SPRACHENIDENTIFIZIERUNG

MASTERSEMINARSUCHMASCHINEN

STEFAN LANGER

CIS, UNIVERSITÄT MÜNCHEN

SOMMERSEMESTER 2021

Wozu Sprachen- und Kodierungserkennung

Interne Verarbeitung

- Kodierungserkennung um überhaupt mit einem Text arbeiten zu können
- Sprachenerkennung zur weiteren linguistischen Verarbeitung
  - Tokenisierung, Lemmatisierung (Stemming) etc etc

Filter – für Suchergebnisse

- Suche nur in einer bestimmten Sprache
- Nachträgliche Einschränkung

Ranking/Relevanz

Sprachenidentifizierung Bsp 1

ویلیام شکسپیر در ۲۶ آوریل سال ۱۵۶۴ در انگلستان در شهر استراتفورد متولد شد. شهرت شکسپیر به عنوان شاعر، نویسنده، بازیگر و نمایشنامه نویس منحصربه‌فرد است و برخی او را بزرگ ترین نمایشنامه نویس تاریخ می‌دانند، اما بسیاری از حقایق زندگی او مبهم است.

fa

Medan Brand var eit drama Ibsen sleit lenge med, kom Peer Gynt omntrent av seg sjølv. Dramaene står i et refleksjonstilhøve til kvarandre

nn

Drama radio enwog gan Dylan Thomas a gyhoeddwyd yn 1954 yw Under Milk Wood. Mae'r ddrama yn disgrifio digwyddiadau mewn un diwrnod yn unig, yn y pentref dychmygol Llareggub, er cred llawer fod nifer o'r cymeriadau yn seiliedig ar bobl go iawn ag oedd yn byw yn Nhalacharn.

cy

Sprachenidentifizierung Bsp 2

(nach Dunning: Statistical Identification of Language, 1994):

e pruebas biquimica

man immunodeficiency

faits se sont produi

er biochemischen Forsch

Zeichensatzerkennung

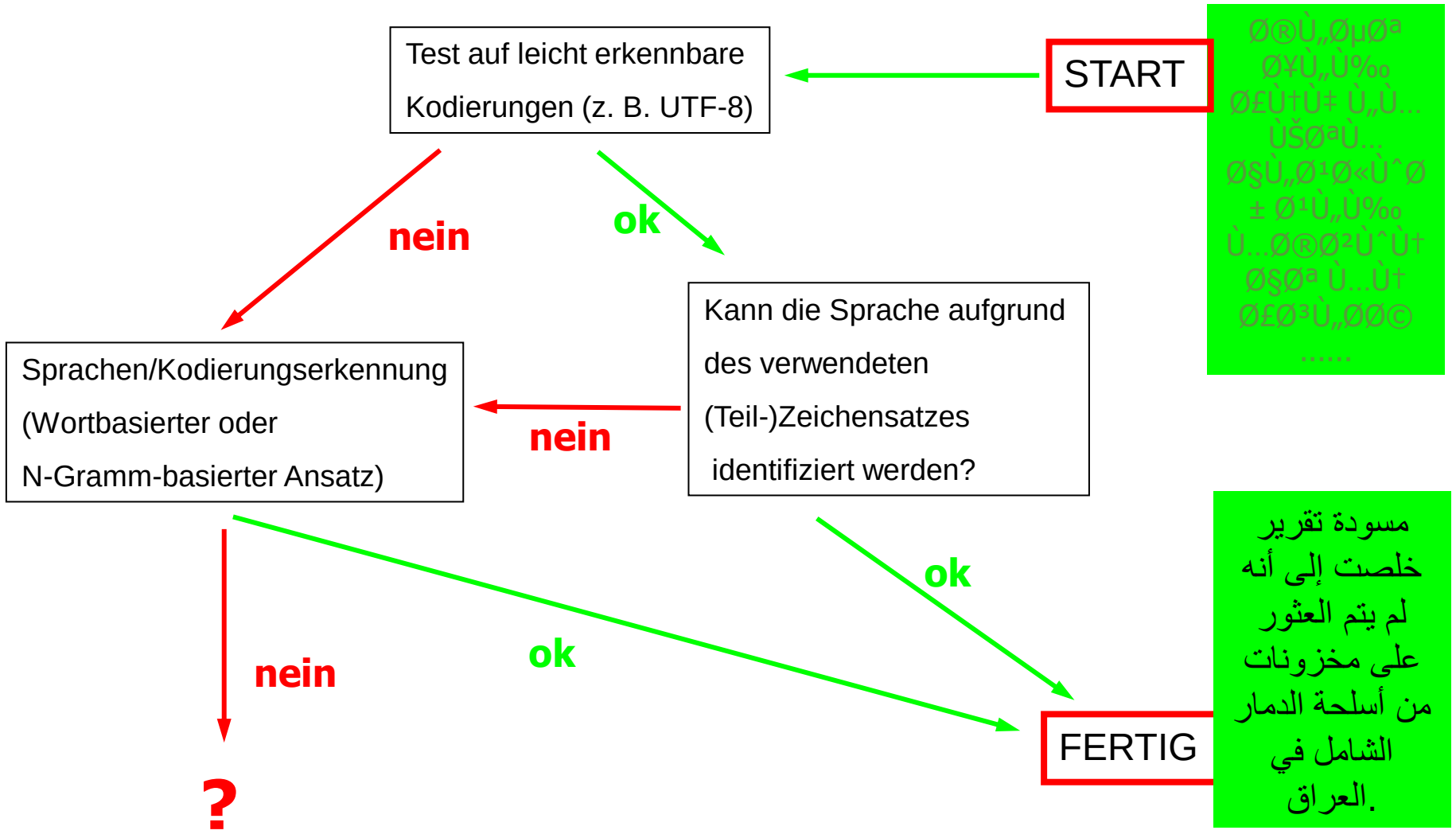
Eng verbunden mit der Sprachenidentifizierung ist die Erkennung der Zeichensatzkodierung eines Dokuments

Bei Unicode-Zeichensatzkodierungen ist die Zeichensatzerkennung relativ einfach und kann in der Regel vor der Sprachenidentifizierung erfolgen

- Byte Order Mark (BOM
- Typische Sequenzen (UTF-8)

Im Falle von 1-Byte Zeichensätzen, bei denen keine oder unzuverlässige Metainformationen vorliegen, lassen sich Zeichensatzkodierungserkennung und Sprachenidentifizierung nicht trennen

Algorithmus Sprachen- und Kodierungserkennung



Wörterbuchbasierte Erkennung

Daten

- Wörterbuch mit 100-10 000 Wörtern (abhängig vom zu klassifizierenden Dokumenttyp und dem morphologischen System einer Sprache) in einer Zeichensatzkodierung
- Konversion des Wörterbuchs in alle Zeichensatzkodierungen, die für eine Sprache relevant sind

Algorithmus

- Vergleiche Wörter im Dokument mit Wörtern im Wörterbuch
- Erkennungswert eines Wortes abhängig von:
  - Worthäufigkeit
  - Eindeutigkeit
  - Länge

Wörterbasierte Erkennung – Datenstrukturen/Algorithmen

Wörterbuch / dictionary – Implementiert als Trie oder Hash

- Information zu jedem Wort: Sprache, Gewichtung

Liste der zu erkennenden Sprachen (Kodierungen)

N-Gramm-basierter Ansatz

Daten

- Für jedes Sprach/Kodierungspaar
  - N-Gramm-Liste mit Häufigkeit

Algorithmus

- Vergleiche N-Gramm-Liste mit N-Grammen aus Dokument
- Berechne Ähnlichkeit zwischen Trainingsdaten und Dokument
  - (Wahrscheinlichkeit der Zugehörigkeit zur Sprache)

N-gramm gestützte Erkennung – Datenstrukturen/Algorithmen

Wörterbuch / dictionary der N-Gramme – Implementiert als Trie oder Hash

- Information zu jedem N-Gramm: Sprache, Gewichtung

Liste der zu erkennenden Sprachen (Kodierungen)

Vergleich der Ansätze

Wortbasiert	**N-Gramm-Ansatz**
Trainingskorpus muss nicht ganz sauber sein, da manuelle Überprüfung möglich	Sauberes Trainingskorpus
Aufwändiges Training, wenn manuell überprüft	Training einfach
Nachträgliche Überprüfung und Korrektur unproblematisch	Nachträgliche Überprüfung / Revision kaum möglich, außer über Trainingskorpus
relative große Datenbasis zur Erkennung	kleine Datenbasis
Neue Kodierungen einfach zu ergänzen	Konversion des Trainingskorpus nötig zur Ergänzung von neuen Kodierungen
Schlecht geeignet für sehr kurze Dokumente oder Listen seltener Wörter	Auch für sehr kurze Dokumente geeignet
Nicht für Sprachen ohne durch Leerzeichen markierte Wortgrenzen (Japanisch, Chinesisch...	Alle Sprachen

Sprachenidentifizierung - Daten

- Datenquellen
  - Wikipedia
  - Wictionary
  - Alle Datenquellen mit monolingualen Dokumenten

Datenaufbereitung:

Parsen der Dokumente und Erstellung von Wort-/N-Grammlisten

Größe der Daten: Textabdeckung als Kriterium!

Algorithmen: mögliche Details

Wie oft wird ein Wort/N-Gramm berücksichtigt?

Wortlänge

Aussortieren unpassender Einträge

- Andere Sprachen

- Anderer Zeichensatz

Algorithmen

Alle klassischen Algorithmen zur Klassifikation lassen sich für die Sprachenerkennung anwenden (Bayes, K-nearest neighbour u.a.)

Eine naiver Algorithmus (größte Vektorenähnlichkeit zwischen Repräsentation der Klasse und des Dokuments, verbunden mit einem Schwellenwert ist meist ausreichend)

Sprachenidentifizierung: problematische Fälle

Multilinguale Dokumente

Sonstige irreguläre Dokumente (Namenslisten u.ä.)

Sehr kurze Texte (z.B. Suchanfragen)

Unterscheidung sehr ähnlicher Sprachen. Beispiele:

- Indonesisch/Malay
- Serbisch (lat), Kroatisch, Bosnisch

Unterscheidung von Sprachvarianten

- Englisch AE/BE

Sprachenidentifizierung in Suchmaschinen

Erkennung der Dokumentsprache

- Relativ einfach für einsprachige, längere Dokumente
- Schwieriger für mehrsprachige Dokumente und sehr kurze Dokumente

Erkennung der Anfragesprache

- Schwierig bis unmöglich aufgrund der Kürze durchschnittlicher Anfragen
- Einige Sprachen lassen sich aufgrund des Zeichensatzes erkennen, z.B. Japanisch, Thai …