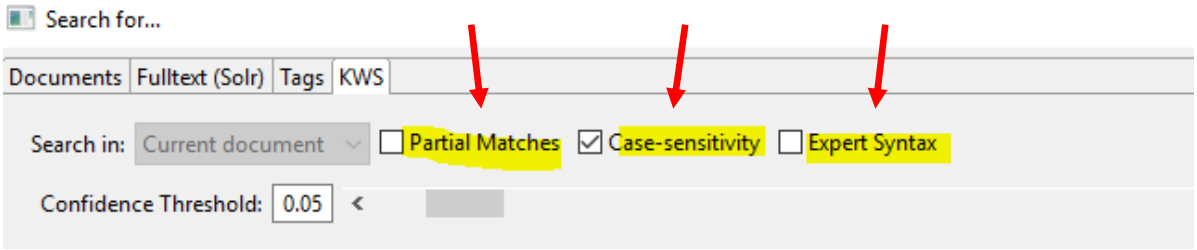
Darstellung 13 Mehrere Keywords gleichzeitig suchen

- Die Resultate werden für jedes Keyword in einem separaten Tab angezeigt.



Darstellung 14 Suchresultate für mehrere Keywords

Weitere Suchoptionen



Darstellung 15 Weitere Suchoptionen

Partial Matches (Teilentsprechungen)

- Wenn Sie diese Option auswählen, sucht das Programm nach allen Wörtern die den Text, den Sie ins Suchfeld eingegeben haben, enthalten. Wenn Sie zum Beispiel nach „**ity**“ suchen, erhalten Sie Ergebnisse wie „**conditionality**“, „**proportionality**“, **indefensibility**“, usw.

Case Sensitivity

- Mit dieser Funktion werden Groß- und Kleinschreibung berücksichtigt. Wenn Sie zum Beispiel nach “Kingdom” suchen, werden Resultate bei denen “Kingdom” mit großem “K” geschrieben ist, vorgezogen.

Expert Syntax

- Anstatt nach Wörtern könne Sie auch nach Ausdrücken suchen.
- Einige Beispiele für Ausdrücke sind:
  - **Datum:** .*(?<KW>[0-3][0-9]\.[0-1][0-9]\.[0-9]{4}).* zeigt jede Zeile die ein Datum im Format TT.MM.JJJJ enthält
  - **Abkürzungen:** .*(?<KW>Dr\.|Doctor).* zeigt jede Zeile die Doctor und die Abkürzung Dr. enthält

  - **Unsicherheiten:** .*(?<KW>(k|c|che|chh)rist?).* zeigt jede Zeile die althochdeutsche Schreibweisen von Christus wie z.B. kris, krist, crist, cherist, chhrist enthält

- Im Gegensatz zu normalen Ausdrücken müssen Suchmuster der ganzen Zeile entsprechen, z.B. .*[0-9]{4,6} zeigt nur Zeilen an, die mit einer Zahl mit mindestens 4 Stellen enden. Um beliebige Zeichen nach 4 Stellen zu erlauben, muss * am Ende hinzugefügt werden: .*[0-9]{4,6}.*
  Dementsprechend zeigt, [0-9]{4,6}.* nur Zeilen, die mit einer vierstelligen Zahl beginnen.

- Standardausdrücke die vom KWS in Transkribus unterstützt werden:

  . jedes Zeichen

  \+ eine weitere Wiederholung des vorherigen Symbols

  * null oder mehr Wiederholungen des Vorherigen Symbols

  [] eine Klasse von Zeichen, z.B. [0-9] jede Zahl zwischen 0 und 9; [aeiou] jeder Vokal; [A-Z] jeder Großbuchstabe

  ? das vorherige Symbol ist optional

  {X} wiederhole das vorherige Symbol X-mal

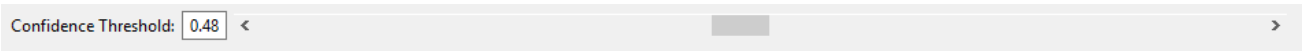
  {X,Y} wiederhole das vorherige Symbol zwischen X- und Y-mal

  | oder Abfrage, z.B. a|b bedeutet entweder a oder b

  () Klammern werden verwendet, um Standardausdrücke zu gruppieren: (a|b)c passt zu ac oder bc während a|bc zu a oder bc passt

Confidence Threshold (Verlässlichkeitswert)

- Es ist möglich den Confidence Threshold Ihrer Keyword Spotting Suche anzupassen.
  Achtung: diese Funktion ist noch nicht aktiviert, steht aber bald zur Verfügung.
- Die Confidence Threshold ist eine Zahl zwischen 0 und 1.
- Wenn die Confidence Threshold 0,5 oder höher ist, bedeutet das, dass das Programm mit hoher Wahrscheinlichkeit Keywords findet, die Ihrer Suchanfrage entsprechen.
- Wenn die Confidence Threshold 0,1 oder niedriger ist, bedeutet das, dass das Programm mehr mögliche Ergebnisse findet, diese aber eine höhere Fehlerwahrscheinlichkeit haben. Es liegt in diesem Fall an Ihnen, die Suchergebnisse auf Ihre Richtigkeit zu überprüfen.



Darstellung 16 Confidence Threshold

Ausblick

Das Transkribus Team arbeitet im Moment an einem Update der KWS Funktion. Die zukünftige Version wird es Nutzern ermöglichen Ihre Resultate zu validieren und die Suchresultate in Tabellen zu exportieren.

Danksagung

Wir möchten den vielen Nutzern danken, die mit ihrem Feedback zur Verbesserung der Transkribussoftware beigetragen haben.

Transkribus wird der Öffentlichkeit im Rahmen des H2020e Infrastrukturprojekts READ (Recognition and Enrichment of Archival Documents) zugänglich gemacht. Dieses Projekt wird von der Europäischen Kommission im Rahmen des Fördervertrags Nr. 674943 finanziert