Machine Learning Day 2017

Applying *Machine Learning* to *Text Mining*

Department of Linguistics - Linguistic Data Processing | Dr. Jürgen Hermes | 06.04.2017

Preliminaries

- Where do I come from?
  - Department of Linguistics – Linguistic Data Processing
- What do we do?
  - Computational Linguistics / Natural Language Processing
- What will I present here?
  - A Machine Learning component for the Job Advertisement Project

Computational Linguistics / Natural Language Processing

- Rough definition: Everything concerning the interface between human language and the computer.
- Development of Human-Machine-Interactions
- Proof of linguistic hypotheses
- Enhance Information Retrieval and Text Mining

The Job Advertisement Challenge

- The Project
- Goals
- Methods
- Framework
- Results
- Discussion

The Job Advertisement Project

- Cooperation with the Bundesinstitut für Berufsbildung (BIBB, en. Federal Institute for Vocational Education and Training)
- First project stage (2015): Implementation and evaluation of a text classifying tool
- Second project stage (ending Mai 2017): Implementation and evaluation of information extraction algorithms.
- Third project stage (starting June 2017): Statistical evaluatuion of the extracted information.

The Data

- Nearly 2 million job ads, more than 400.000 added each year
- Full texts with additional metadata (date of publication, region, branch of industry…)
- 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- Raw data can not be published due to privacy and copyright reasons.

Goals of the project

- Relevant data from the Job Ads full texts should be captured and „coded“ before 2015, because the BIBB has to delete the original data 15 years after being recieved from the BfA.
- Relevant information of the Job Ads full text 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.

Methods: 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

Methods: Information Extraction / Templates

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

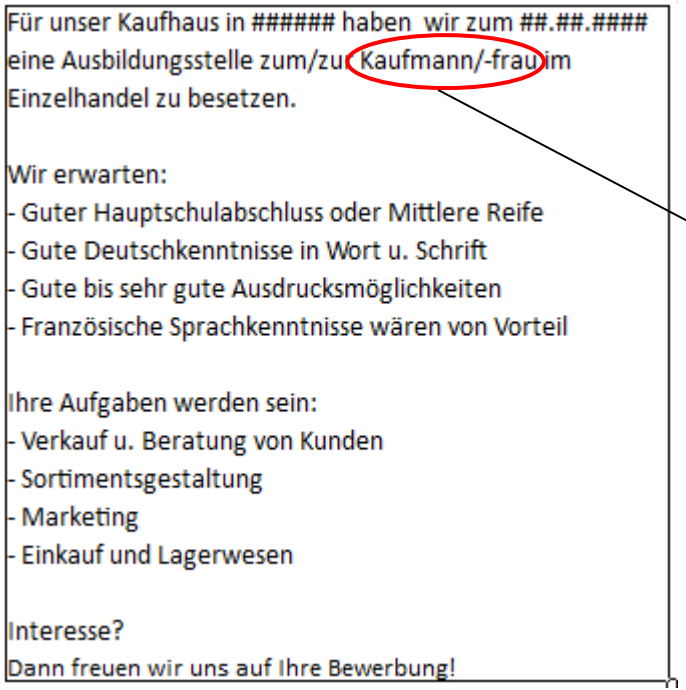
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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

Methods: Information Extraction / Templates



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

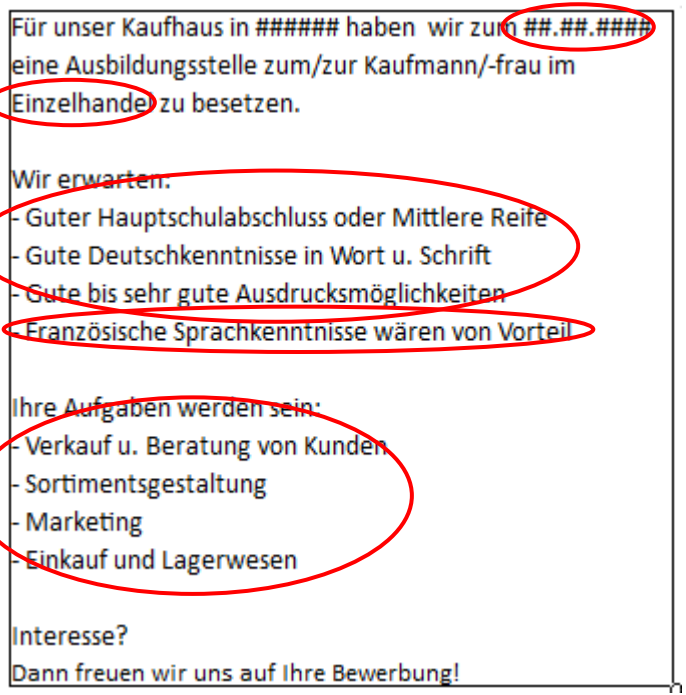
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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

Methods: Information Extraction / Templates



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

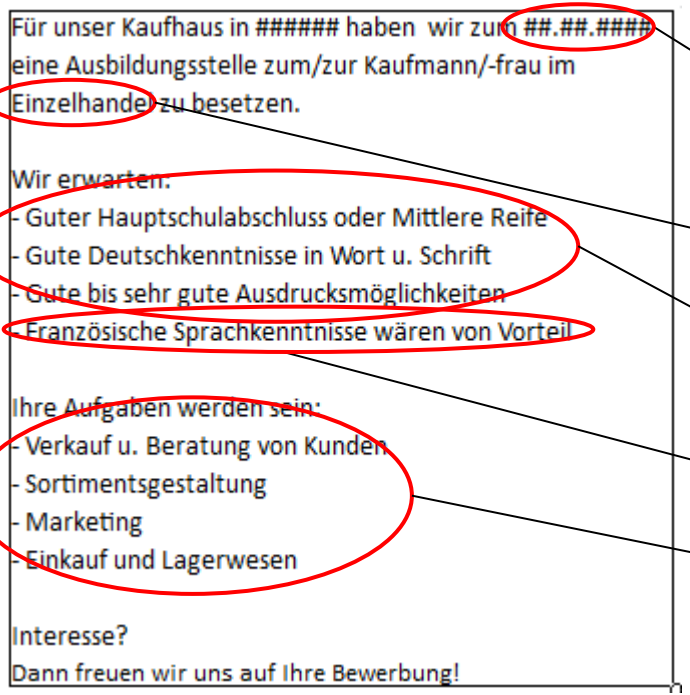
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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
Kaufmann/-frau	Job Title
	Requirements
	Optional
	Tasks

Methods: Information Extraction / Templates



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

Methods: 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
Mittlere Reife / …	Requirements
Französisch / …	Optional
Verkauf / …	Tasks

Methods: Text Classification / 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!

Methods: Text Classification / 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!

Methods: Text Classification / Zone Analysis

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

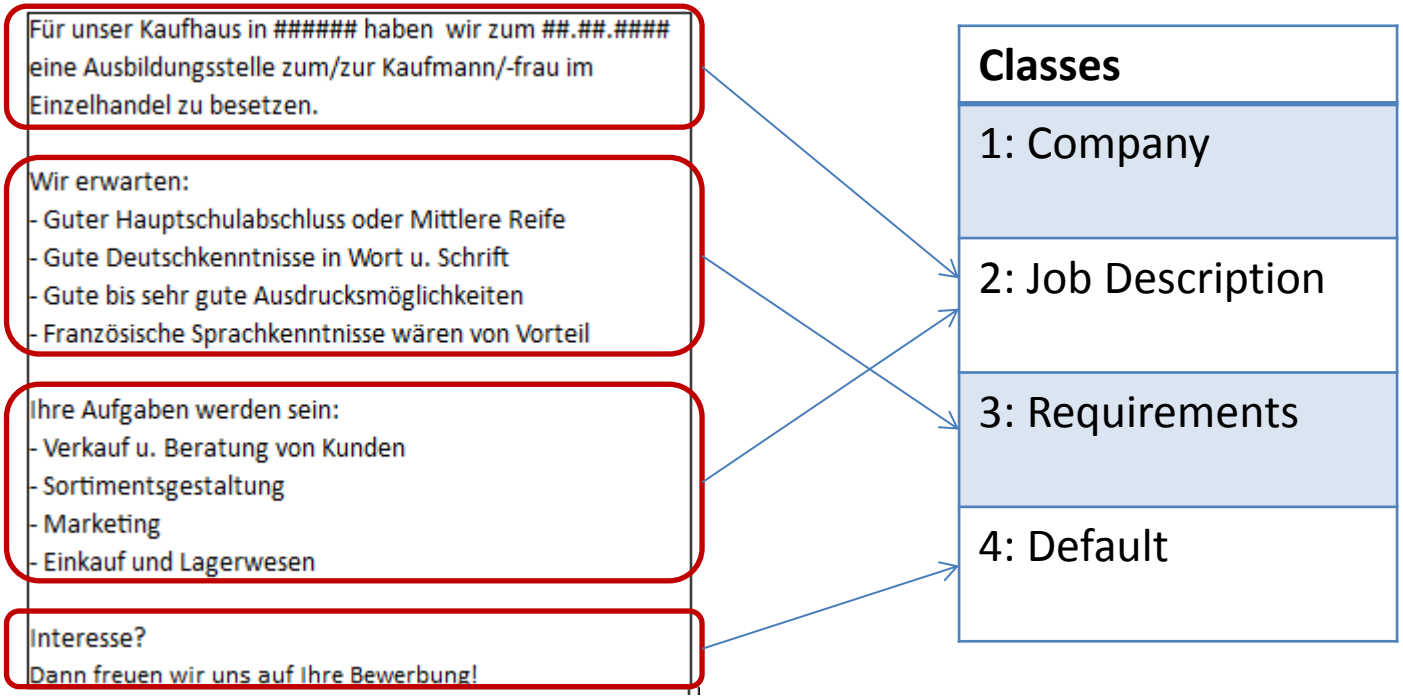
Wir erwarten:
- Guter Hauptschulabschluss oder Mittlere Reife
- Gute Deutschkenntnisse in Wort u. Schrift
- Gute bis sehr gute Ausdrucksmöglichkeiten
- Französische Sprachkenntnisse wären von Vorteil

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

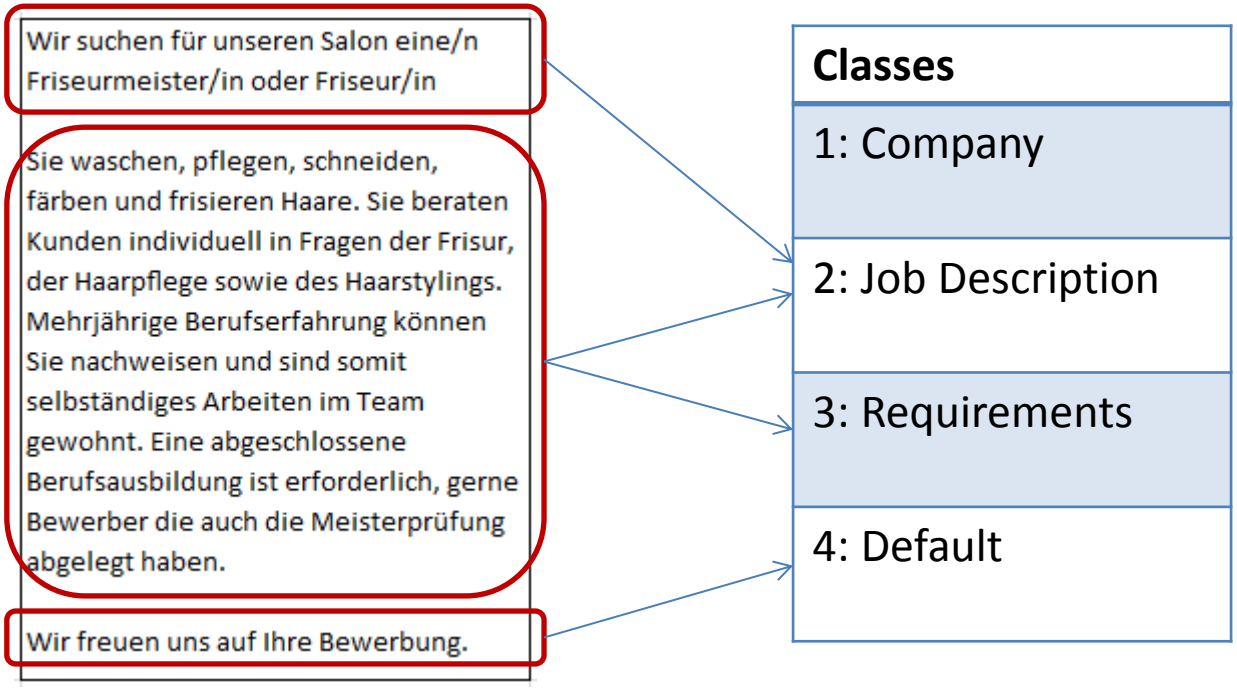
Interesse?
Dann freuen wir uns auf Ihre Bewerbung!

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

Methods: Text Classification / Zone Analysis



Methods: Text Classification / Zone Analysis



Methods: Zone Analysis Requirements

- Available: Database with nearly 2 million job ads
- Required:
  - Connection of a zone analysis tool to the database
  - Evaluated multiple-class-classifyier
  - Models for classes to train/test the classifier
  - Anonymized, manually preclassified paragraphs

Methods: Classification Approaches

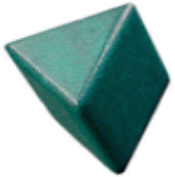
Two different (but still combinable) approaches:

1. Rule based classifiers
   - Based on domain specific rules
   - Require manual rule encoding
   - Empirical formula: High precision – low recall
2. Machine Learning classifiers
   - Based on learning from training data
   - Require manual preparation of the training data
   - Empirical formula: High recall – low precision

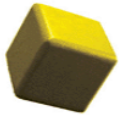
Methods: Machine Learning Classifiers – The fundamental principle

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

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



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

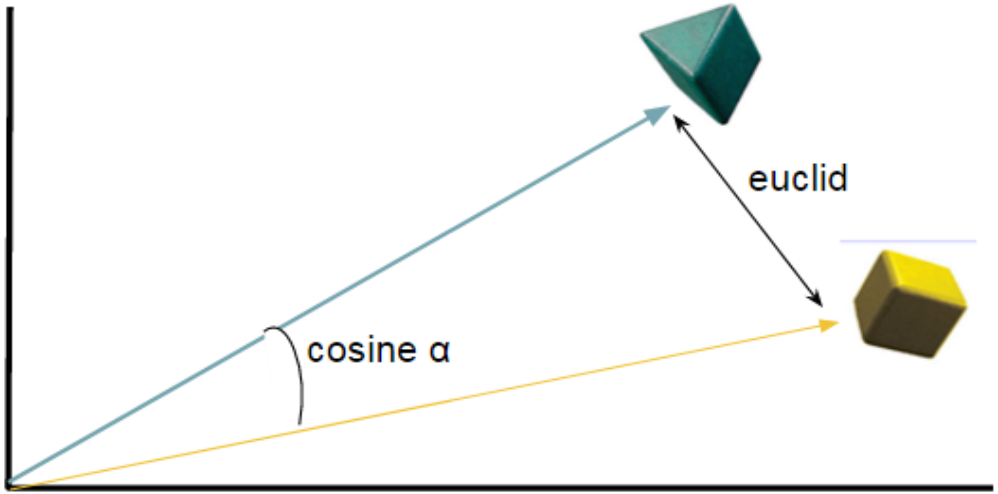


$\begin{pmatrix} 8 \\ 100 \end{pmatrix}$

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

Methods: Machine Learning Classifiers – The fundamental principle

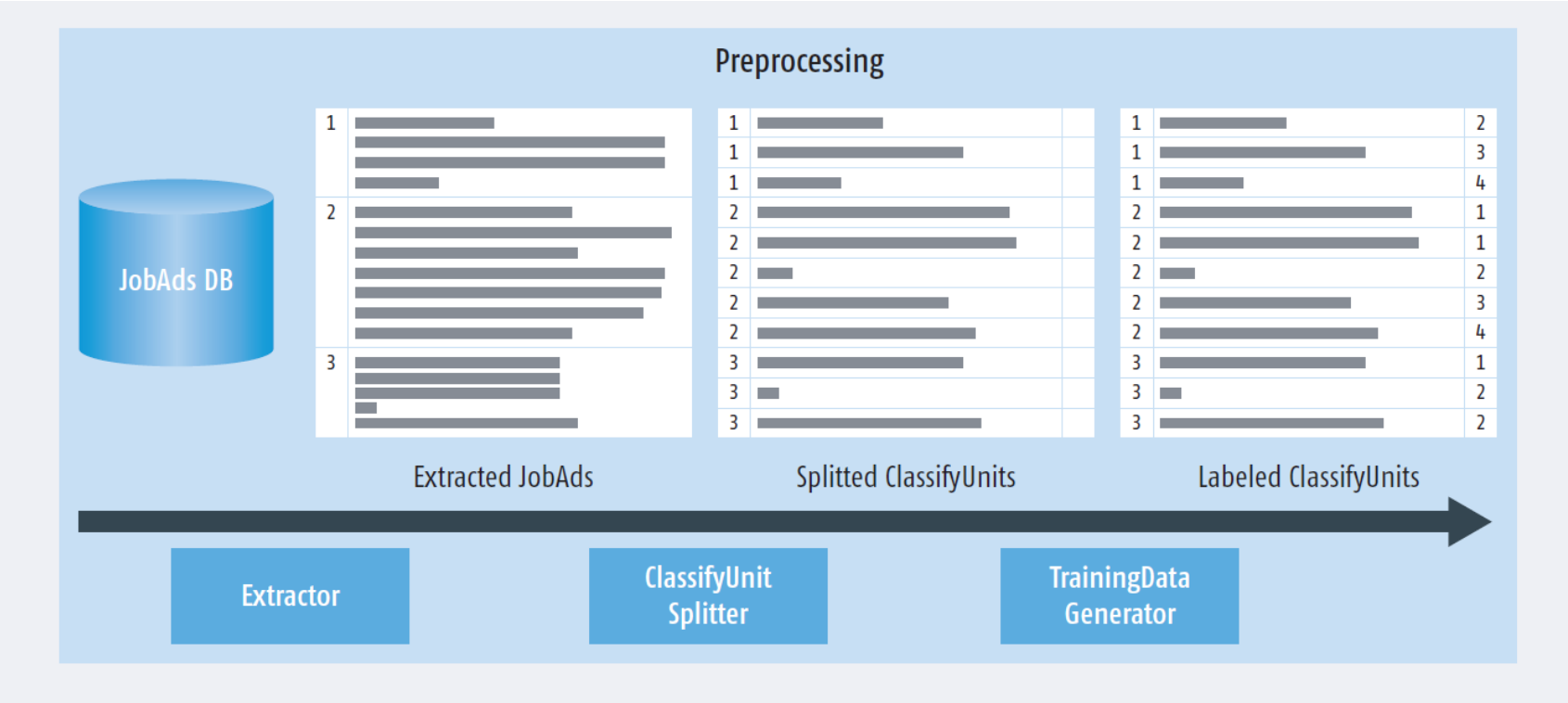
2. Calculation of similarity measures between objects using vector similarity measures (eg. Euclidean or cosine distance)



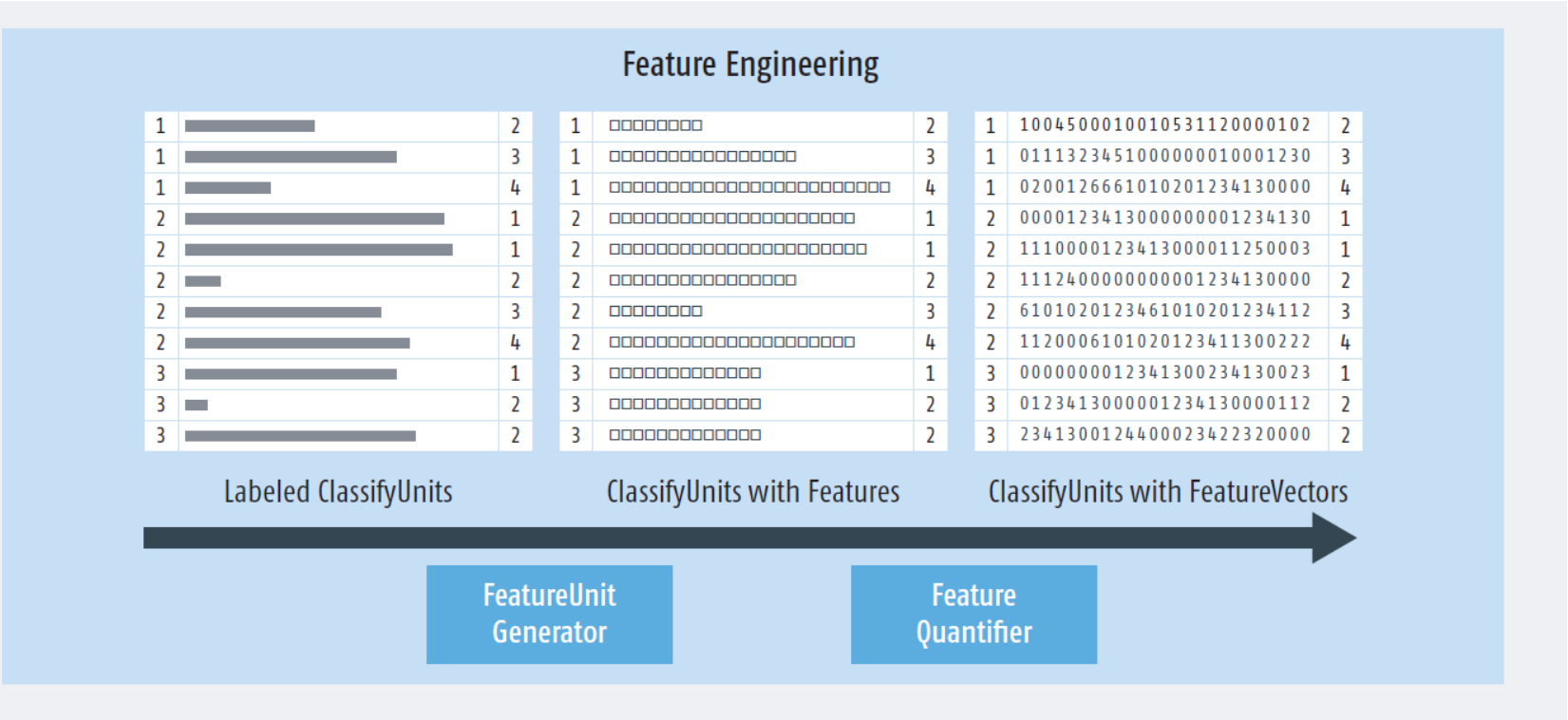
The Framework JASC (Job Ad Section Classifier)

1. Preprocessing and training corpus
2. Feature selection
3. Feature quantification
4. Classification
5. Evaluation
6. Ranking of methods / experiments

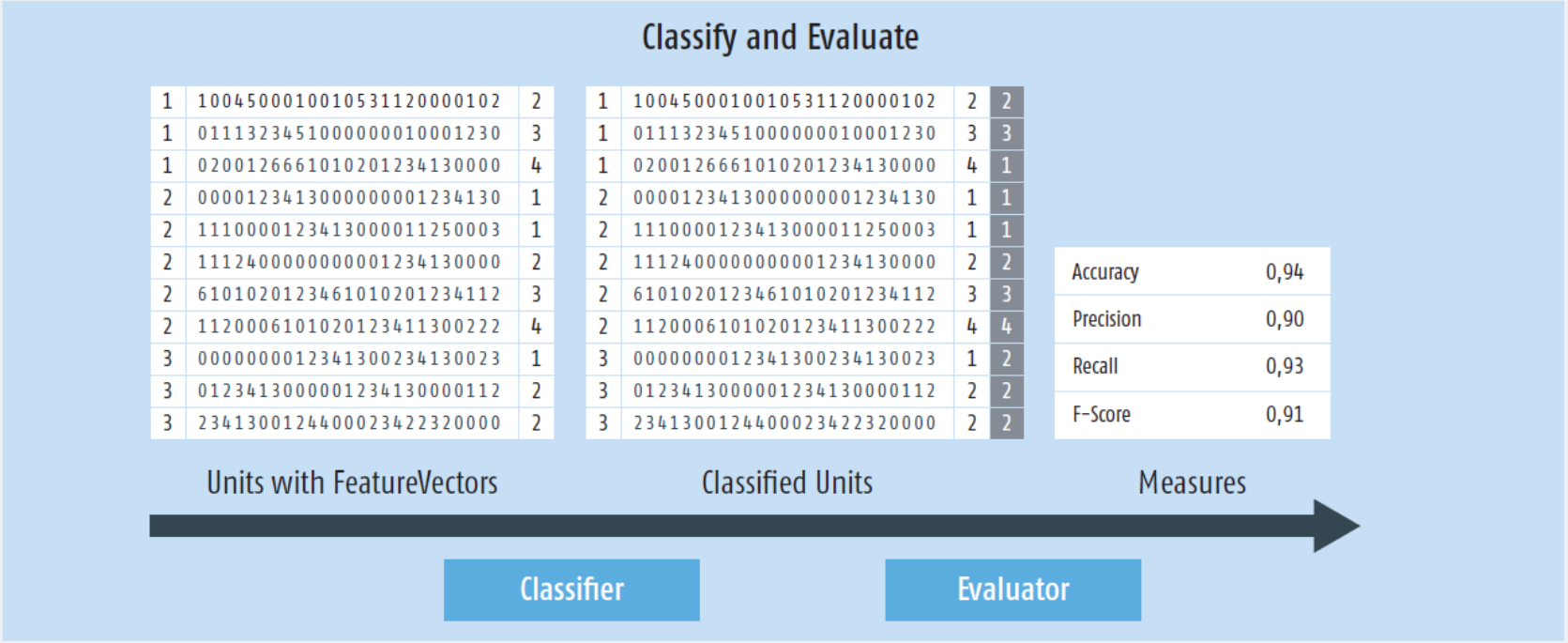
Preprocessing and training data generation



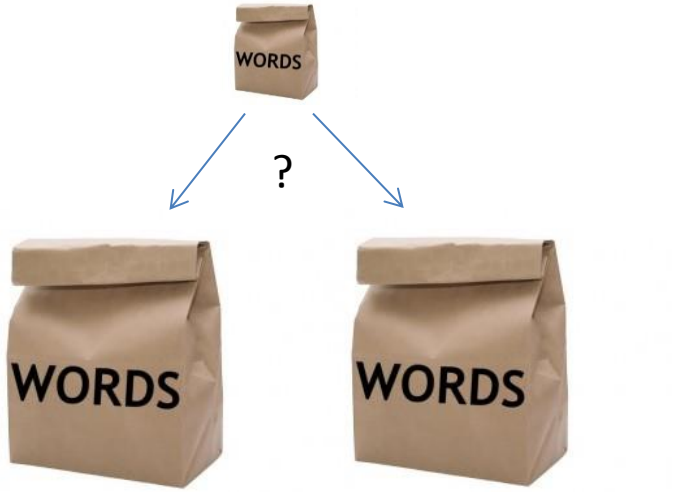
Feature generation and weighting



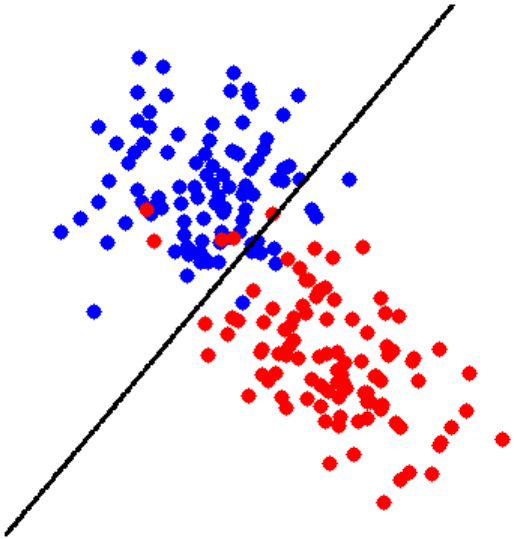
Classification and evaluation



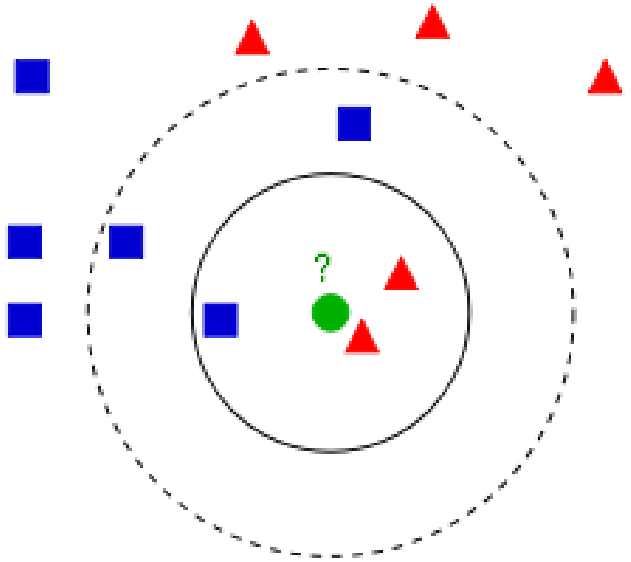
Naive Bayes Classifier



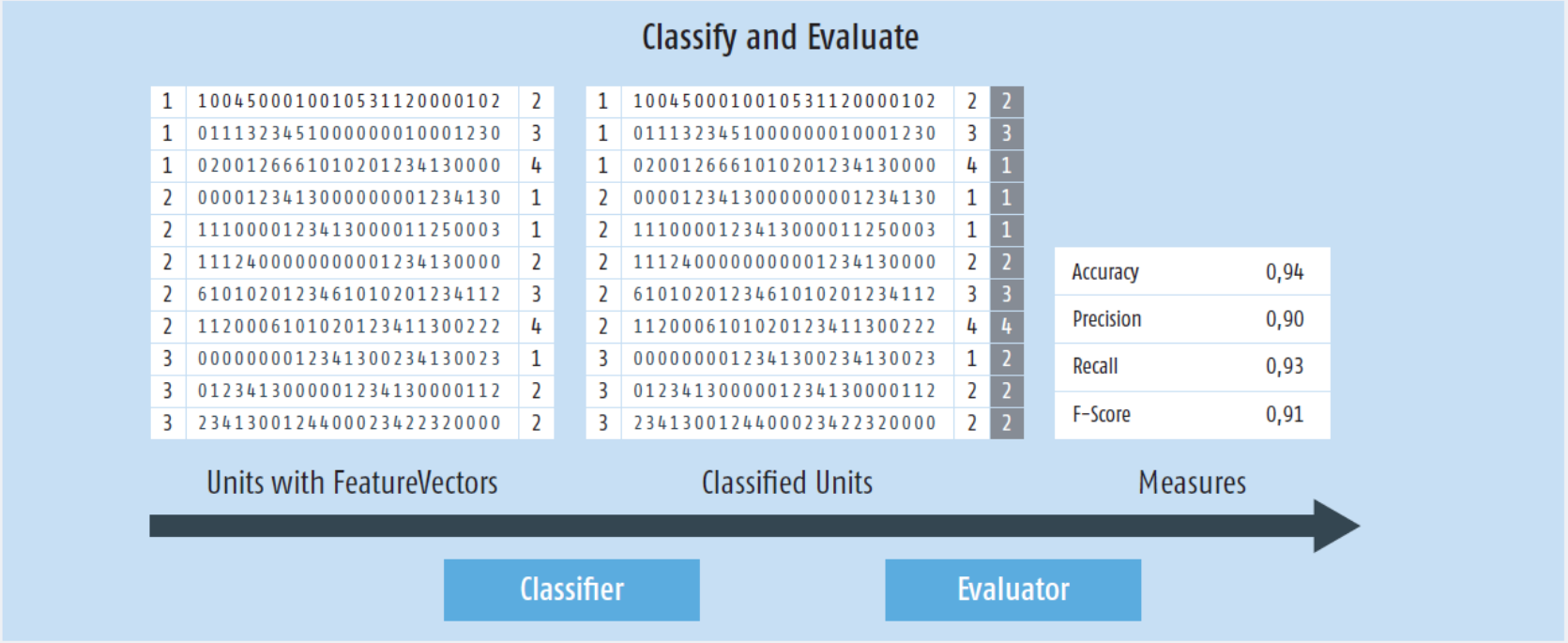
Linear Classifiers



KNN Classifier



Classification and evaluation



Results / Ranking of experiments

Nr.	F-score	Preci-sion	Recall	Accu-racy	Classifier	Distance	Quan-tifier	Feat-ures
1	0.95	0.98	0.92	0.97	KNN (k=4)	Cosinus	LogLike	2 & 3 grams
2	0.93	0.96	0.91	0.93	SVM	-	LogLike	3 & 4 grams
3	0.92	0.92	0.92	0.92	Rocchio	Cosinus	LogLike	2 & 3 grams
4	0.66	0.54	0.85	0.67	Naive Bayes	-	-	3 grams

Results: Insights

1. As expected, Naive Bayes classifier delivers the baseline results
2. KNN performs best, but is also the slowest algorithm
3. Linear classifiers also deliver respectable – and faster – results (but poorer ones than KNN). SVM performs slightly better than Rocchio.

Discussion

- Problems while adapting the algorithm to the BIBB-Database
  - Anonymized models vs. Not anonymized original data to classify
  - Mix of different encodings inside the database
  - Time requirements of the KNN-classifier
- Current tasks
  - Using classified sections for information extraction process
  - Analysis of trends for extracted tools, tasks, and competences


THANK
YOU