Winfried Gödert / Klaus Lepsky

Laborpraktikum Automatisches Indexieren – Wiederholungsfragen

Die Wiederholungsfragen dienen der Vertiefung der im Laborpraktikum behandelten Materie. Sie ergänzen die Übungsaufgaben, die sich im Buch am Ende jedes Kapitels finden. Es wäre nützlich, wenn solche Fragen nicht nur von uns gestellt würden. Zusätzliche Wiederholungsfragen aus dem Kreis der Teilnehmerinnen und Teilnehmer der Laborpraktika sind immer willkommen.

I Automatische Schlagwortvergabe mit Midos (Kapitel 5.2)

1. Warum ist es nicht zweckmäßig, die Automatische Schlagwortvergabe für alle Kategorien eines Datensatzes durchzuführen?

2. Welchen Verweisungstyp enthält die Datei `auto-sw.wtx` (bzw. `synonym.wtx`) und wie erkennt man die Richtung der Verweisung?

3. In welcher Form enthält die Datei `synonym.txt` diese Verweisungen?

4. Geben Sie zwei Datensätze an, an denen man die Wirkung der Verweisungen auf die Zuteilung der Auto-Schlagwörter sehen kann?

5. Wäre es sinnvoll, mit der Datei `synonym.wtx` zu einem Unterbegriff auch Oberbegriffe zuteilen zu lassen?

6. Was müsste man tun, um zu Unterbegriffen auch Oberbegriffe zu erzeugen?

7. Nennen Sie drei Beispiele, für die das Erzeugen von Oberbegriffen sinnvoll wäre.

II Das Indexierungssystem Lingo (Kapitel 5.3)

IIa tokenizer und Konfigurationsdatei (Kapitel 5.3.2)

8. Wie wird in einer Konfigurationsdatei `*.cfg` festgelegt, dass ein Attendee mitarbeitet oder nicht?

9. Wie wird in einer Konfigurationsdatei `*.cfg` festgelegt, ob eine Ergebnisdatei geschrieben wird?

10. Welche Attendees müssen bei einem Verarbeitungslauf mindestens eingeschaltet sein, damit Lingo sinnvolle Ergebnisse produzieren kann?

11. Welche Zeichen benutzt der Attendee `tokenizer` zur Bestimmung von Wörtern?

12. Ist die Zeichenkette „3-D-Simulation“ für den `tokenizer` ein Wort?

13. Arbeitet der `tokenizer` regelbasiert oder gestützt auf Wortlisten?

14. Welche Endungen haben die Ergebnisdateien, die jeweils von den folgenden Attendees produziert werden?
    - `sequencer`
    - `decomposer`
    - `multiworder`

15. Welche Wortklassenkennung besitzen die Wörter der `*.non` Datei in der `*.log` Datei ?

16. An welcher Stelle im Abschnitt „Inhalte verarbeiten“ der `*.cfg`-Datei muss die Angabe `out: datei` erfolgen, damit die erzeugten Ergebnisse vom `vector_filter` als `in: datei` verarbeitet werden können?

IIb Grundformerzeugung und Wortklassenerkennung (Kapitel 5.3.2)

17. Welche Bedeutung hat das "?" hinter Wörtern in der `*.log`-Datei?

18. Welchen Nutzen kann es haben, die Wörter der `*.non`-Datei einer Analyse zu unterziehen?

19. Woher bezieht Lingo die Wortklassenkennungen, die in der `*.log`-Datei angegeben werden?

20. Welcher Attendee erzeugt Grundformen und wie macht er das?

21. Kann Lingo das Wort „Menschern“ verarbeiten und eine korrekte Grundform generieren?

22. Bilden Sie aus der Grundform Mensch durch Anhängen eines Suffixes ein Wort (kein Kompositum!), das vom `word_searcher` nicht erkannt wird.

23. Ist die Reihenfolge der Attendees beliebig oder nicht? Gilt diese Aussage für alle Attendees gleichermaßen?

24. Besitzt der `word_searcher` eine Kenntnis darüber, ob ein erkanntes Wort ein korrektes Wort der deutschen Sprache ist?

25. Warum gibt es keine Wortklassenkennung für Homonyme?

26. Wie muss der `vector_filter` für die Erzeugung der `*.vec`-Datei eingestellt werden, damit nur
    – Adjektive
    – Verben
    – Komposita
    ausgegeben werden?

IIc Kompositumerkennung, Longest Matching, Wörterbücher (Kapitel 5.3.3)

27. Welche Wörter werden nach dem `word_searcher` vom `decomposer` weiterverarbeitet?

28. Werden durch den `decomposer` immer nur **zwei** Zerlegungsbestandteile eines Kompositums ermittelt? In welcher Datei kann man darauf durch welche Einstellung Einfluss nehmen?

29. Welchen Einfluss hat es auf das Zerlegungsergebnis von Komposita, ob das Verfahren des *Longest Matching* von links oder von rechts durchgeführt wird? Welche Variante wird von Lingo verwendet? Erklären Sie das folgende Ergebnis:

    ```
    lex:)  <Wirkungsorte|KOM = [(wirkungsorte/k), (sorte/s+),
    (wirkung/s+)]>
    ```

30. Wie kann man erreichen, dass „Wirkungsorte“ korrekt zerlegt wird?

31. Welche Wortklassenkennungen können die durch den `decomposer` erkannten Zerlegungsbestandteile haben?

32. Warum kennt Lingo kein Wörterbuch, das nur *Komposita* enthält?

33. Erläutern Sie den Zusammenhang zwischen `usr-dic` und `user-dic.txt` in der Datei `de.lang`.

34. Was bedeutet der Eintrag `source:` in einer `*.cfg`-Datei?

35. Was muss man tun, damit der `decomposer` ein Benutzerwörterbuch verwendet?

36. Was bedeutet der Eintrag `mode: first` in einer `*.cfg`-Datei?

37. Ist es möglich, ein Wörterbuch zu gestalten, das nur *Adjektive* enthält und in die Konfiguration einer Lingo-Verarbeitung einzubinden?

IId Semantische Analyse: multiworder, sequencer, synonymer (Kapitel 5.3.4)

38. Welcher Eintrag muss im Benutzerwörterbuch für Mehrwortgruppen vorhanden sein, damit die Phrase „Menschen für Menschen“ vom Attendee `multi_worder` erkannt und ausgegeben wird?

39. Ist es empfehlenswert, für den Attendee `multi_worder` in der `*.cfg`-Datei neben dem Wörterbuch `sys-mul` auch das Wörterbuch `sys-dic` anzugeben? In welcher Reihenfolge?

40. Mit welcher Einstellung des Attendee `sequencer` lässt sich die Phrase „Menschen für Menschen“ identifizieren? Wird bei der Ausgabe das Wort „Menschen“ oder das Wort „Mensch“ ausgegeben?

41. Welche Einstellung für den Attendee `sequencer` erzeugt aus der Wortfolge „Kommunale öffentliche Bibliothek“ die Ausgaben:

    a) Bibliothek, kommunal öffentlich
    b) öffentlich Bibliothek, kommunal
    c) kommunal öffentlich Bibliothek

42. Welche der Varianten empfiehlt sich für den Aufbau eines Suchindex in einer Retrievalumgebung? Welche für den Aufbau eines Mehrwortgruppen-Wörterbuchs, das vom Attendee `multi_worder` benutzt werden soll?

43. Wie müsste ein Eintrag im Benutzerwörterbuch für den Attendee `synonymer` aussehen, damit aus „Bücherei“ im Text das Ergebnis „Bibliothek“ erzeugt wird?

44. Wie müsste ein Eintrag im Benutzerwörterbuch für den Attendee `synonymer` aussehen, damit aus „Schlagwort“ im Text die Ergebnisse „Deskriptor“ und „Vorzugsbenennung“ erzeugt werden? Geht das auch wechselweise, z. B. aus „Deskriptor“ im Text sollen „Schlagwort“ und „Vorzugsbenennung“ generiert werden?

45. Ist „deskriptoren=deskriptor“ ein sinnvoller Eintrag in einem Synonym-Wörterbuch?

46. Dürfen Synonyme auch Komposita sein?

IIe LIR-Konfiguration (Kapitel 5.3.6)

47. Warum ist es nicht zweckmäßig, die aus einer Datenbank exportierten Datensätze mit der `lemma.cfg` zu indexieren, obwohl sie alle in einer Datei stehen?

48. Wie zweckmäßig wäre es, die exportierten Datensätze einer Datenbank jeweils in getrennte Dateien zu schreiben, um sie mit der `lemma.cfg` indexieren zu können?

49. Warum werden für die Zwecke der Indexierung nicht alle Felder der Datenbank exportiert? Unterscheidet sich die Antwort von der Antwort auf Frage 1?

50. Was muss beim Export der Datensätze aus einer Datenbank beachtet werden, damit durch Lingo erzeugte Indexierungsdaten dem zutreffenden Datensatz zugeordnet werden können?

51. Welche Unterschiede gibt es zwischen der automatischen Schlagwortvergabe mit MIDOS und der Indexierung mit Lingo:

    - ☐ MIDOS-AutoSW kann keine Adjektiv-Substantiv-Verbindungen erzeugen
    - ☐ MIDOS-AutoSW kann alle Singular-/Plural-Wendungen erkennen
    - ☐ MIDOS-AutoSW kann Synonyme auf ihre Vorzugsbenennung abbilden
    - ☐ Lingo kann alle synonymen Wortformen als Indexterme erzeugen
    - ☐ Lingo kann Flektionsformen von Substantiven im Plural erkennen
    - ☐ MIDOS-AutoSW kann Flektionsformen von Kompositabestandteilen erkennen
    - ☐ Lingo kann Flektionsformen von Kompositabestandteilen erkennen

52. Ist es sinnvoll, die Ergebnisse einer automatischen Schlagwortvergabe mit MIDOS mit Indexierungsergebnissen von Lingo für einen gemeinsamen Suchindex einer Retrievalanwendung zusammenzufassen?

53. Ist es sinnvoll, die Inhalte der `*.vec` Datei und der `*.mul` Datei einer Lingo-Indexierung zu einem gemeinsamen Suchindex zusammenzufassen?

54. Kann ich mir eine Liste von Indextermen ausgeben lassen, die für mindestens drei Datensätze erzeugt wurden?