

FEATUREAUSWAHL AM BEISPIEL SPRACHENIDENTIFIZIERUNG

SEMINAR: KLASSIFIKATION

DOZENT: STEFAN LANGER

CIS, UNIVERSITÄT MÜNCHEN

Wozu Sprachen- und Kodierungserkennung

Interne Verarbeitung

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

Filter

- Wähle Dokumente nur aus einer bestimmten Sprache aus
- Finde Teildokumente bei mehrsprachigen Dokumenten

Sprachenidentifizierung

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

fa

Medan Brand var eit drama Ibsen sleit lenge med, kom Peer Gynt omtrent 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

Sprachenerkennung – Beispiel

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

e pruebas biquimica

man immunodeficiency

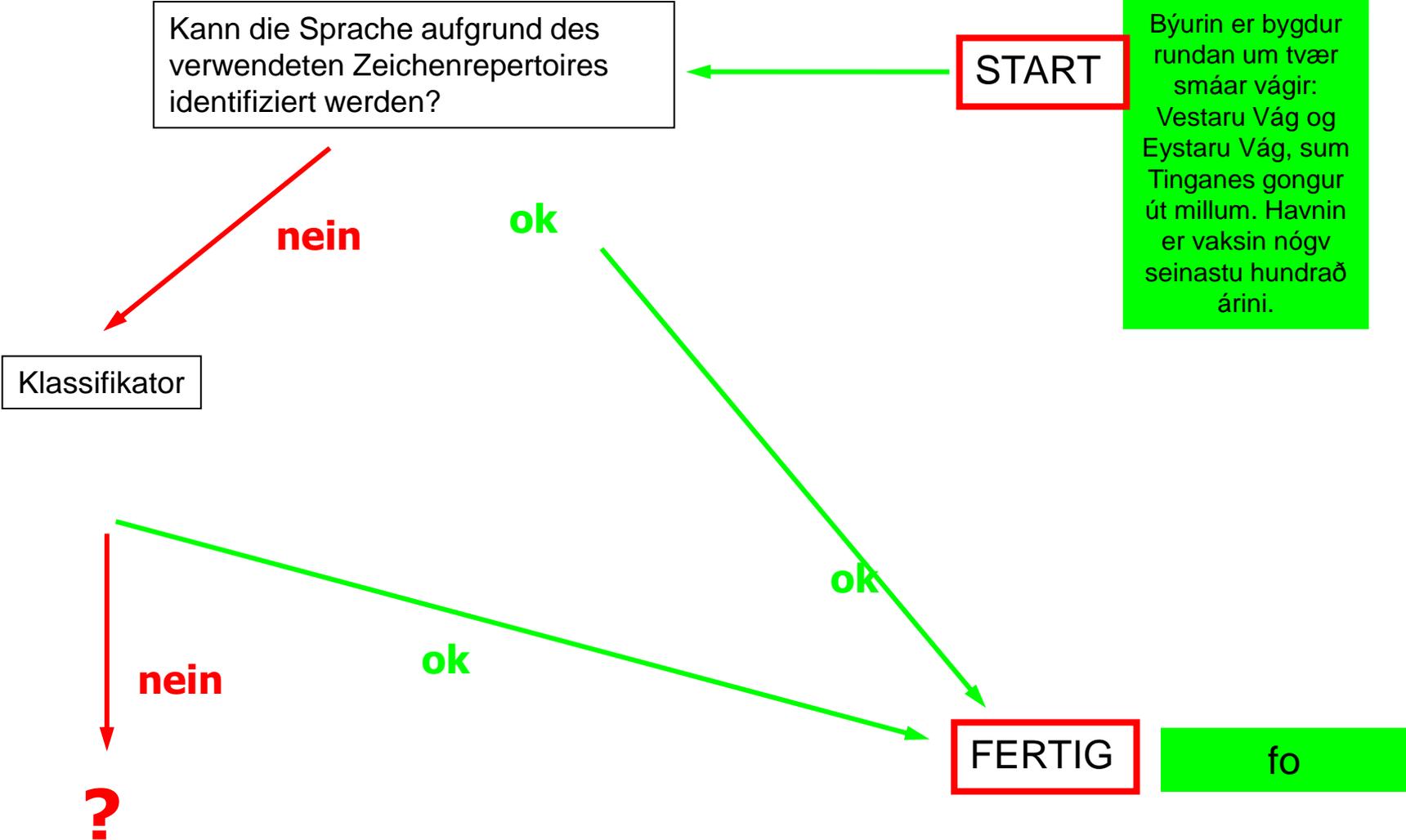
faits se sont produi

er biochemischen Forsch

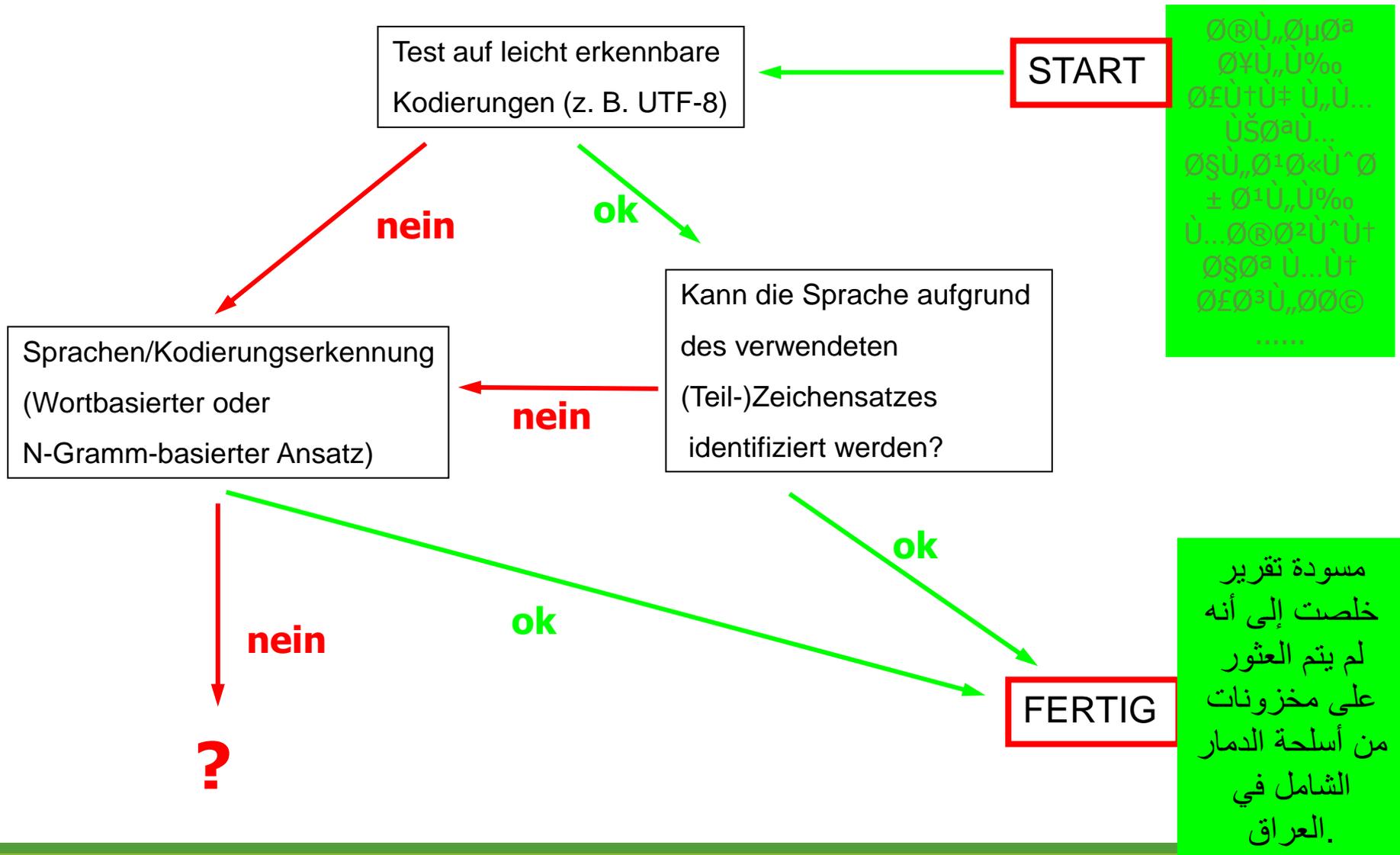
Sprachenerkennung – Features

- Semantische Analyse nicht notwendig
- Worte oder kleinere Einheiten ausreichend

Algorithmus Sprachenerkennung



Algorithmus Sprachen- und Kodierungserkennung



Featureauswahl: Wortbasierte 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

- Vergleiche Wörter im Dokument mit Wörtern im Wörterbuch / Modell
- Erkennungswert eines Wortes abhängig von:
 - Worthäufigkeit (z.B. naive Bayes)
 - Eindeutigkeit
 - Länge

N-Gramm-basierter Ansatz

Daten

- Für jedes Sprach/Kodierungspaar
 - Buchstaben N-Gramm-Liste mit Häufigkeit
 - $N = 2-4$

Vergleich der Featureauswahl

Wortbasiert	N-Gramm-Ansatz
Trainingskorpus muss nicht ganz sauber sein, da manuelle Überprüfung möglich	Sauberer 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

Schwierige Fälle

Sehr kurze Texte

Mehrsprachige Texte

Text mit zahlreichen Eigennamen

Kein Fließtext sondern Listen (z.B. Produktlisten, Namenslisten)