Technology
Arts Sciences
TH Köln

Institut für Informationswissenschaft

Fakultät für Informations- und Kommunikationswissenschaften

Prof. Dr. Klaus Lepsky
klaus.lepsky@th-koeln.de

Automatisches Erschließen [BdK 6.2 – DIS 25b]

Tutorial

Oktober 2022

Modulinhalte

Das Modul „Automatisches Erschließen (BdK 6.2 – DIS 25b)“ vermittelt durch theoretische Inhalte und eine praktische Aufgabenstellung die Grundlagen für linguistisch und statistisch arbeitende Verfahren der automatischen Indexierung. Dazu gehören: Möglichkeiten und Grenzen automatischer Erschließungsverfahren; Linguistisch basierte Verfahren; Statistisch basierte Verfahren; Gewichtungsverfahren; Retrievaltests; Automatische Klassifizierung; Clustering.

In der praktischen Aufgabenstellung werden mit eigenen Daten unterschiedliche Techniken und Verfahren zur automatischen Indexierung bzw. Erschließung angewendet. Dazu gehört die Erstellung einfacher Stichwortlisten und Register aus Volltexten mit einem Textverarbeitungsprogramm. Der Zusammenhang zwischen Zeichenketten in Dokumenten und Indextermen für das Retrieval wird untersucht. Es werden Indexierungsläufe für eine vorgegebene Dokumentkollektion durchgeführt. Die erzielten Ergebnisse werden analysiert und miteinander in Bezug auf Retrievalverbesserungen verglichen.

Dieses Tutorial orientiert sich an der praktischen Aufgabenstellung für das Modul und dient als Basis für ein Selbststudium der Lehrinhalte.

Inhaltsverzeichnis

1 Einführung in die Thematik

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 5.1 des Kapitels 5 des Lehrbuchs Informationserschließung und Automatisches Indexieren.[1]

Beachten Sie: Im Verlauf des Tutorials werden in den „Lektüre“-Abschnitten immer wieder Hinweise auf Abschnitte des Lehrbuchs gegeben, die für den entsprechenden Teil des Arbeitsprogramms relevant sind. Dies kann dazu führen, dass spätere Abschnitte des Kapitels 5 zu einem eher frühen Zeitpunkt vorgeschlagen werden. Sollten Ihnen in diesen Fällen Fachbegriffe noch unbekannt sein, können Sie das „Sachverzeichnis“ auf den S. 427ff. verwenden, das für die im Buch verwendeten Fachbegriffe die Stellen nachweist, an denen diese behandelt werden. Zusätzlich kann für die begriffliche Klärung von Fachterminologie der Thesaurus Informationserschließung herangezogen werden.

1.1 Einrichten der Datenbank „literatur.dbm“ (Midos 6)

Für die Arbeit an der praktischen Aufgabenstellung wird eine bereits eingerichtete Arbeitsumgebung mit einer Datenbank und einer Umgebung für die Automatische Indexierung zur Verfügung gestellt.

Laden Sie den *zip*-Ordner „lit-lingo.zip“ hier herunter. Entpacken Sie den *zip*-Ordner und speichern Sie den Ordner „lit-lingo“ auf der Hauptebene der Festplatte `c:\lit-lingo`. Der Ordner „lit-lingo“ enthält zwei Unterverzeichnisse: „lingo-work“ mit einer vorbereiteten Arbeitsumgebung für die Indexierung mit *Lingo* (Abschnitt 3ff.) und „midos-literatur“ mit einer *Midos*-Datenbank mit bibliografischen Referenzdaten:

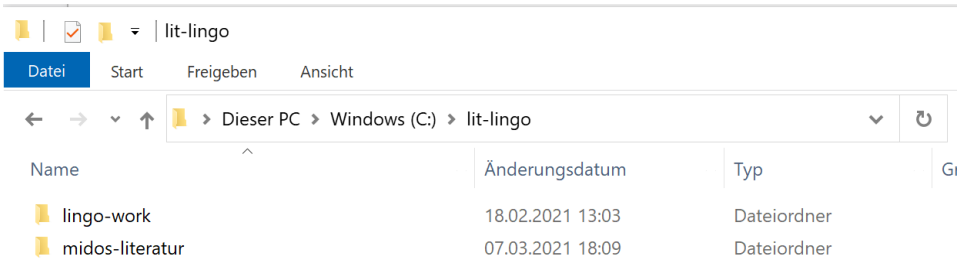

Abbildung 1: Verzeichnisstruktur „lit-lingo“

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 8.1 des Kapitels 8.

[1] Winfried Gödert, Klaus Lepsky und Matthias Nagelschmidt, *Informationserschließung und Automatisches Indexieren : ein Lehr- und Arbeitsbuch*, X.media.press (Berlin [u.a.]: Springer, 2012). Das Lehrbuch kann als PDF innerhalb des Hochschulnetzes (bzw. von außerhalb über eine VPN-Verbindung) kostenlos heruntergeladen werden.

Midos 6 installieren

Laden Sie das Datenbanksystem *Midos 6* herunter und installieren es auf Ihrem Rechner: Entpacken Sie dazu die Archivdatei und führen Sie „setup.exe“ aus.

Datenbank „literatur.dbm“ öffnen

Starten Sie *Midos* und öffnen Sie über „Datenbank – Öffnen“ die Datenbank „literatur.dbm“ im Verzeichnis `c:\lit-lingo\midos-literatur`. Nach dem Navigieren zum Ordner der Datenbank und dem Öffnen der Datenbank wird im linken Frame der Datenbanksoftware ein Lesezeichen für die Datenbank angelegt, um diese zukünftig schnell öffnen zu können. Die Datenbank enthält 100 Datensätze zum Themenbereich „Automatisches Indexieren“ aus der Datenbank Literatur zur Informationserschließung.

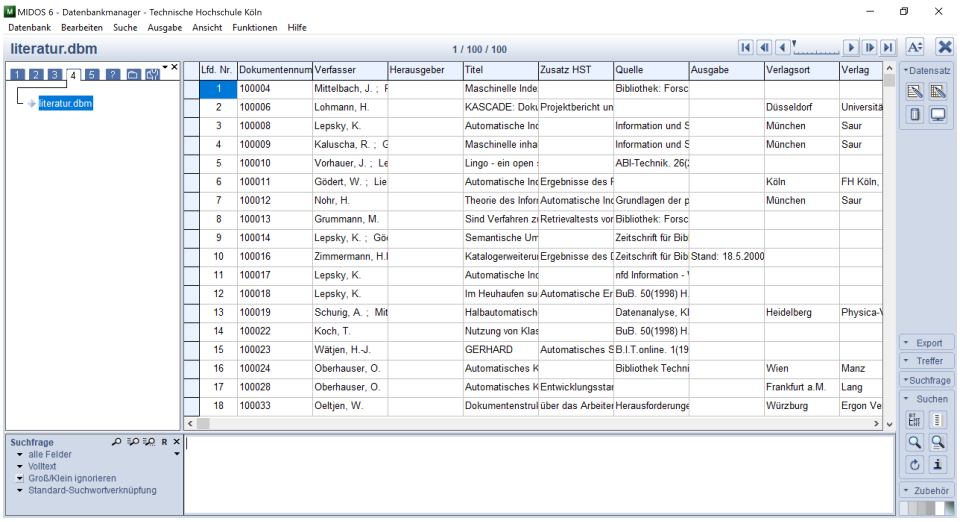

Abbildung 2: *Midos 6 – Datenbankmananger*

1.2 Erfassen eigener Dokumente in der Datenbank (Midos 6)

Die Datenbank soll um 10 Dokumente erweitert werden.

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 8.2 des Kapitels 8.

Auswahl eines Datensets

Laden Sie die Datei rohdaten-001-050.zip herunter und wählen Sie eines der nummerierten 50 enthaltenen Datensets. Jedes Set enthält die bibliografischen Daten von 10 Zeitschriftenaufsätzen aus der Datenbank Literatur zur Informationserschließung.

Tragen Sie Ihre Auswahl in der Form „Nummer des Datensets | Matrikelnummer“ in diese Tabelle ein: bdk62-dis25b-auswahl-des-datensets-2023w

Achtung: Bereits gewählte Datensets können nicht mehr gewählt werden! Die Auswahl des Datensets ist bindend für die Bearbeitung des Lernportfolios und damit Bestandteil der Modulprüfung.

Erfassen der Datensätze in der Datenbank

Erfassen Sie mit dem *Datenbankeditor* von *Midos 6* die Daten zu den 10 Zeitschriftenaufsätzen.

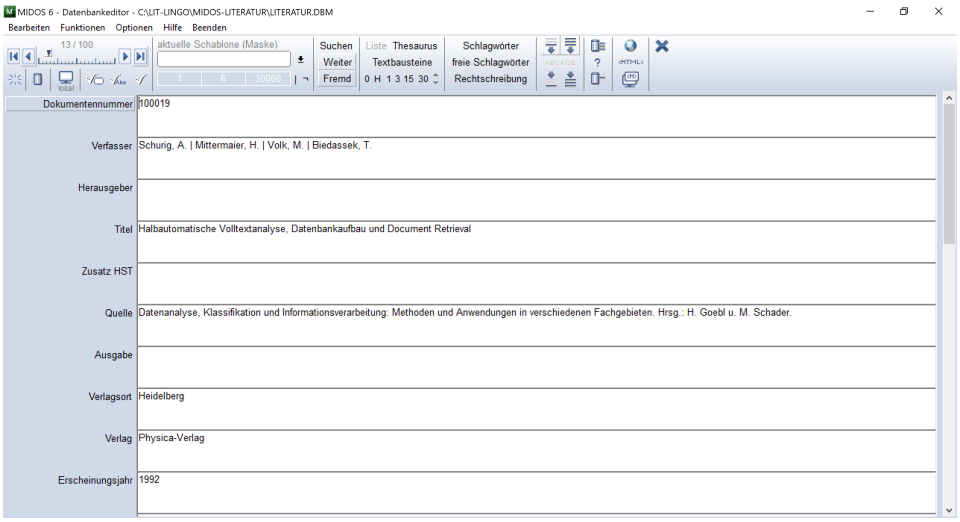

Abbildung 3: *Midos 6 – Datenbankeditor*

Beachten Sie: Das Erfassen von Daten in einer Datenbank folgt immer dem Prinzip der Einheitlichkeit. Orientieren Sie sich bei der Erfassung an den Erfassungsprinzipien, die sich aus den 100 bereits in der Datenbank vorhandenen Datensätzen bzw. aus den Dokumentbeschreibungen in der Datenbank Literatur zur Informationserschließung ableiten lassen.

Lernportfolio – Teil 0 Exportieren Sie die 10 von Ihnen erfassten Dokumentbeschreibungen im Ausgabeformat „Vollanzeige“ und übernehmen Sie die 10 Dokumentbeschreibungen in den „Anhang“ Ihres Lernportfolios (keine Screenshots).

1.3 Automatische Indexerstellung

Die in Abschnitt 5.1.2 des Kapitels 5 beschriebene Vorgehensweise soll mit dem Titel und dem Abstract eines eigenen Datensatzes durchgeführt werden.

Kopieren Sie dazu Titel und Abstract einer der 10 von Ihnen in der Datenbank erfassten Dokumentbeschreibungen in eine Textverarbeitung (z. B. *Word* oder *LibreOffice*). Erzeugen Sie über geeignete „Suche-und-Ersetze“-Befehle eine Liste aller Einzelwörter aus dem Text.

Analysieren Sie das Ergebnis. Achten Sie insb. auf Auffälligkeiten, die dadurch verursacht wurden, dass Sie für Ihre „Suche-und-Ersetze“-Vorgänge zu *schwache* Regeln für die Begrenzung von Zeichenketten verwendet haben.

Versuchen Sie das Ergebnis durch eine Erweiterung dieser Regeln zu verbessern.

Lernportfolio – Teil 1 Nehmen Sie die Dokumentbeschreibung und das Ergebnis Ihrer Umwandlung in eine Zeichenketten-Liste in Ihr Lernportfolio auf (keine Screenshots). Beschreiben Sie Ihre Vorgehensweise und machen Sie deutlich, welche Regelergänzungen Sie vorgenommen haben und warum Sie diese für sinnvoll erachten.

1.4 Intellektuelle Registererstellung

Die in Abschnitt 5.1.3 des Kapitels 5 beschriebene Vorgehensweise soll mit dem Titel und dem Abstract eines eigenen Datensatzes durchgeführt werden.

Kopieren Sie dazu Titel und Abstract der in Abschnitt 1.3 verwendeten Dokumentbeschreibung in eine Textverarbeitung (z. B. *Word* oder *LibreOffice*). Markieren Sie dabei die von Ihnen für sinnvoll erachteten Verzeichniseinträge im Text und verwenden Sie die jeweilige Funktion für die Festlegung von Verzeichniseinträgen:

- in *Word* durch „Verweise – Eintrag festlegen“; dort lässt sich durch „Haupteintrag“ und „Untereintrag“ ein hierarchischer Eintrag erzeugen; falls die Ansicht für Feldfunktionen im Dokument aktiv ist, werden die erzeugten Verzeichniseinträge in geschweiften Klammern im Klartext im Text angezeigt; dort können sie auch verändert/erweitert werden;
- in *LibreOffice* durch „Einfügen – Verzeichnis – Verzeichniseintrag“; dort unter „Verzeichnis“ den Typ „Stichwortverzeichnis“ auswählen; unter „Eintrag“ lässt sich der zum markierten Term gehörende Verzeichniseintrag festlegen (dieser muss nicht dem Wort im Text entsprechen); mit „1. Schlüssel“ und „2. Schlüssel“ können hierarchische Einträge erstellt werden.

Orientieren Sie sich bei der Auswahl und Festlegung der Verzeichnis- oder Indexeinträge allein an dem möglichen Nutzen, den diese für eine Suche versprechen. Relevant sind also nur solche Terme, die einen inhaltlichen Bezug zum Dokument haben.

Wenn Sie alle Verzeichniseinträge markiert haben, können Sie unterhalb der Dokumentbeschreibung das daraus generierte Stichwortverzeichnis erzeugen:

- in *Word* durch „Verweise – Index einfügen“;
- in *LibreOffice* durch „Einfügen – Verzeichnis – Verzeichnis“; dort als Typ „Stichwortverzeichnis“ wählen.

Bei Überarbeitungen der ausgewählten Verzeichniseinträge müssen die entsprechenden Verzeichnisse dann nur noch aktualisiert werden.

Lernportfolio – Teil 2 Nehmen Sie die Dokumentbeschreibung und das von Ihnen erzeugte Stichwortverzeichnis in Ihr Lernportfolio auf (keine Screenshots). Beschreiben Sie, welche Kriterien Sie für die Auswahl der Registerbegriffe angewendet haben und warum Sie diese für sinnvoll erachten. Vergleichen Sie die Eigenschaften des Stichwortverzeichnisses und der erzeugten Termliste aus Abschnitt 1.3 in Bezug auf ihre Eignung für Suchprozesse und halten sie charakteristische Unterschiede fest.

2 Automatische Schlagwortvergabe (Midos 6)

Die 110 Datensätze der Datenbank sollen mit der *Midos*-Funktion „Automatische Schlagwortvergabe“ automatisch erschlossen werden.

Lektüre Lesen Sie zur Einführung in die Thematik Abschnitt 5.2 des Kapitels 5.

2.1 Automatische Schlagwortvergabe

Führen Sie für alle Datensätze eine „Automatische Schlagwortvergabe“ durch. Orientieren Sie sich bei der Vorgehensweise an der Beschreibung in Abschnitt 5.2 des Kapitels 5.

Midos verwendet für die Automatische Schlagwortvergabe eine sog. Synonymliste „synonym.txt“, die durch einen Export von Synonymrelationen eines Thesaurus entstanden ist (vgl. Thesaurus Informationserschließung). Dabei handelt es sich um eine reine Textdatei, die auf jeder Zeile am Zeilenanfang jeweils den Deskriptor und dahinter, durch Semikolon voneinander getrennt, ein oder mehrere Synonyme enthält. Diese muss für die Funktionen der Automatischen Schlagwortvergabe zunächst in eine sog. *Midos*-Indexdatei umgewandelt werden.

Öffnen Sie dazu über „Bearbeiten – Automatische Schlagwortvergabe“ den Dialog für die Konfigurierung und Ausführung der Automatischen Schlagwortvergabe (vgl. Abbildung 4). Wählen Sie über den Button „txt > wtx“ die Synonymliste „synonym.txt“ aus dem Datenbankverzeichnis aus; diese wird dann in eine *Midos*-Indexdatei mit der Dateiendung „wtx“ umgewandelt und als „Positivliste“ geladen.

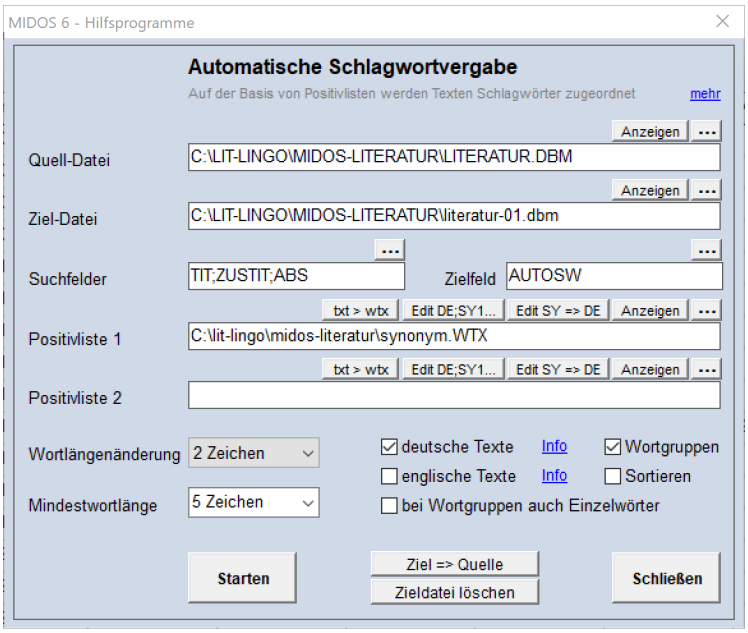

Abbildung 4: *Midos 6 – Dialog Automatische Schlagwortvergabe*

Für die Durchführung der Automatischen Schlagwortvergabe sind jetzt noch die Angaben von „Quelldatei“ („literatur.dbm“) und „Zieldatei“ („literatur-01.dbm“) erforderlich; *Midos* unterstellt für automatische Prozesse immer einen Workflow, bei dem die Originaldatei („Quelldatei“) unverändert bleibt und die Ergebnisse des Prozesses in einer Kopie der Datenbank abgelegt werden („Zieldatei“). Diese Kopie liegt im selben Verzeichnis und ist zweckmäßigerweise durch eine Zählung vom Original zu unterscheiden, hier „literatur-01.dbm“.

Die Leistungsfähigkeit der „Automatischen Schlagwortvergabe“ hängt ab von den im Einstellungsdialog festgelegten „Suchfeldern“. Hier sollten ausschließlich Felder ausgewählt werden, die einen *inhaltlichen* Bezug zum Dokument haben. Die für die automatische Schlagwortvergabe zu durchsuchenden Felder werden in „Suchfeld“, das Feld für die Ablage der erzeugten Schlagwörter in „Zielfeld“ vereinbart. Wählen Sie als Suchfelder die Felder für Titel und Abstract: „TIT“, „ZUSTIT“, „ABS“. Wählen Sie als Zielfeld das Feld „AUTOSW“.

Hinweis: Je nach eingesetztem Rechner kann die Verschlagwortung einige Zeit beanspruchen.

Öffnen Sie nach Beendigung der Automatischen Schlagwortvergabe über „Datenbank – Öffnen“ die neue erzeugte Datenbank „literatur-01.dbm“. Dies ist nun Ihre aktuelle Arbeits-Datenbank.[2]

Lernportfolio – Teil 3 Nehmen Sie die Dokumentbeschreibung für eines der von Ihnen erfassten Dokumente mit den erzeugten Schlagwörtern in Ihr Lernportfolio auf (kei-

[2]Die aktuelle Arbeitsdatenbank wird im Arbeitsprogramm noch mehrfach geändert. Belassen Sie bitte die Lesezeichen der älteren Versionen in der Liste, um bei möglicherweise unerwünschten Resultaten zu früheren Versionen der Datenbank zurückkehren zu können.

ne Screenshots). Analysieren Sie das Ergebnis der Automatischen Schlagwortvergabe und untersuchen Sie insb., welche Wörter in der Dokumentbeschreibung zu welchen Schlagwörtern geführt haben. Geben Sie auch für nicht übereinstimmende Ergebnisse (beispielsweise Plural im Text → Schlagwort im Singular) an, welche Gründe zur Erzeugung des Schlagworts geführt haben. Beziehen Sie den Inhalt der Synonymliste und die „Wortlängenänderung“ in Ihre Betrachtung ein.

3 Automatisches Indexieren I – Grundformerzeugung, Wortklassenerkennung, Kompositumerkennung (Lingo)

Lektüre Lesen Sie zur Einführung in die Thematik die Abschnitte 5.3 bis 5.3.3 des Kapitels 5.

3.1 Durchführen einer Testindexierung mit Lingo-Web

Ein erster Eindruck einer linguistisch basierten automatischen Indexierung mit *Lingo* lässt sich sehr einfach durch die Verwendung einer Web-Version von *Lingo* gewinnen. Öffnen Sie die Webseite Lingo-Web und geben in das linke Feld „Input“ ein deutschsprachiges Abstract aus den von Ihnen erstellten Dokumentbeschreibungen ein (vgl. Abbildung 5). Durch den Button „Start processing“ wird der Text automatisch indexiert und das Ergebnis der Verarbeitung erscheint im rechten Fenster. Dort wird die Protokolldatei des Indexierungslaufs mit *Lingo* angezeigt, in der für alle Zeichenketten des Abstracts die durchgeführten Operationen zeilenweise aufgeführt sind.

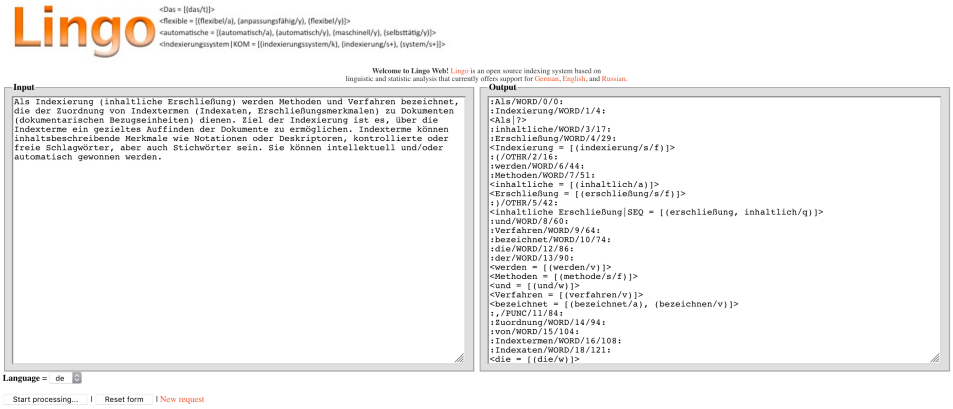

Abbildung 5: Lingo-Web

Lernportfolio – Teil 4 Übernehmen Sie das verwendete Abstract und das Indexierungsergebnis in Ihr Lernportfolio (keine Screenshots). Analysieren Sie das Ergebnis der Indexierung unter Verwendung der Legende links unten („Legend“). Verschaffen Sie sich ein Bild vom Indexierungsergebnis anhand der verwendeten Kürzel bzw. des Markups für die erkannten Bestandteile des Quelltextes und zeigen Sie anhand einzelner Zeilen der Protokolldatei die realisierten Funktionen (für jede Funktion nur ein Beispiel).

3.2 Einrichten und Test einer eigenen Lingo-Arbeitsumgebung

Falls noch nicht geschehen, muss zunächst *Ruby* auf Ihrem System installiert werden und anschließend *Lingo*. In Abhängigkeit vom verwendeten Betriebsystem gibt es dafür unterschiedliche Wege; für *Windows* ist die Installation in Abschnitt 10.2 ausführlich beschrieben.

Lingo verfügt über keine grafische Benutzeroberfläche. Sämtliche Programmabläufe werden über die Kommandozeile gestartet. Für die Arbeit mit *Lingo* muss eine eigene Arbeitsumgebung (in Form eines eigenen Verzeichnisses) vorhanden sein.

Diese ist im Ordner `/lit-lingo` bereits eingerichtet und befindet sich im dortigen Unterverzeichnis `/lingo-work`.

Im Verzeichnis `/lingo-work` befinden sich vier Unterverzeichnisse:

`/config` enthält mitgelieferte Konfigurationsdateien;
`/dict` enthält benutzerspezifische Wörterbücher;
`/lang` enthält sprachspezifische Konfigurationsdateien;
`/txt` enthält zu indexierende Daten und die Ergebnisdateien.

Für die Indexierung im Rahmen dieses Moduls sollen speziell angepasste Konfigurationsdateien verwendet werden. Diese befinden sich im Verzeichnis `/config` der *Lingo*-Arbeitsumgebung.

Testen Sie Ihre *Lingo*-Arbeitsumgebung mit der bereits vorhandenen Textdatei „artikel.txt“ im Unterverzeichnis `/txt` von `/lingo-work` und der Konfigurationsdatei „lingo.cfg“ mit folgendem Befehl (Achtung: der Befehl muss aus dem Verzeichnis `/lingo-work` aufgerufen werden!):

```
$ lingo -c lingo.cfg txt/artikel.txt
```

Die Indexierung der Datei „artikel.txt“ ist erfolgreich beendet worden, wenn folgender Hinweis auf der Kommandozeile ausgegeben wird (der Pfad kann natürlich individuell verschieden sein):

```
/.../lingo-work/txt/artikel.txt: progress done.
```

Alle Ergebnisdateien einer Indexierung mit *Lingo* werden ebenfalls im Verzeichnis `/txt` abgelegt; sie beginnen immer mit dem Dateinamen der Quelldatei (hier „artikel“) und unterscheiden sich nur hinsichtlich ihrer Dateiendungen.

Auf gängigen Betriebssystemen sind Dateiendungen in der Regel mit zugehörigen Programmen verknüpft, um durch einen Doppelklick direkt die „richtige“ Anwendung starten zu können. Für die Analyse der Ergebnisdateien einer *Lingo*-Indexierung (wie auch allgemein für das reine Ansehen von Dateien mit beliebigen Dateiendungen) ist dieses Verhalten nicht hilfreich. Es empfiehlt sich daher, für die Arbeit mit *Lingo* die Installation eines alternativen Dateimanagers mit integriertem Dateibetrachter, beispielsweise den Dateimanager *Double Commander* (vgl. Abbildung 6).

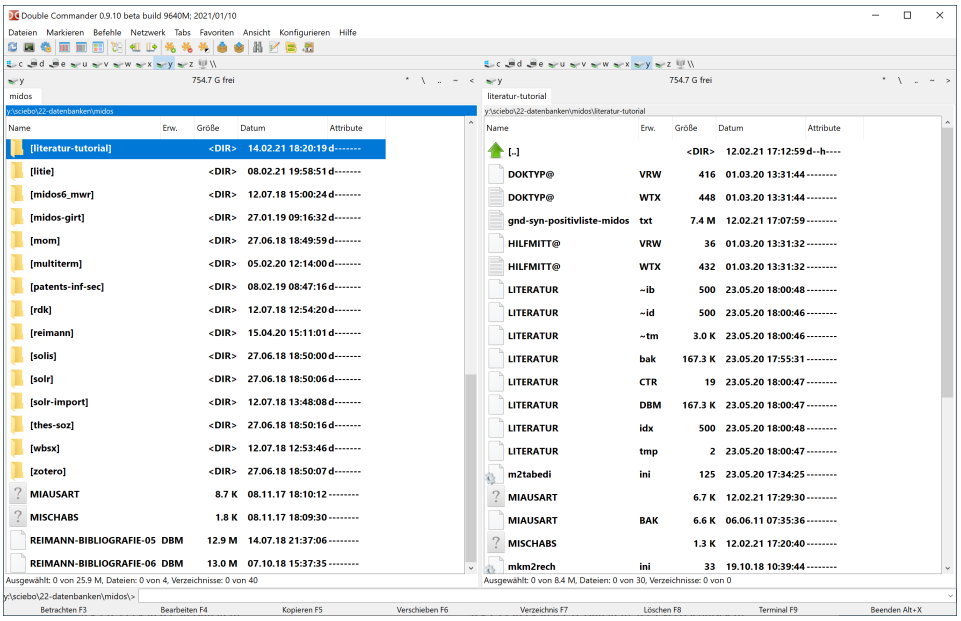

Abbildung 6: Double Commander

Im Doppelfenster dieses Dateimanagers steht über „F3“ eine Funktion zum reinen Betrachten (nur-Lese-Modus) von beliebigen Dateien zur Verfügung, wodurch die Analyse der *Lingo*-Ergebnisdateien sehr vereinfacht wird.

Lernportfolio – Teil 5 Sehen Sie sich alle erzeugten Ergebnisdateien an (Dateien im Verzeichnis `/txt` deren Dateinamen mit „artikel“ beginnen). Erstellen Sie eine Liste der unterschiedlichen Dateitypen und geben an, welchen Inhalt sie jeweils haben und nach welchem Kriterium sie sortiert sind. Nehmen Sie die Dateiliste in Ihr Lernportfolio auf (keine Screenshots).

3.3 Durchführung einer ersten Indexierung mit den Datensätzen der Datenbank

Die Testindexierung mit der Datei „artikel.txt“ diente in erster Linie dem Test der Funktionalität von *Lingo* und einer ersten Gewöhnung an das „Handling“. Bei der Datei handelt es sich um den Volltext eines Zeitschriftenaufsatzes.

Für eine Indexierung der Datensätze Ihrer Datenbank müssen Sie die Daten zunächst mit einem geeigneten Ausgabeformat aus *Midos* exportieren. Dieses ist bereits vorbereitet unter dem Namen „lingo-export“. Es extrahiert aus den Datensätzen die „Identnummer“ zur eindeutigen Identifizierung eines Datensatzes (in eckigen Klammern und mit einem Punkt dahinter, um sie nicht durch die Indexierung zu verarbeiten und ggf. unkenntlich zu machen) und den Inhalt der Kategorien „Titel“ (TIT), „Zusatztitel“ (ZUSTIT) und „Abstract“ (ABS):

```
[0123456789.]
Text Text Text
```

Exportieren Sie alle Datensätze der Datenbank mit dem Ausgabeformat „lingo-export“ in eine Textdatei. Verwenden Sie dazu die Funktion „Ausgabe – Datei“ – diese Funktion gibt es im Hauptmenü des Datenbankmanagers und in der Vollanzeige. Stellen Sie sicher, falls Sie aus dem Datenbankmanager heraus die Funktion gewählt haben, dass das *aktuelle* Ausgabeformat das Format „lingo-export“ ist. Aktivieren Sie im Ausgabe-Dialog die Option „UTF-8“. *Midos* schlägt für die Exportdatei den Namen „export.txt“ vor, den wir beibehalten.

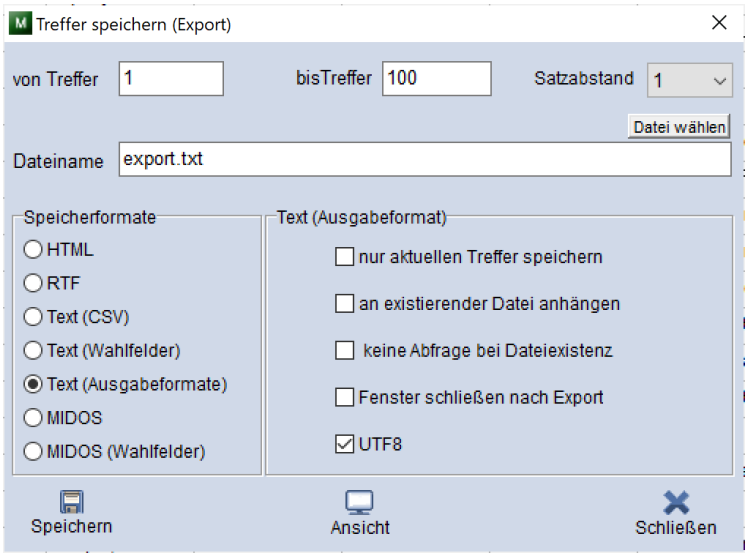

Abbildung 7: Midos – Ausgabe einer Datei

Kopieren Sie für die Indexierung mit *Lingo* die exportierte Datei aus dem Verzeichnis `/midos-literatur` in das Verzeichnis `/txt` des *Lingo*-Arbeitsverzeichnisses `/lingo-work`. Starten Sie die Indexierung mit:

```
$ lingo -c lit-lingo-lem.cfg txt/export.txt
```

Lernportfolio – Teil 6 Übernehmen Sie die ersten zehn Einträge der Ergebnisdateien „export.ven“ und „export.non“ in Ihr Lernportfolio (keine Screenshots). Untersuchen Sie die Einträge hinsichtlich der Fragestellungen (1) und (2) und halten Sie die Ergebnisse im Lernportfolio fest.

(1) Beurteilen Sie für die Einträge der „export.ven“-Datei den möglichen Nutzen für ein Retrieval. Warum sind diese zehn Einträge die häufigsten?

(2) Warum sind die zehn häufigsten Einträge aus der „export.non“-Datei nicht von *Lingo* erkannt worden?

4 Automatisches Indexieren II – Semantische Analyse (Lingo)

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 5.3.4 des Kapitels 5.

Die Realisierung der grammatikalischen Grundfunktionalitäten – Grundformerzeugung, Kompositumerkennung und die damit verbundene Wortklassenerkennung – ist Voraussetzung für Verfahren, die über das einzelne Wort hinausgehen. Mit „semantischer Analyse“ sind hier Funktionen gemeint, die zusammengehörige Wörter in einem gemeinsamen Bedeutungszusammenhang betrachten. *Lingo* stellt dafür zwei Module zur Verfügung, die auf unterschiedliche Art und Weise eine Erkennung von sog. „Mehrwortgruppen“ erlauben. Der „sequencer“ identifiziert auf der Basis der Wortklassen-Taggings des „wordsearchers“ Muster von Wortfolgen, beispielsweise alle im Text vorkommenden Adjektiv-Substantiv-Verbindungen. Der „multiworder“ identifiziert Mehrwortgruppen auf der Basis von Einträgen eines spezifischen Mehrwort-Wörterbuchs. Beide Module haben ihre jeweils eigene Berechtigung und können sinnvoll auch gemeinsam eingesetzt werden: So kann es etwa für eine neue Dokumentkollektion zunächst nützlich sein, mit dem „sequencer“ alle darin enthaltenen Mehrwortgruppen zu identifizieren und auf der Basis dieser „Vorab-Analyse“ die für die Kollektion wichtigsten Mehrwortgruppen in ein Wörterbuch zu übernehmen, um für zukünftige Identifizierungsläufe nur noch „erwünschte“ Mehrwortgruppen zu identifizieren.

Führen Sie eine Indexierung der Datensätze Ihrer Datenbank mit den beiden zusätzlichen Attendees „sequencer“ und „multiworder“ durch. Verwenden Sie wieder die Datei „export.txt“ für die Indexierung. Verwenden Sie die Konfigurationsdatei „lit-lingo-sem.cfg“:

```
$ lingo -c lit-lingo-sem.cfg txt/export.txt
```

Lernportfolio – Teil 7 Übernehmen Sie die ersten zehn Einträge der Ergebnisdateien „export.seq“ und „export.mul“ in Ihr Lernportfolio (keine Screenshots). Untersuchen Sie die Einträge hinsichtlich der Fragestellungen (1) und (2) und halten Sie die Ergebnisse im Lernportfolio fest.

(1) Beurteilen Sie für die Einträge der „export.seq“-Datei deren möglichen Nutzen als Einträge für ein Mehrwort-Wörterbuch. Warum sind diese zehn Einträge die häufigsten?

(2) Beurteilen Sie für die Einträge der „export.mul“-Datei deren Nutzen für eine Dokumenterschließung.

5 Automatisches Indexieren III – Wörterbucharbeit (Lingo)

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 5.3.7 des Kapitels 5.

Ein wörterbuchbasiertes System zur automatischen Indexierung – wie *Lingo* – hängt in seiner Leistungsfähigkeit entscheidend von der Qualität der in den Wörterbüchern vorhandenen Terminologie ab. Für den Einsatz des Systems, insb. wenn fachlich spezifische Dokumentkollektionen indexiert werden sollen, ist daher eine Erweiterung der Wörterbuchbasis um nicht vorhandene Fachterminologie und eine laufende Pflege der Wörterbücher nötig.

Dies geschieht nicht in den mit dem System ausgelieferten Wörterbüchern, sondern in eigens dafür angelegten Benutzerwörterbüchern. Diese Benutzerwörterbücher befinden sich im Verzeichnis `/dict/de` des Arbeitsverzeichnisses `/lingo-work`.

Wählen Sie aus der „export.non“-Datei Ihres letzten Indexierungsergebnisses jeweils 5 Einzelwörter und Komposita aus, die nicht erkannt wurden, deren Identifizierung Sie aber für nützlich halten würden. Legen Sie für diese Terme Einträge im Benutzerwörterbuch „user-dic.txt“ an, die zu einer Identifizierung führen. Indexieren Sie Ihre Daten erneut, um den Effekt zu testen und zu bestätigen. Überprüfen Sie das Ergebnis anhand der aktualisierten „export.non“-Datei.

Lernportfolio – Teil 8 Übernehmen Sie die 10 von Ihnen ausgewählten Terme in das Lernportfolio und begründen Sie die Auswahl. Übernehmen Sie auch die von Ihnen erzeugten Einträge des Benutzerwörterbuchs und die Zeilen der „export.log“-Datei in das Lernportfolio (keine Screenshots).

6 Automatisches Erschließen mit GND-Daten I (Midos 6)

Für eine automatische Erschließung der Datensätze der Datenbank existiert eine Synonymliste für *Midos* mit den Sachschlagwörtern der GND und ihren jeweiligen Synonymen. Die Synonymliste umfasst mehr als 200.000 Zeilen in der Version als reine Textdatei (vgl. „gnd-syn-midos.txt“ im Datenbankverzeichnis) und aufbereitet als „*.wtx“-Datei für *Midos* mehr als 20 MB.

Führen Sie eine Automatische Schlagwortvergabe mit *Midos* für die Datensätze Ihrer Datenbank durch. Verwenden Sie dafür die Positivliste „gnd-syn-midos.wtx“ und als „Suchfelder“ Titel und Abstract („TIT“, „ZUSTIT“, „ABS“). Für die Ergebnisse der Verschlagwortung wählen Sie das Feld „AUTOGND“ aus der Datenbeschreibung.

Wählen Sie als „Quelldatei“ für die Automatische Schlagwortvergabe „literatur-01.dbm“ (sollte bereits voreingestellt sein, da „literatur-01.dbm“ Ihre aktuelle Datenbank ist) und als Zieldatei „literatur-02.dbm“.

Hinweis: Je nach eingesetztem Rechner kann die Verschlagwortung einige Zeit beanspruchen (30 Min. und mehr).

Öffnen Sie nach Beendigung der Automatischen Schlagwortvergabe über „Datenbank – Öffnen“ die Datenbank „literatur-02.dbm“. Dies ist nun Ihre aktuelle Arbeits-Datenbank.

Lernportfolio – Teil 9 Übernehmen Sie eine der von Ihnen erfassten Dokumentbeschreibungen und den zugehörigen Inhalt der Kategorie „AUTOGND“ in Ihr Lernportfolio (keine Screenshots). Analysieren Sie das Ergebnis der Automatischen Schlagwortvergabe und machen Sie anhand konkreter Beispiele deutlich, welche Ergebnisse Sie aus der Sicht einer Dokumenterschließung als nützlich und welche Ergebnisse Sie als problematisch betrachten.

7 Automatisches Erschließen mit GND-Daten II (Lingo)

Die in Abschnitt 6 verwendete Synonymliste mit GND-Schlagwörtern lässt sich auch für eine Indexierung mit *Lingo* verwenden. Dazu wurde diese in ein *Lingo*-Synonym-Wörterbuch vom Typ „KeyValue“ konvertiert, in dem jeder Eintrag auf einer Zeile mit einem Synonym beginnt, gefolgt von einem Trennzeichen „*“ und der Vorzugsbenennung, d. h. dem GND-Schlagwort (vgl. „wtx“-Datei von *Midos* und die Datei „gnd-syn-lingo.txt“ im Verzeichnis `/dict/de`. Die Indexierung mit dem GND-Vokabular benötigt die Konfigurationsdatei „lit-lingo-gnd.cfg“:

```
$ lingo -c lit-lingo-gnd.cfg txt/export.txt
```

Lernportfolio – Teil 10 Übernehmen Sie die ersten zehn Einträge der Ergebnisdatei „export.syn“ in Ihr Lernportfolio (keine Screenshots). Untersuchen Sie die Einträge hinsichtlich der Fragestellungen (1) und (2) und halten Sie die Ergebnisse im Lernportfolio fest.

(1) Warum sind diese zehn Einträge die häufigsten?

(2) Welchen Nutzen haben die zehn häufigsten Einträge für eine Dokumenterschließung?

8 Automatische Indexierung von Datensätzen (Midos 6, Lingo)

Lektüre Lesen Sie zur Einführung in die Thematik den Abschnitt 5.3.6 des Kapitels 5.

Alle bisher erzeugten Teilergebnisse sollen abschließend als Inhalte in Feldern der *Midos*-Datenbank vorhanden sein. Für die Ergebnisse der Automatischen Schlagwortvergabe ist das vergleichsweise einfach, weil die Inhalte von *Midos* selbst übernommen werden. Für die Übernahme der von *Lingo* erzeugten Indexierungsergebnisse ist ein Import in die *Midos*-Datenbank erforderlich.

Zunächst ist jedoch ein erneuter Indexierungslauf nötig, um Ergebnisdateien zu erzeugen, die für einen Import in eine Datenbank geeignet sind. Für die Indexierung von sehr vielen Datensätzen aus einer Datenbank wird ein Modus benötigt, bei dem Datensatz-spezifisch indexiert wird, d. h. die Indexate müssen dem jeweiligen Datensatz, zu dem sie gehören, zugeordnet werden. Eine Datensatz-spezifische Indexierung wird mit der Konfigurationsdatei „lit-lir.cfg“ erreicht:

```
$ lingo -c lit-lir.cfg txt/export.txt
```

Die Indexierungsergebnisse werden in einem *csv*-konformen Dateiformat mit zwei Spalten abgelegt, bei denen auf jeder Zeile ein Datensatz steht mit der Identnummer in der ersten Spalte und den Indexaten in der zweiten Spalte (als Spalten- oder Feldtrenner wird das Zeichen „*“ verwendet, mehrere Einträge in einem Feld sind durch Pipe („|“) getrennt):

```
1*text|text|text
```

Dieses Format kann von *Midos* importiert werden. Der Import erfolgt in mehreren Schritten – Konvertierung der Ergebnisdateien von UTF-8 zu ANSI; Umwandlung der Ergebnisdateien in *Midos*-Datenbankdateien; Importieren der *Midos*-Datenbankdateien mit den *Lingo*-Ergebnissen in die Datenbank „literatur-02.dbm“. Alle Schritte sind bereits vorbereitet und als sog. „Jobdatei“ von *Midos* abgespeichert.

Öffnen Sie in *Midos* über „Funktionen – Job starten“ den Job „lingo-to-midos“. Bestätigen Sie die „Abarbeitung der Jobdatei“ mit „OK“. Nach der erfolgreichen Abarbeitung des Jobs öffnet sich das Fenster des Programms *Midos Update* mit den Einstellungen des Jobs. Das Fenster können Sie schließen (ggf. die dort enthaltenen Einzelschritte auch näher ansehen).

Öffnen Sie über „Datenbank – Öffnen“ die Datenbank „literatur-04.dbm“.[3] Diese enthält nun alle durch Automatische Schlagwortvergabe und *Lingo*-Indexierung erzeugten Ergebnisse und ist Ausgangspunkt für den folgenden, letzten Abschnitt der Aufgabenstellung.

9 Erstellen einer Retrievalanwendung und Durchführen von Testrecherchen (Midos 6)

Midos verfügt mit dem Modul zur Erstellung einer „Windows-Retrievalanwendung“ über eine leistungsfähige Funktion, mit der sich die Ergebnisse von Indexierung und Erschließung im Kontext einer Such- und Finde-Umgebung ansehen und überprüfen lassen. Eine Einführung in die Thematik gibt Abschnitt 2.10 des Kapitels 2 und den Abschnitt 8.8 des Kapitels 8. Die Funktionalitäten und Konfigurationsmöglichkeiten der Retrievalanwendung finden Sie mit zahlreichen Screenshots auf den Folien 73-86 dieses Vorlesungsskripts.

Abschließend soll für die zuletzt erzeugte Version der Datenbank („literatur-04.dbm“) eine solche Windows-Retrievalanwendung erzeugt werden, um mit ihr einige vergleichende Suchen durchzuführen.

Öffnen Sie über „Funktionen – Windows-Retrieval-Anwendung herstellen“ den entsprechenden Dialog. Dort sind für die insgesamt 6 Teildialoge bereits Einstellungen vorhanden, mit denen erfolgreich eine Retrievalanwendung erstellt werden kann. Starten Sie diesen Prozess über den Button „Gesamte Anwendung erzeugen“. Nach einigen Schritten mit mehreren Info-Fenstern öffnet sich die Retrievalanwendung mit der Suchmaske.

Wählen Sie aus der Indexliste für das Feld „Deskriptoren“ (Button links neben dem Suchfeld „Deskriptoren“ in der Suchmaske) 5 Deskriptoren aus. Wählen Sie dabei auch Deskriptoren, die thematisch zu den von Ihnen erfassten Dokumenten passen.

[3] Es befindet sich auch noch eine Datei „literatur-03.dbm“ im Datenbankverzeichnis, die vom Job als Zwischenschritt erzeugt wird. Falls es bei der Abarbeitung des Jobs nicht zu Fehlermeldungen kommt, kann diese Version ignoriert werden.

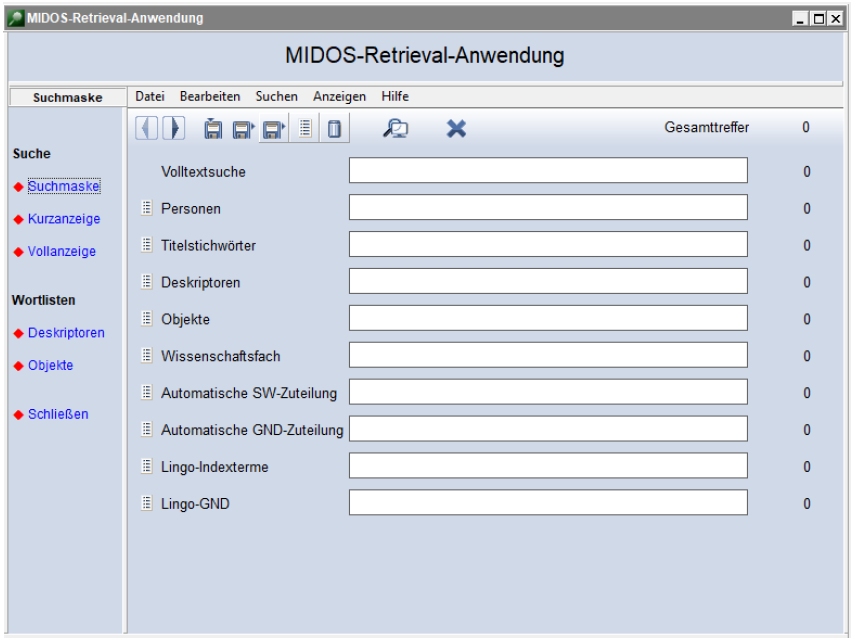

Abbildung 8: Midos – Windows-Retrievalanwendung

Führen Sie mit jedem dieser Deskriptoren jeweils eine Suche im Feld „Deskriptor“ durch und analysieren Sie die Treffermengen in Bezug auf die in ihnen enthaltenen relevanten bzw. nicht relevanten Dokumente. Halten Sie das Ergebnis der Analyse in einer Tabelle fest.

Führen Sie mit jedem dieser Deskriptoren jeweils auch Suchen in der Volltextsuche und in den vier automatisch erstellten Kategorien durch und analysieren Sie die Treffermengen in Bezug auf die in ihnen enthaltenen relevanten bzw. nicht relevanten Dokumente. Übernehmen Sie die Ergebnisse ebenfalls in die Tabelle.

Bewerten Sie die Suchergebnisse auf quantitativer Basis durch Verwendung von Standardmaßen wie *Recall* und *Precision*.

Lernportfolio – Teil 11 Übernehmen Sie die Ergebnisse der durchgeführten Suchen und deren Bewertung in Ihr Lernportfolio (keine Screenshots). Bewerten Sie den Nutzen der Suchen in den unterschiedlichen automatischen Erschließungkategorien, insb. im Vergleich zu einer Volltextsuche und im Vergleich zu einer Deskriptorsuche. Erklären Sie den möglichen Erfolg bzw. Misserfolg bei Suchen in den unterschiedlichen automatisch erstellten Suchkategorien.

10 Anhang

10.1 Modulprüfung

Die Modulprüfung besteht aus der Abgabe des vollständigen Lernportfolios als Hausarbeit (*pdf*) bis zum angegebenen Termin im *Moodle*-Kursraum.

Kriterien für die Bewertung des Lernportfolios sind:

- Problembewältigung (30%): Aufgaben vollständig ausgeführt; sachliche Richtigkeit;
- Reflexion (50%): Bezug zur Aufgabe; Antworten und Ideen entwickeln; Zusammenhänge erkennen und darstellen; kritische Beurteilung;
- Formalia (20%): Sprache, Grammatik, Rechtschreibung, Layout.

Formalia

Voraussetzung für die Teilnahme an der Modulprüfung ist die Auswahl eines Datensets aus der Datei rohdaten-001-050.zip und der Eintrag der Auswahl in der Tabelle bdk62-dis25b-auswahl-des-datensets-2023w (vgl. Abschnitt 1.2).

Die Bearbeitung eines Datensets, das nicht dem gewählten entspricht, wird als ungültige Abgabe gewertet.

Das Lernportfolio ist eine individuelle Prüfungsleistung. Identische Antworten oder Teil-Antworten werden als Täuschungsversuch bewertet.

10.2 Installationsanleitung für Ruby und Lingo unter Windows

Installation von Ruby

Auf Windows-Rechnern lässt sich Ruby über das Gesamtpaket *RubyInstaller* installieren (getestet mit Windows 10):

`https://rubyinstaller.org/downloads/`

Dort bitte die höchste 2er-Version (derzeit (05/22) *Ruby+Devkit 2.7.X (x64)*) herunterladen und starten. Alle Konfigurations- bzw. Einstellmöglichkeiten können bei den voreingestellten Standardwerten bleiben.

Nach einem Neustart kann die installierte Version auf der Kommandozeile abgefragt werden:

```
$ ruby -v
```

Installation von *Lingo*

Lingo liegt als sog. „gem“ vor; „Gems“ ist das offizielle Paketsystem für die Programmiersprache Ruby. *Lingo* wird als „gem“ mit folgendem Befehl installiert:

```
$ gem install --no-document lingo
```

Test mit Versionsabfrage:

```
$ lingo -v
```

Falls „lingo v1.10.2“ ausgegeben wird, war die *Lingo*-Installation erfolgreich.