2.4 API für das Suchen mit SEMPRIA-Search

Wenn man SEMPRIA-Search als Software-as-a-Service (Saas) betreiben möchte, kann man die in diesem Abschnitt beschriebene API verwenden. Man benötigt die Information, wo der SaaS erreichbar ist. In diesem Dokument nehmen wir die folgende fiktive *Service-URL*: https://finde.maschine/ . Jede URL für die API muss mit dieser Service-URL beginnen. Nach dem abschließenden Fragezeichen folgen ein oder mehrere Parameter, deren Bedeutung und Format im Folgenden beschrieben wird. Reale Service-URLs enthalten oft eine Portnummer, z.B. https://finde.maschine:1234/ Alle Service-URLs verwenden das sichere Protokoll HTTPS (statt HTTP), so dass weder die Anfragen noch die Ergebnisseiten von Dritten mitgelesen werden können. (Dies kann nur gewährleistet werden, solange der Computer, von dem aus die API angesprochen wird, nicht kompromittiert ist, also nicht in der Sicherheit durch Sicherheitslücken, Hackerangriffe o.ä. eingeschränkt ist.) Weitergehende Verschlüsselungen können individuell vereinbart werden.

Falls der Zugang Passwort-geschützt vereinbart wurde, müssen an jede URL ein Parameter &pw=IhrPasswort und ein Parameter &corpus=IhreCorpusId angehängt werden. Falls ein IP-beschränkter Zugang eingerichtet wurde, so muss die IP des anfragenden Computers in der beim SaaS-Server hinterlegten Whitelist enthalten sein.

2.4.1 Suchen (API-Kommando sempria-search)

2.4.1.1 Suchanfrage (sentence) Die eigentliche Suchanfrage ist der zentrale Parameter der API und darf nicht fehlen. Beispielsuche nach Oboe, in allen Wortformen und möglichen semantisch verwandten Formulierungen: https://finde.maschine/sempria-search?sentence=Oboe

Die restlichen Parameter sind optional für den Aufruf der Suchmaschine. Je nachdem welche Metadaten Ihre Dokumentenkollektion enthält, ist nur ein Teil der Parameter sinnvoll.

2.4.1.2 Autor (author) Treffer müssen zusätzlich zur Suchanfrage auch die Bedingung erfüllen, dass der Autor des Textes mit dem des Parameterwerts übereinstimmt. Beispiel: https://finde.maschine/sempria-search?sentence=Oboe&author=Max

2.4.1.3 Zeitliche Einschränkung (date-begin und date-end) Die Treffer werden eingeschränkt bezüglich ihres Schreibedatums (date). Dazu kann date-begin, date-end oder beides verwendet werden. Die Datumsangabe muss in der kompakten Variante von ISO 8601 geschehen, also 20191231 für den 31. Dezember 2019. Die angegebenen Werte sind inklusiv zu verstehen, so dass das folgende Beispiel bedeutet, dass Dokumente vom 31. Dezember 2019 oder danach gesucht sind: https://finde.maschine/sempria-search?sentence=Oboe&date-begin=20191231

2.4.1.4 Adressat (addree) Das Adressaten-Feld ist meist nur in Dokumentenkollektionen mit Kommunikationscharakter sinnvoll besetzt. Solche Kollektionen können Mailsammlungen, Briefsammlungen o.ä. sein. Im folgenden Beispiel sind nur Treffer mit dem Empfänger Max erwünscht: https://finde.maschine/sempria-search?sentence=Oboe&addree=Max

2.4.1.5 Stichwortphrasen (keytopics) Falls ihre Dokumentenkollektion automatisch von SEMPRIA-Search berechnete (oder manuell gepflegte) Stichwortphrasen enthält, so kann der Parameter keytopics verwendet werden. Leerzeichen innerhalb einer Phrase müssen als Unterstrich geschrieben werden. Beispiel: https://finde.maschine/sempria-search?sentence=Oboe&keytopics=San_Remo

2.4.1.6 Zeichensatzkodierung Die API verwendet per Default UTF-8 für die Eingabe und die Ausgabe. Andere Kodierungen können auf Nachfrage über spezielle Parameter eingestellt werden oder bei der Dokumentenkollektion hinterlegt werden.

2.4.1.7 Ausgabeformat Die Ergebnisse einer Suchanfrage mit den oben beschriebenen Parametern werden in einem HTML-Format geliefert. Andere Format als html kann man über den Parameter oformat anfordern. Es werden unterstützt: opensearch (ein XML-Format für Suchergebnisse, s. https://github.com/dewitt/opensearch), json (JSON-Format), s-exp (s-expressions, s-Audrücke).

2.4.1.8 Trefferformat Das inhaltliche Format der einzelnen Treffer wird momentan beim Einrichten der Suchmaschine festgelegt. Es kann noch nicht als Parameter übergeben werden.

2.4.2 Testzugang per Browser (API-Kommando sempria)

Sofern für die Dokumentenkollektion freigeschaltet, findet man ein einfaches Webformular zum Abschicken von Suchanfragen unter folgender URL: https://finde.maschine/sempria? Es dient meist dem Testen oder Debuggen der Suchmaschine.

2.4.3 Hilfetext (API-Kommando sempria-help)

Zu jeder SEMPRIA-Search als SaaS gibt es einen Hilfetext unter folgender URL: https://finde.maschine/sempria-help?corpus=IHRE-CORPUS-ID&format=htmlfrag Als Ergebnis bekommt man ein HTML-Fragment zum Einbinden an geeigneter Stelle. Geplante weitere Formate: html und txt.

2.5 API für die Systemverwaltung von SEMPRIA-Search

2.5.1 Index-Update (API-Kommando sempria-update-index)

Der Index der Suchmaschine wird in den meisten Fällen in regelmäßigen, vereinbarten Abständen aktualisiert. Wenn eine Echtzeit-Aktualisierung mit Sitemap vereinbart wurde, kann die Aktualisierung mit dem Aufruf sempria-update-index asynchron angestoßen werden. Die Laufzeit der Aktualisierung hängt stark vom Umfang der geänderten Daten ab. Beispiel-Aufruf: https://finde.maschine/sempria-update-index?corpus=IHRE-CORPUS-ID Dieser API-Aufruf ist ohne die Parameter pw und corpus nicht möglich. Der Rückgabewert ist 0, falls die Aktualisierung erfolgreich angestoßen wurde, sonst der Fehlercode 1.

2.5.2 Index-Stand (API-Kommando sempria-indexed)

Das Kommando sempria-indexed liefert die URLs aller momentan indexierten Dokumente zurück. Beispiel-Aufruf: https://finde.maschine/sempria-indexed?corpus=IHRE-CORPUS-ID Dieser API-Aufruf ist ohne die Parameter pw und corpus nicht möglich. Der optional Parameter oformat beschreibt das Ausgabeformat. Unterstützt werden hier html und htmlfrag, geplant sind json und s-exp.