Klassifikation von Textabschnitten

Am Beispiel von Stellenanzeigen

(**JASC** - **J**ob **A**ds **S**ection **C**lassifier)

Jürgen Hermes, Alena Geduldig
Sprachliche Informationsverarbeitung
Institut für Linguistik

Universität zu Köln

Gliederung

1. Einführung: Zu welchem Zweck machen wir das?
2. Klassifikation – ein kurzer Überblick
3. Umsetzung: Der Job Ads Section Classifier (JASC)
4. Systemdemonstration
5. Probleme und Ausblick

Informationsextraktion aus Jobangeboten

- Auftraggeber: Bundesinstitut für Berufsbildung (BIBB).
- Zuständig u.a. für den Zuschnitt von Lehr- und Ausbildungsberufen
- Datenbank der Bundesanstalt für Arbeit (BfA) über 1.000.000 Stellenzeigen, wachsend
- Strenge Datenschutzauflagen verlangen „Vercodung“
- Empirischer Ansatz: Unterstützung von Expertengruppe durch realwirtschaftliche Daten

Informationsextraktion: Templates

Für unser Kaufhaus in ###### haben wir zum ##.##.#### eine Ausbildungsstelle zum/zur Kaufmann/-frau im Einzelhandel zu besetzen.

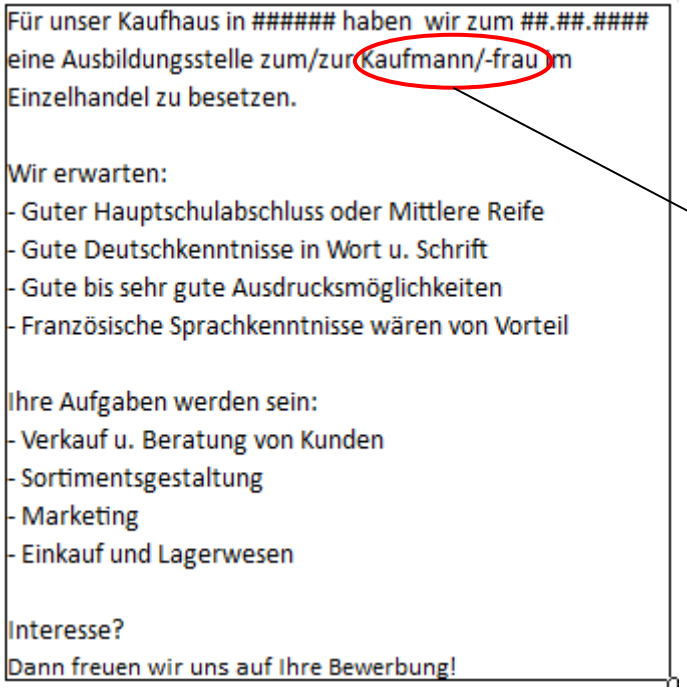
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

Ihre Aufgaben werden sein:
- Verkauf u. Beratung von Kunden
- Sortimentsgestaltung
- Marketing
- Einkauf und Lagerwesen

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

Wert	Attribut
	Datum
	Branche
	Beruf
	Gefordert
	Gewünscht
	Aufgaben

Informationsextraktion: Templates



Wert	Attribut
	Datum
	Branche
	Beruf
	Gefordert
	Gewünscht
	Aufgaben

Informationsextraktion: Templates

Für unser Kaufhaus in ###### haben wir zum ##.##.#### eine Ausbildungsstelle zum/zur Kaufmann/-frau im Einzelhandel zu besetzen.

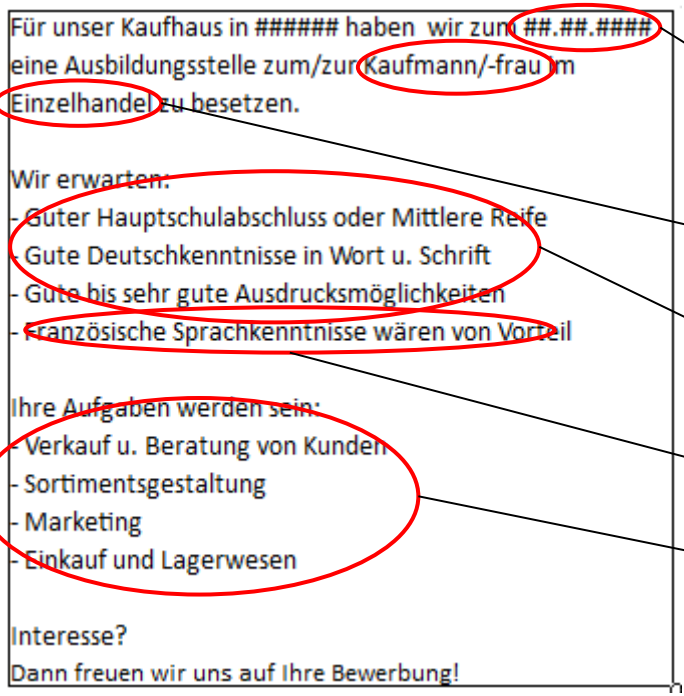
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

Ihre Aufgaben werden sein:
- Verkauf u. Beratung von Kunden
- Sortimentsgestaltung
- Marketing
- Einkauf und Lagerwesen

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

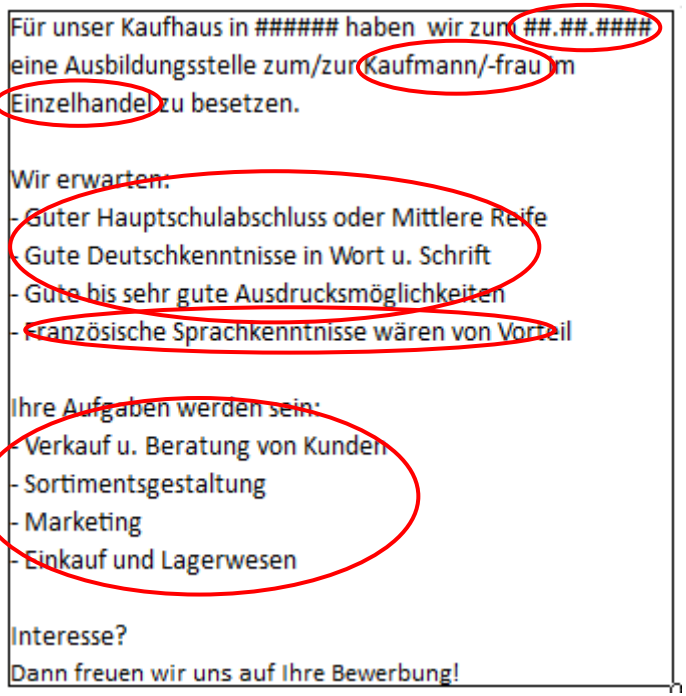
Wert	Attribut
	Datum
	Branche
Kaufmann/-frau	Beruf
	Gefordert
	Gewünscht
	Aufgaben

Informationsextraktion: Templates



Wert	Attribut
	Datum
	Branche
Kaufmann/-frau	Beruf
	Gefordert
	Gewünscht
	Aufgaben

Informationsextraktion: Templates



Für unser Kaufhaus in ###### haben wir zum ##.##.#### eine Ausbildungsstelle zum/zur Kaufmann/-frau im Einzelhandel zu besetzen.

Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

Ihre Aufgaben werden sein:
- Verkauf u. Beratung von Kunden
- Sortimentsgestaltung
- Marketing
- Einkauf und Lagerwesen

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

Wert	Attribut
##.##.####	Datum
Einzelhandel	Branche
Kaufmann/-frau	Beruf
Hauptschule / …	Gefordert
Fränzösisch	Gewünscht
Verkauf / …	Aufgaben

Vorstufe: Klassifikation von Abschnitten

Für unser Kaufhaus in ###### haben wir zum ##.##.#### eine Ausbildungsstelle zum/zur Kaufmann/-frau im Einzelhandel zu besetzen.

Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

Ihre Aufgaben werden sein:
- Verkauf u. Beratung von Kunden
- Sortimentsgestaltung
- Marketing
- Einkauf und Lagerwesen

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

Klassen
1: Firma
2: Job
3: Anforderungen
4: Sonstiges

Vorstufe: Klassifikation von Abschnitten

Für unser Kaufhaus in ###### haben wir zum ##.##.#### eine Ausbildungsstelle zum/zur Kaufmann/-frau im Einzelhandel zu besetzen.

Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

Ihre Aufgaben werden sein:
- Verkauf u. Beratung von Kunden
- Sortimentsgestaltung
- Marketing
- Einkauf und Lagerwesen

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

Klassen
1: Firma
2: Job
3: Anforderungen
4: Sonstiges

Vorstufe: Klassifikation von Abschnitten

Voraussetzungen und Ziele

- **Was ist vorhanden?**
  - **Datenbank mit >1 Mio datengeschützten, nicht ausgezeichneten Stellenanzeigen**
- **Was wird gebraucht?**
  - **Eine Anbindung eines Klassifikationswerkzeugs an die Datenbank für Stellenanzeigen, dafür benötigt man …**
  - **Einen Klassifizierer, der getestet und für brauchbar befunden wurde, dieser benötigt …**
  - **Modelle für die Klassen, in die Abschnitte eingeordnet werden sollen, was wiederum voraussetzt …**
  - **Ein Textkorpus mit > 1000 anonymisierten, vorklassifizierten Abschnitten.**

Automatische Klassifikation - Ansätze

Einordnung von Objekten in vorgegebene Klassen

2 Ansätze:

1. Maschinelles Lernen

- Lernen aus Erfahrung/Beispielen
- Voraussetzung: Trainingskorpus

2. Regelbasierte Klassifikation

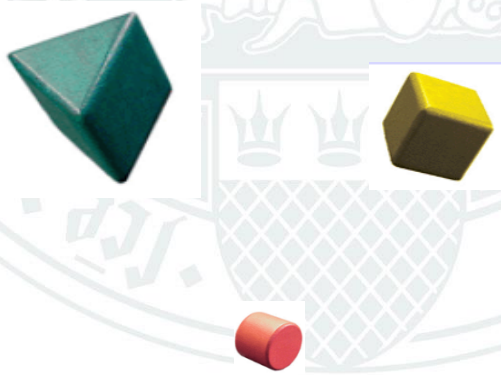
- Formulierung fester Regeln

Klassifikation - Grundprinzip

Voraussetzungen für den Vergleich von Objekten

1. Definition von Merkmalen
2. numerische Repräsentation der Merkmale (Vektor)

Beispiel: Bauklötze



M1 = Form → **Anzahl der Ecken**

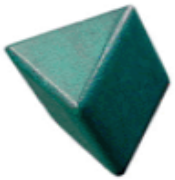
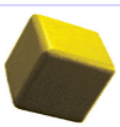
M2 = Größe → **Gewicht in Gramm**

:

Klassifikation - Grundprinzip

Beispiel: Bauklötze

M1 = Form → Anzahl der Ecken
M2 = Größe → Gewicht in Gramm

 $\approx \begin{pmatrix} 6 \\ 150 \end{pmatrix}$  $\approx \begin{pmatrix} 8 \\ 100 \end{pmatrix}$  $\approx \begin{pmatrix} 0 \\ 50 \end{pmatrix}$

Ähnlichkeit = Vektorähnlichkeit
(z.B. euklidische Distanz oder Cosinus-Ähnlichkeit)

JASC: Workflow

1. Preprocessing & Trainingskorpus
2. Feature Engineering
3. Feature Quantifying
4. Classifying
5. Evaluation
6. Ranking der Verfahren

1. Preprocessing

- **Import der Stellenanzeigen aus der BIBB-Datenbank bzw. CSV-Datei**

  → `DatabaseConnector bzw. BIBBReader`

- **Zerlegung in Paragraphen** (*ClassifyUnits*)

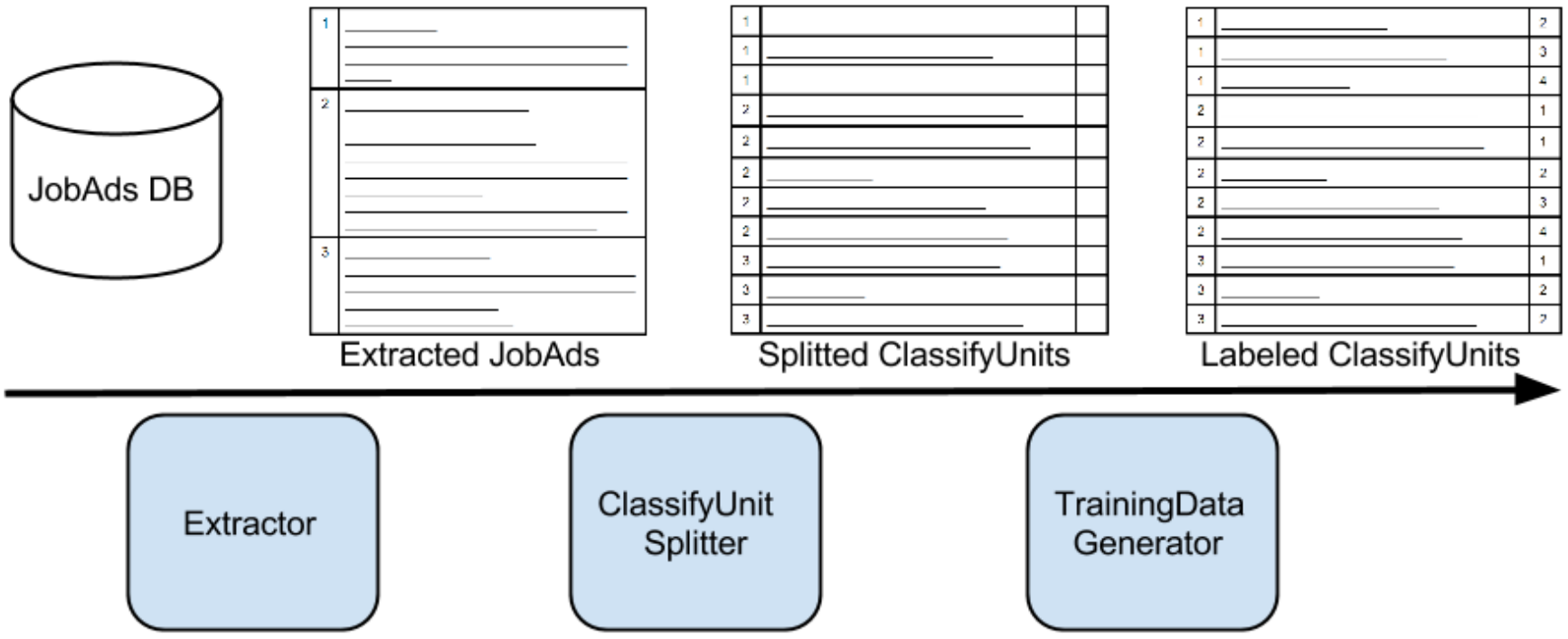
  → `ClassifyUnitSplitter`

- **Trainingskorpus erstellen**

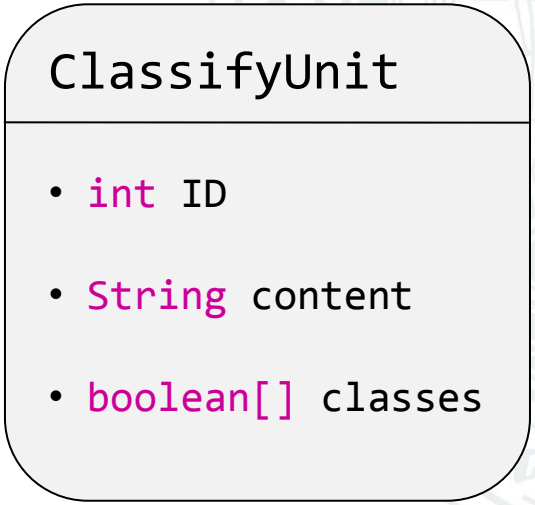
  (manuelle Klassifikation von ca. 280 Jobangeboten, mit insgesamt etwa 1500 Abschnitten)

  → `TrainingdataGenerator`

Preprocessing



ClassifyUnit nach Preprocessing



2. Merkmalsextraktion

Schritt 1: Tokenisierung

→Merkmale = Strings

Schritt 2: Filtern

→Stoppwörter filtern?

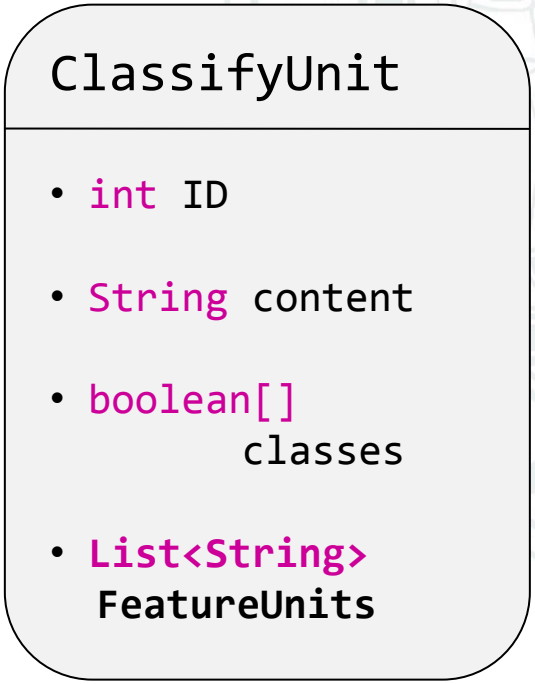
→‚unwichtige' Wörter filtern?

Schritt 3: Modifizieren

→Stemming? Normalisieren?

→NGramme?

ClassifyUnit nach Merkmalsextraktion



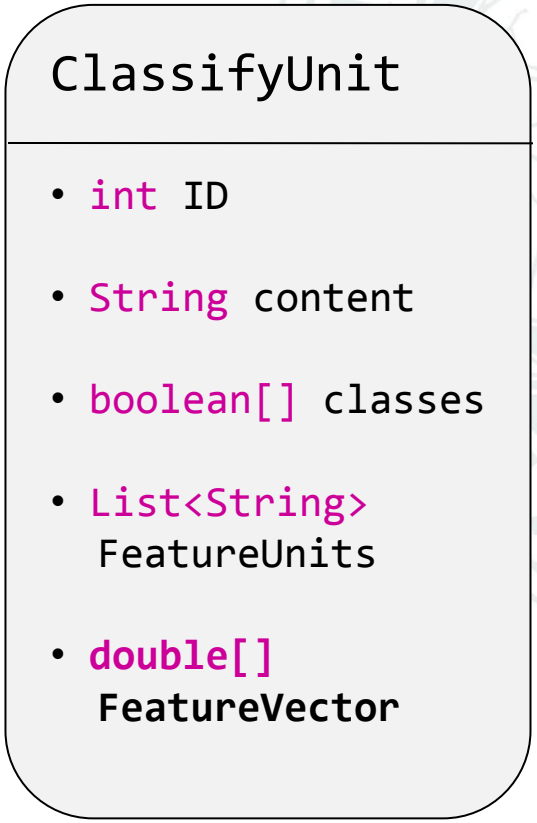
3. Merkmalsgewichtung

Ziel: numerische Repräsentation der `FeatureUnits`

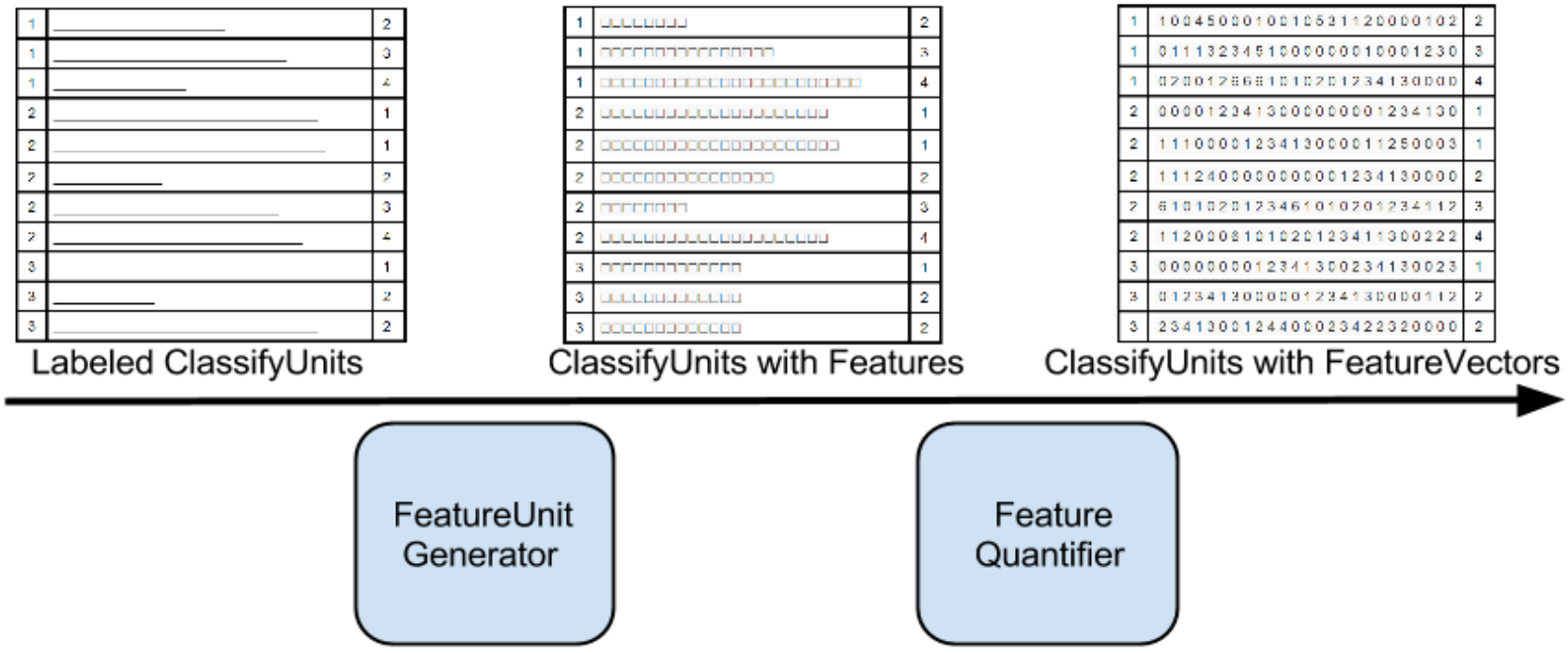
4 Möglichkeiten:

1. absolute Häufigkeit
2. relative Häufigkeit (rel. zur Textlänge)
3. Tf-Idf-Wert
4. LogLikelihood-Wert

ClassifyUnit nach Merkmalsgewichtung



Feature Engineering



Jürgen Hermes, Alena Geduldig
Sprachliche Informationsverarbeitung
Institut für Linguistik

Universität zu Köln

4. Classifying

Implementation von vier verschiedener Klassifikatoren

1. K-nearest-neighbor-Klassifikator
2. Rocchio-Klassifikator
3. NaiveBayes-Klassifikator
4. (RegEx-Klassifikator)

Rocchio-Klassifikation

- **Objekte = Feature-Vektoren**
- **Trainingsphase:**

  Bilde für jede Klasse einen Zentroid-Vektor

  (= Mittelwert aller Klassenobjekte)
- **Klassifikation eines neuen Objekts X:**

  Ordne X der Klasse mit dem nächsten Zentroid-Vektor zu
- **Konfiguration**

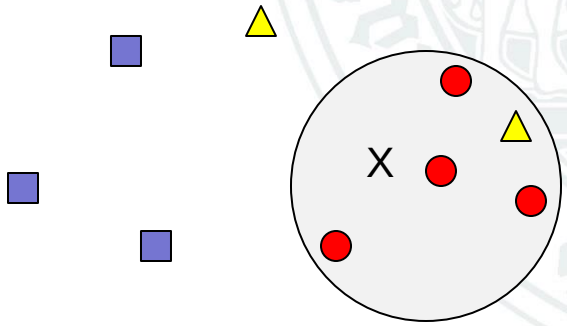
  - Distanzmaß

K-nearest-Neighbor-Klassifikation

- **Objekte = Feature-Vektoren**
- **Klassifikation eines neuen Objekts X:**
    1. Ermittle die zu X k nächsten/ähnlichsten Objekte im Trainingskorpus
    2. Ordne X der Klasse zu, die unter den k nächsten Nachbarn am häufigsten vertreten ist
- **Konfiguration:**
    - Distanzmaß
    - Anzahl der Nachbarn (k)

K-nearest-neighbor-Klassifikation

K = 5



X =

Naive Bayes Klassifikation

- Arbeitet mit *FeatureUnits* (Objekt = String-List)
- Beruht auf dem Bayes‘schen Theorem

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

- Berechnet Klassenwahrscheinlichkeiten auf der Grundlage von Wort-Häufigkeiten

5 Evaluation

Zahlreiche Konfigurationsmöglichkeiten:

- Tokens vs. NGramme
- Stoppwortfilter vs. MI-Filter vs. kein Filter
- Stemming?
- Normalisieren?
- Rocchio vs. NaiveBayes vs. KNN-Classifyer
- Distanzmaß?

→ > 5000 Konfigurationsmöglichkeiten

Welche liefert das beste Ergebnis?

5 Evaluation

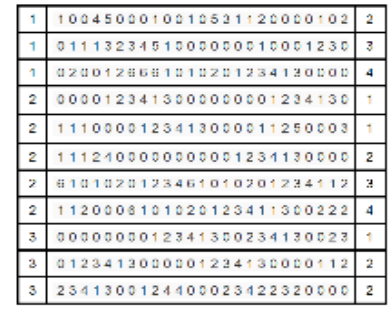
Kreuzvalidierung

- Aufteilung des Trainingskorpus in Trainings- und Testset
- Vergleich der manuell zugeordneten Klassen mit den maschinell zugeordneten

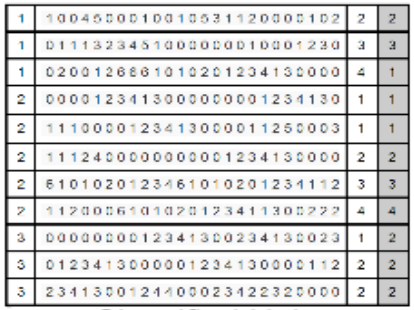
Evaluationsmaße

- Precision
- Recall
- Accuracy
- F-Measure

Classify and Evaluate



Units with FeatureVectors



Classified Units

Accuracy	0.94
Precision	0.90
Recall	0.93
F-Score	0.91

Measures

Classifier

Evaluator

Jürgen Hermes, Alena Geduldig
Sprachliche Informationsverarbeitung
Institut für Linguistik

Universität zu Köln

6. Ranking

- **mehr als 5000 Klassifikationsergebnisse**
- **Jeweils 4 Evaluationswerte**

→ **Implementation einer Ranking-Komponente**

Erstellt für jeden Evaluationswert eine Ranking-Tabelle

6. Ranking: F-Measure

fscore	prec	rec	accu	Classifier	Dist	Quant	NGrams	MI	SW	Norm
0,9882	0,9913	0,9852	0,9935	KNN(k=1)	COS	LogLike	-2-3	0	false	true
0,9882	0,9913	0,9852	0,9935	KNN(k=1)	COS	LogLike	-2-4	0	false	true
0,9882	0,9925	0,9840	0,9935	KNN(k=1)	COS	LogLike	-2-3-4	0	false	true
0,9615	0,9845	0,9395	0,9790	KNN(k=4)	COS	LogLike	-2-4	0	false	true
0,9190	0,9102	0,9278	0,9530	NaiveBayes	null	null	null	60	true	false
0,9190	0,9102	0,9278	0,9530	NaiveBayes	null	null	null	60	true	false

7. Ergebnisse

- **Das KNN-Verfahren klassifiziert am besten, braucht aber am längsten**
  - F-Score = 0,99 bei k = 1 (0,96 bei k = 4)
  - Cosinus-Distanz
  - NGramme
  - LogLikeliHood
- **Naive Bayes wesentlich performanter und gering schlechter**
  - F-Score = 0,92

Systemvorführung

Applications:

- `DefaultExperimentsGenerator`: Ausführung von 2500 verschiedenen Experimenten mit unterschiedlicher Konfiguration
- `SingleExperimentExecutor`: Ausführung eines zu konfigurierenden Experiments
- RankResulsApplication: Ranken der durchgeführten Experimente anhand ihrer Evaluationswerte.
- `ConfigurableDatabaseClassifier`: Klassifizieren von Datensätzen aus einer Datenbank anhand ausgewählter Klassifikatoren.

Code (Repo wird veröffentlicht nach Klärung zum Datenschutz):

https://github.com/spinfo/bibb-jasc

Ausblick

- Probleme bei der Übertragung auf die BIBB-Datenbank:
  - Anonymisierte Modelle vs. nicht anonymisierte Klassifikationsaten
  - Gemischtes Encoding in der Datenbank
  - Zeitbedarf des KNN-Klassifikators

- Möglichkeiten der weiteren Kooperation:
  - Behandlung Encoding, Modelle über nicht-anonymisierte Trainingsdaten
  - Erprobung schnellerer Klassifikatoren
  - Einbeziehung von Metainformationen aus der Datenbank
  - Studie Informationsextraktion
  - Auswertung gesuchter Kompetenzen und Kompetenz-Kombinationen