Mining a Corpus of Job Ads

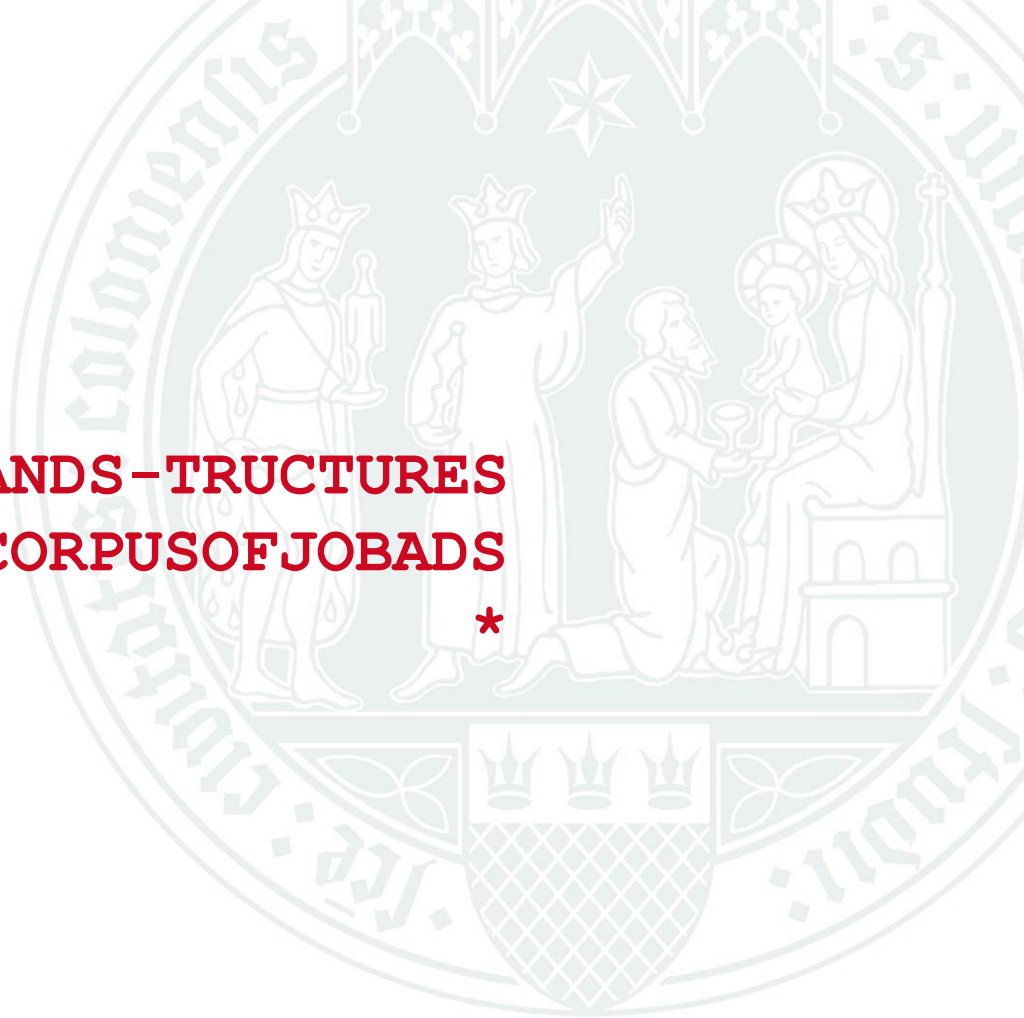
Workshop

Strings and Structures –

Computational Biology & Linguistics

Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne



```
Workshop      STRINGSANDS-TRUCTURES
Presentation  MININGACORPUSOFJOBADS
                       ***        *

Matches: 4/21
```

Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

Mining a Corpus of Job Ads

Using Strings to Structure Documents

Workshop

Strings and Structures –

Computational Biology & Linguistics

Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

Workshop -----STRINGSANDSTRUCTURE--------S
Presentation USINGSTRINGSTO-STRUCTUREDOCUMENTS
******* ********* *

Matches: 17/33

Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

Structure of the presentation

1. Project
2. Corpus
3. Goals
4. Methods
5. Framework
6. Results
7. Discussion

1. The Project

- Cooperation with the *Bundesinstitut für Berufsbildung* (BIBB- en. *Federal Institute for Vocational Education and Training*)

- First project stage (finished 12/2014): Evaluation and implementation of a text classifying tool.

- Second project stage (starting 07/2015): Evaluation and implementation of information extraction algorithms.

- Third project stage (scheduled for 2016): Statistical evaluation of the extracted information.

2. The Corpus

- Nearly 2 million job advertisements, more than 400.000 added each year
- Full texts with additional metadata (date of publication, region, branch of industry ASO)
- Source: Database from the *Bundesanstalt für Arbeit* (BfA, en. *Federal Labour Office*), where more than 60% of all job ads from Germany are recorded.
- Raw data can not be published due to privacy and copyright reasons, so we had to evaluate our methods based on anonymized samples.

3. The Goals

- **Relevant data from the Job Ads full texts should be captured and coded before 2025, because the BIBB has to delete the texts 15 years after being received from the BfA.**

- **Relevant information of the Job Ads full texts should be accessible in a performant and user-friendly way.**

- **Due to the quantity of the collected data, Information Extraction can‘t be carried out manually – thus Machine Learning methods have to be developed.**

3. Information Extraction: 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!

Value	Attribute
	Date
	Job Area
	Job Title
	Requirements
	Optional
	Tasks

3. Information Extraction: 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!

Value	Attribute
	Date
	Job Area
	Job Title
	Requirements
	Optional
	Tasks

3. Information Extraction: Templates

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

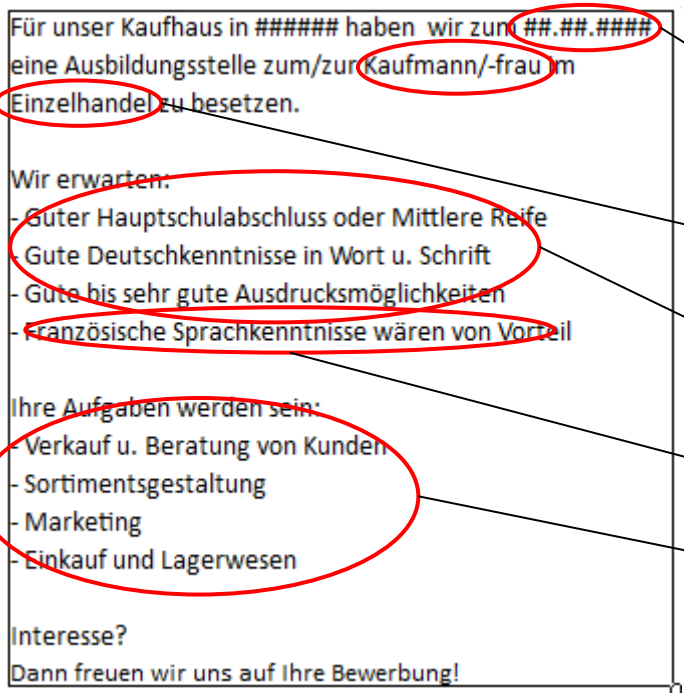
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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

Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

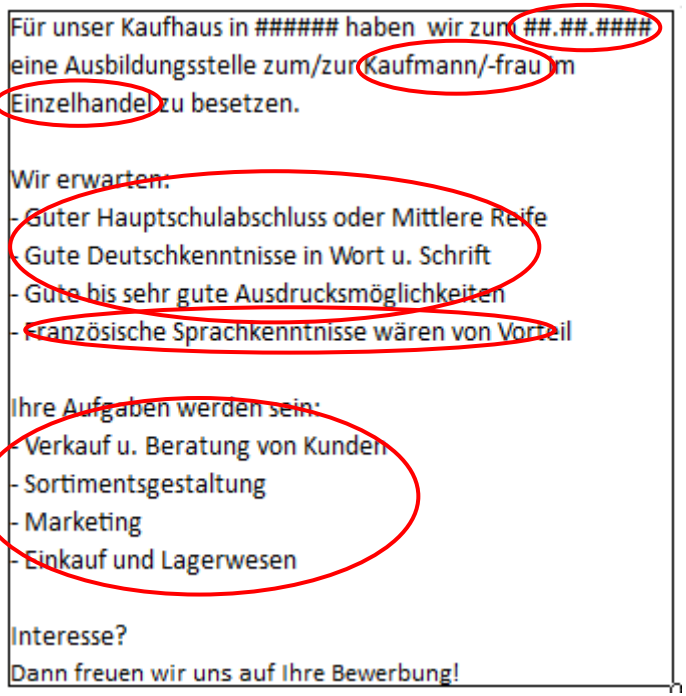
Value	Attribute
	Date
	Job Area
	Job Title
	Requirements
	Optional
	Tasks

3. Information Extraction: Templates



Value	Attribute
	Date
	Job Area
Kaufmann/-frau	Job Title
	Requirements
	Optional
	Tasks

3. Information Extraction: 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!

Value	Attribute
##.##.####	Date
Einzelhandel	Job Area
Kaufmann/-frau	Job Title
Hauptschule / …	Requirements
Französisch	Optional
Verkauf / …	Tasks

3. Preliminary Stage: Zone Analysis

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!

3. Preliminary Stage: Zone Analysis

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!

3. Preliminary Stage: Zone Analysis

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

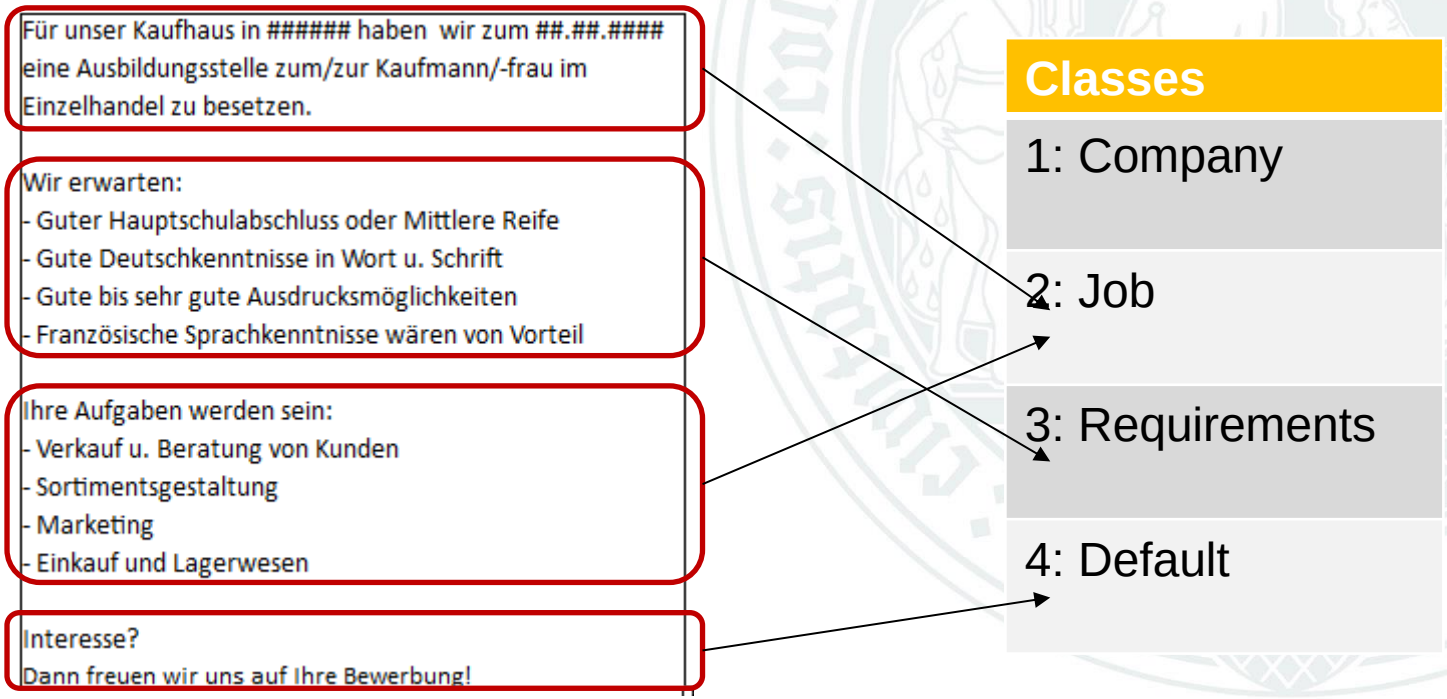
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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

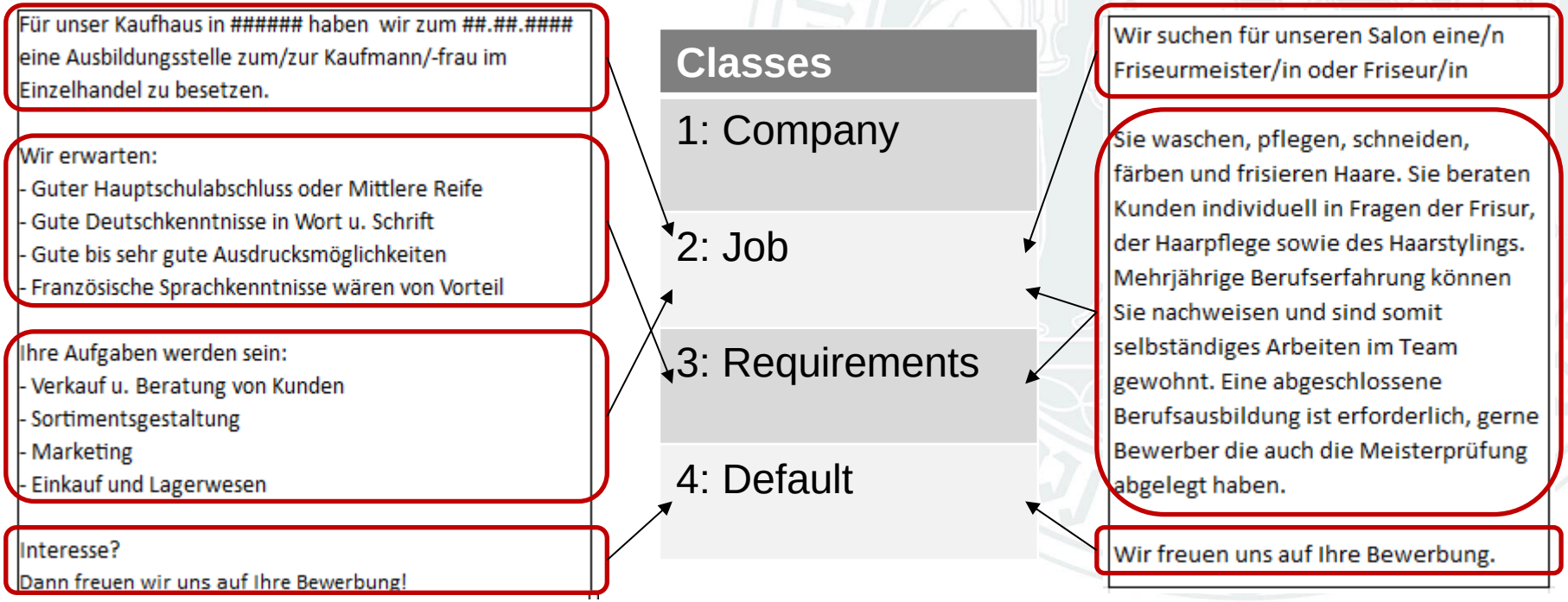
Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

Classes
1: Company
2: Job Description
3: Requirements
4: Default

3. Preliminary Stage: Zone Analysis



3. Preliminary Stage: Zone Analysis



3. Zone Analysis Requirements

- **Available:**
  - **Database with nearly 2 million job ads**
- **Required:**
  - **Connection of a zone analysis tool to the database, which requires …**
  - **An evaluated multiple-class-classificator, which requires …**
  - **Models for classes for training the classificator, which requires …**
  - **A corpus of more than 1000 anonymized, manually preclassified paragraphs from job ads.**

4. Automatic Classification – Approaches

Two different (but still combinable) approaches:

1. **Rule based classificators**
   - Based on domain specific rules
   - Require manual encoding of rules
   - Empirical formula: High precision, low recall

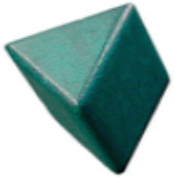
2. **Machine learning approaches**
   - Based on learning from training data
   - Require manual preparation of the training data
   - Empirical formula: High recall, low precision

4. Machine Learning – Basics

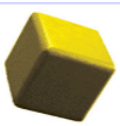
Definition and numeric representation of features for every object to classify.

Example: Bricks

$$\begin{pmatrix} number\_of\_corners \\ weight\_in\_grams \end{pmatrix}$$



$$\begin{pmatrix} 6 \\ 150 \end{pmatrix}$$



$$\begin{pmatrix} 8 \\ 100 \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 50 \end{pmatrix}$$

Calculating similarities between objects using vector similarity measures (eg. euclidean or cosine distance)

Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

5. The Framework: JASC

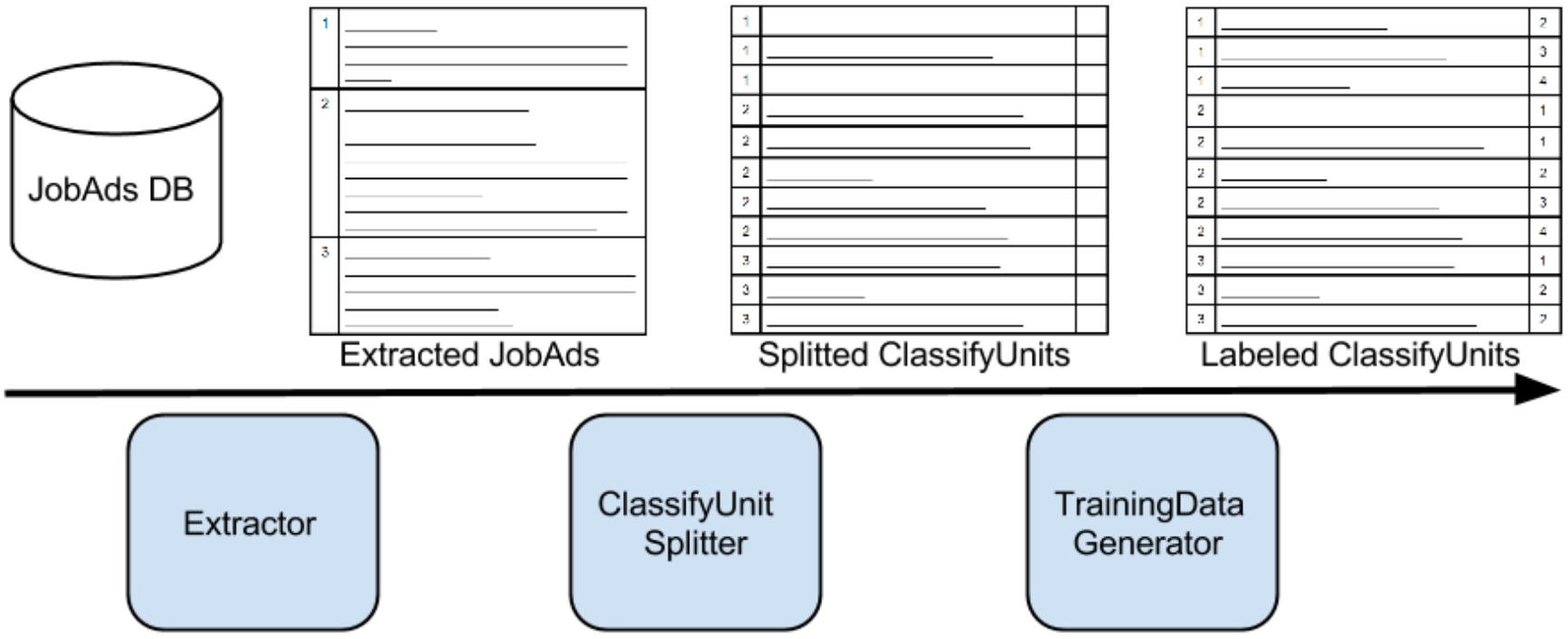
(JASC stands for Job Advertisement Section Classifier)

a. **Preprocessing & training corpus**
b. **Feature selection**
c. **Feature quantifying**
d. **Classification**
e. **Evaluation**
f. **Ranking of experiments**

5a. Preprocessing

- **Import of job ads from the database and from an export file of anonymized data**
- **Splitting each job ad into paragraphs** (*ClassifyUnits*)
- **Manual classification of training data**
  - **approximate 300 job ads → about 1500 ClassifyUnits**

Preprocessing



Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

5b. Feature Selection

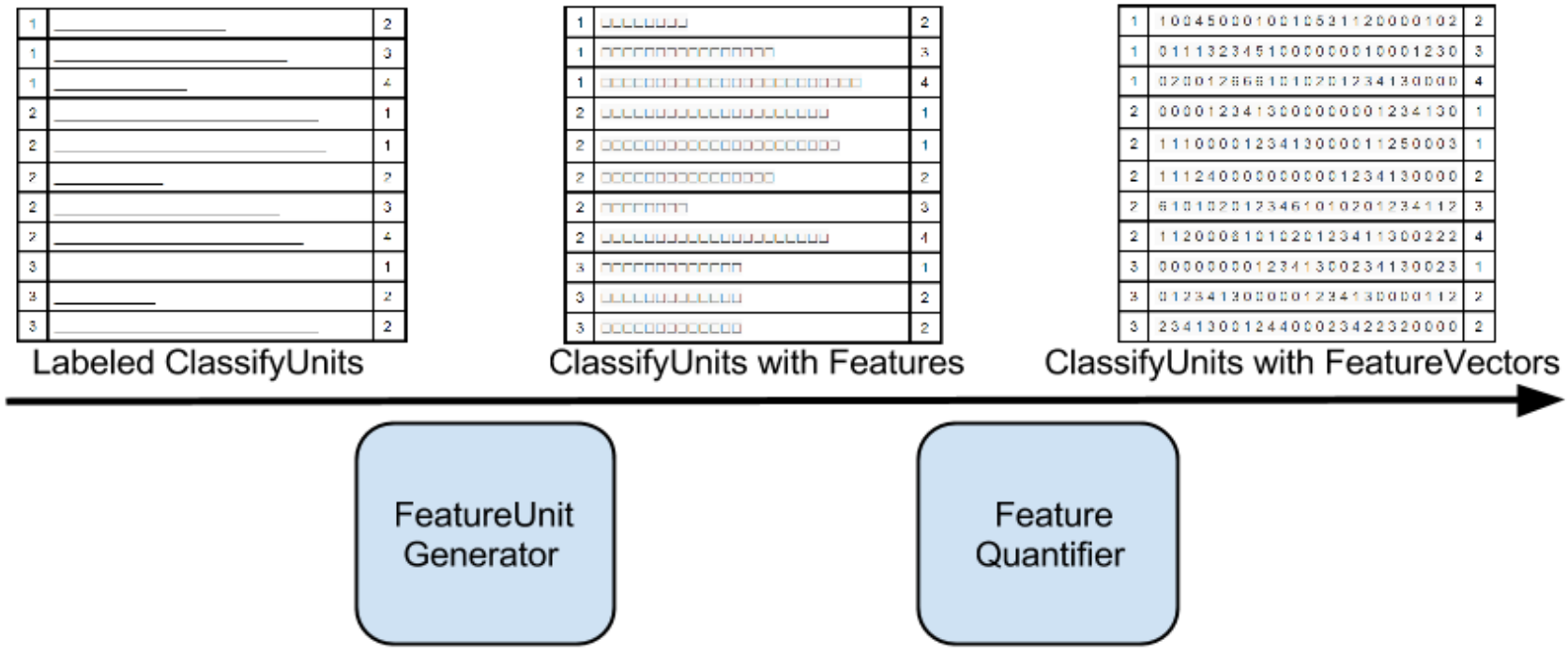
- Using the content of the paragraphs as features
  - Tokenization – Using words or ngrams of letters?
  - Should Stopwords or „irrelevant“ words be filtered out?
  - Should words be stemmed/lemmatized?
  - Should the input otherwise be normalized, eg. numbers?
  - Should word combinations be used as features?

5c. Feature Quantifying

- **Quantify features and transform them into a vector format**

- **Different options to calculate feature weigths:**
  - Absolute count
  - Relative count
  - Term Frequency / Inverse Paragraph Frequency
  - Log Likelyhood

Feature Engineering



Jürgen Hermes
Linguistic Data Processing
Department of Linguistics

University of Cologne

5d. Classification

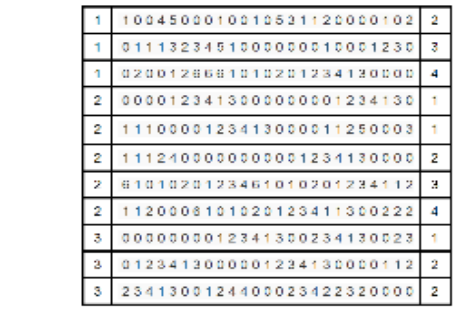
Using a combination of a RegEx-Classifier (perfect precision, recall 0.5) and 4 different Machine Learning Classifiers:

1. K Nearest Neighbor (KNN)
2. Rocchio (R)
3. Naive Bayes (NB)
4. Support Vector Machine (SVM)

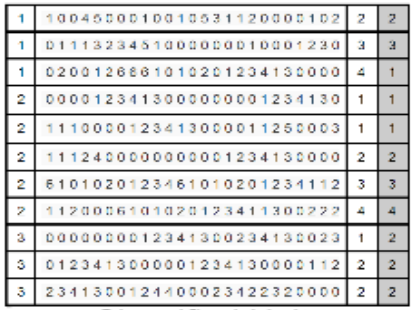
5e. Evaluation

- **Cross validation on more than 5000 experiments with different configurations**
- **As previously noted, paragraphs may be linked to more than one class. Consequently, each link should be considerd in the evaluation.**
- **The default class (class 4) should be ignored.**
- **Measures:**
  - Precision (Number of TP / Number of TP+FP)
  - Recall (Number of TP / Number of TP+FN)
  - Accuracy (Number of TP+TN / Number of all TP+TN+FP+FN)
  - F-Measure (Harmonic mean of precision and recall)

Classify and Evaluate



Units with FeatureVectors



Classified Units

Accuracy	0.94
Precision	0.90
Recall	0.93
F-Score	0.91

Measures

6. Results: Ranking of F-Measures

F-score	Preci-sion	Recall	Accu-racy	Classi-fier	Dist-ance	Quanti-fier	N-grams
0.98	0.99	0.98	0.99	KNN 1	COS	Log-Like	3
0.98	0.98	0.98	0.98	KNN 4	COS	Log-Like	2 & 3
0.93	0.93	0.93	0.97	SVM	-	Log-Like	3
0.92	0.92	0.92	0.96	Rocchio	COS	Log-Like	3
0.92	0.92	0.92	0.96	Rocchio	EUK	Log-Like	-
0.91	0.91	0.91	0.95	Naive Bayes	-	-	-
0.85	0.83	0.89	0.92	Naive Bayes	-	-	3

6. Results: Insights

- **KNN** performs best, but it is also the slowest algorithm**.**
  - F-Score = 0,98 with k = 1 (0,98 with k = 4)
  - Distance: Cosine better than Euklid
  - Quantifying: LogLikelyhood better than tf/idf
  - Selection: n-grams better than words, other variations have minor effects
- **Linear classifiers** also deliver respectable results (but poorer ones than KNN), SVM performs slightly better than Rocchio.

7. Discussion

- Problems while adapting the algorithm to the BIBB-Database
  - Anonymized models vs. not anonymized data to classify
  - Encoding-Mix inside the database
  - Time requirements of the KNN-Classifier

- Future tasks:
  - Generating models for not anonymized training data
  - Evaluate faster classificators (SVM) with the original data
  - Include metadata
  - Information Extraction


THANK
YOU