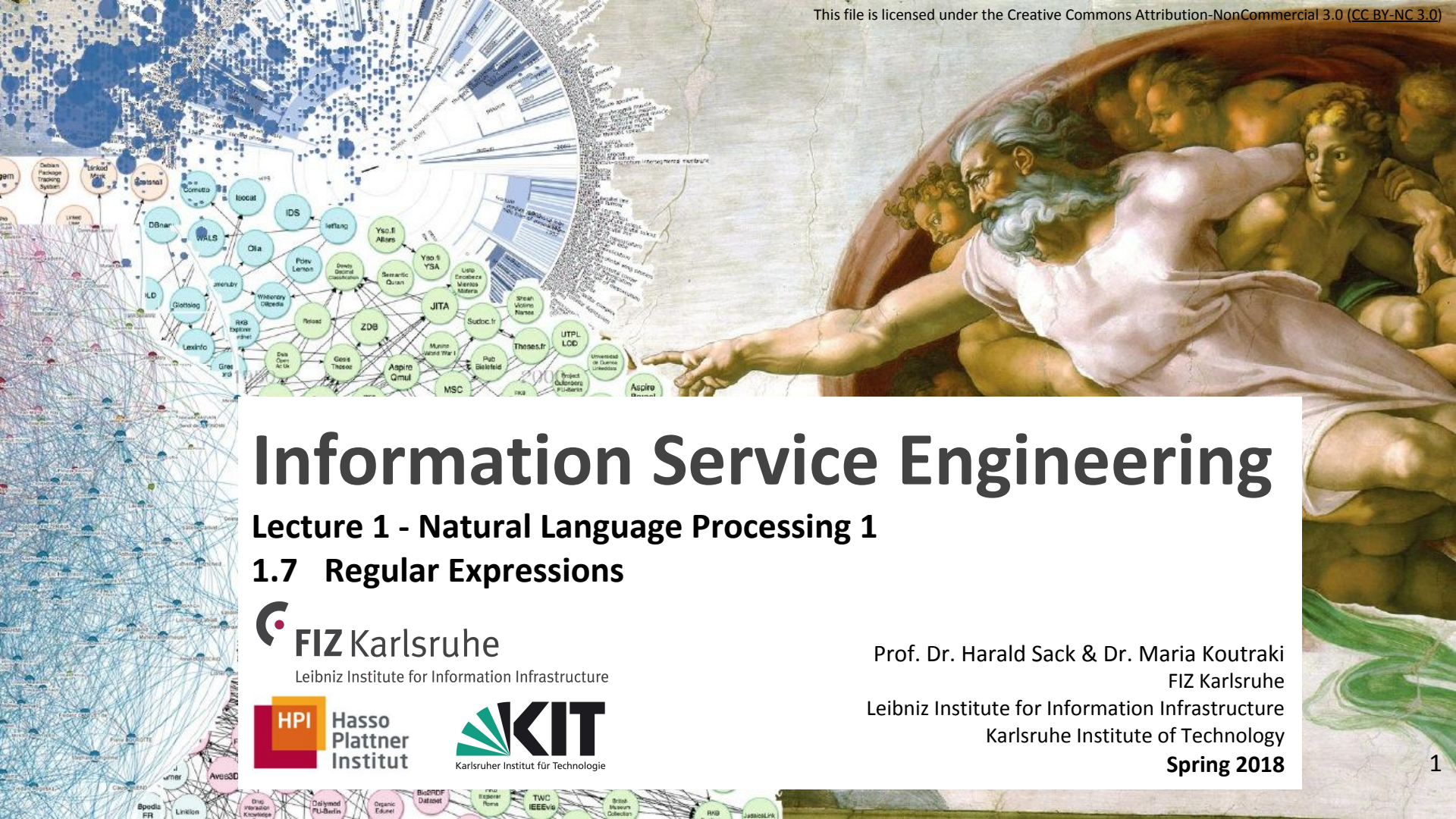




Information Service Engineering

Lecture 1 - Natural Language Processing 1

1.7 Regular Expressions



FIZ Karlsruhe

Leibniz Institute for Information Infrastructure



Hasso
Plattner
Institut



Karlsruher Institut für Technologie

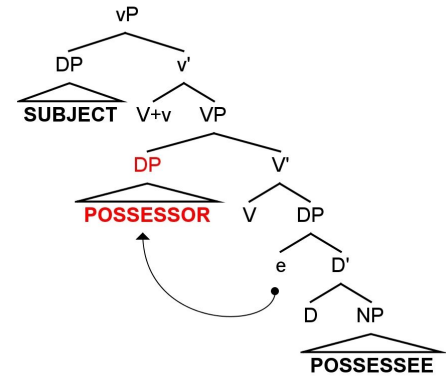
Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe

Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 1: Natural Language Processing

- 1.1 Information and Knowledge
- 1.2 Language and Understanding
- 1.3 NLP Techniques and Challenges
- 1.4 Linguistic Basics
- 1.5 Morphology
- 1.6 Evaluation, Precision and Recall
- 1.7 Regular Expressions**



Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Regular Expressions (RE)** are a formal language to define search patterns.
- RE can be used in **UNIX tools**
 - *grep, sed, awk,...*
- as well as in **programming languages**, as e.g.
 - *Python, Java, .NET*, etc.
- introduced by Kleene (1956), used for text search first by Thompson (1968)

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- RE are an algebraic notation that specifies simple classes of **strings**
- A **string** is defined as a sequence of symbols from an alphabet
- RE search requires a **pattern** that is to be searched and a **corpus** of texts to search through
- Most simple example with `grep`:
 - `grep Beethoven Beethovencorpus.csv`

Alternatively use online  tool:
<http://regexr.com/>

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- RE are an algebraic notation that specifies simple classes of **strings**
- A **string** is defined as a sequence of symbols from an alphabet
- RE search requires a **pattern** that is to be searched and a **corpus** of texts to search through
- Most simple example with `grep`:
 - `grep Beethoven Beethovencorpus.csv`

Alternatively use online  tool:
<http://regexr.com/>

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- Most simple example with grep:

- `grep Beethoven Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep Beethoven Beethovencorpus.csv
The Andante favori is a work for piano solo by Ludwig van Beethoven.
The other great passion of the young Mirabehn was the music of van Beethoven.
L.V. Beethoven spent the better part of his life in Vienna.
Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
Among the few composers writing for the orphica was Ludvig von Beethoven
Naue went to Vienna to study briefly with von Beethoven.
Bonn is the birthplace of Ludwig van Beethoven (born 1770).
"Johann van Beethoven joined the court, primarily as a singer, in 1764."
Camper van Beethoven were inactive between late 1990 and 1999.
"Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round."
KA000019:Downloads hsa$ █
```

- `grep -c Beethoven Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep -c Beethoven Beethovencorpus.csv
10
KA000019:Downloads hsa$ █
```

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions



Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Disjunction** example with `grep`:

- `grep 199[02] Beethovencorpus.csv`

```
[KA000019:Downloads hsa$ grep 199[02] Beethovencorpus.csv  
Camper van Beethoven were inactive between late 1990 and 1999.  
Beetehoven hit theaters in april 1992.  
KA000019:Downloads hsa$ █
```

- **Range** specification example with `grep`:

- `grep 199[0-9] Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep 199[0-9] Beethovencorpus.csv  
Camper van Beethoven were inactive between late 1990 and 1999.  
Beetehoven hit theaters in april 1992.  
KA000019:Downloads hsa$ █
```

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions



Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Disjunction** example with `grep`:

- `grep 199[02] Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep 199[02] Beethovencorpus.csv
Camper van Beethoven were inactive between late 1990 and 1999.
Beethoven hit theaters in april 1992.
KA000019:Downloads hsa$ █
```

- **Range specification** example with `grep`:

- `grep 199[0-9] Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep 199[0-9] Beethovencorpus.csv
Camper van Beethoven were inactive between late 1990 and 1999.
Beethoven hit theaters in april 1992.
KA000019:Downloads hsa$ █
```

- | | |
|--------------|-----------------------|
| [0-9] | any single digit |
| [a-z] | any lower case letter |
| [A-Z] | any upper case letter |

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Negation** example with `grep`:

- `grep 199[^2] Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep 199[^2] Beethovencorpus.csv
Camper van Beethoven were inactive between late 1990 and 1999.
KA000019:Downloads hsa$
```

- | | |
|---------------------|-------------------------|
| <code>[^0-9]</code> | not a digit |
| <code>[^a-z]</code> | not a lower case letter |
| <code>[^A-Z]</code> | not upper case letter |
- | | |
|--------------------|-------------------|
| <code>[^sS]</code> | neither s nor S |
| <code>[^\.]</code> | not a period |
| <code>[e^]</code> | either e or ^ |
| <code>a^b</code> | the pattern "a^b" |

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Wildcard** example with `grep`:

- `grep 199. Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep 199. Beethovencorpus.csv
Camper van Beethoven were inactive between late 1990 and 1999.
Beethoven hit theaters in april 1992.
KA000019:Downloads hsa$
```

- `.` matches any character (except carriage return)

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Repetitive pattern** example with `grep`:

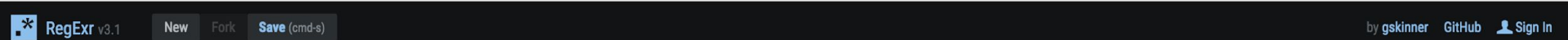
- `grep Be+thoven Beethovencorpus.csv`

- for Mac OSX: `grep "Be\+thoven" Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep "Be\+thoven" Beethovencorpus.csv
The Andante favori is a work for piano solo by Ludwig van Beethoven.
The other great passion of the young Mirabehn was the music of van Beethoven.
L.V. Beethoven spent the better part of his life in Vienna.
Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
Among the few composers writing for the orphica was Ludvig von Beethoven
Naue went to Vienna to study briefly with von Beethoven.
Bonn is the birthplace of Ludwig van Beethoven (born 1770).
"Johann van Beethoven joined the court, primarily as a singer, in 1764."
Camper van Beethoven were inactive between late 1990 and 1999.
"Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round."
KA000019:Downloads hsa$ █
```

- would also include “Bethoven”, “Beethoven”, “Beeethoven”, etc.

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions



/Beethoven/g

Text

10 matches (0.3ms)

The Andante favori is a work for piano solo by Ludwig van Beethoven. The other great passion of the young Mirabeau was the music of van Beethoven. L.V. Beethoven spent the better part of his life in Vienna. Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962. Among the few composers writing for the orchestra was Ludwig van Beethoven. Beethoven, too, used this key extensively in his second piano concerto. Naue went to Vienna to study briefly with von Beethoven. Bonn is the birthplace of Ludwig van Beethoven (born 1770). Johann van Beethoven joined the court, primarily as a singer, in 1764. Camper van Beethoven were inactive between late 1990 and 1999. Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round. Beethoven hit theaters in April 1992.

<http://regexpr.com/>

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

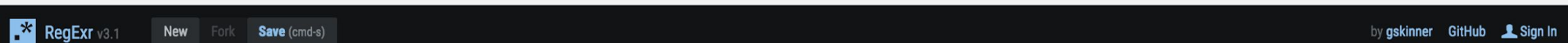
- **Optional and repetitive pattern example with grep:**

- `grep Beete*hoven Beethovencorpus.csv`

```
[KA000019:Downloads hsa$ grep Beete*hoven Beethovencorpus.csv
The Andante favori is a work for piano solo by Ludwig van Beethoven.
The other great passion of the young Mirabehn was the music of van Beethoven.
L.V. Beethoven spent the better part of his life in Vienna.
Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
Among the few composers writing for the orphica was Ludvig von Beethoven
Naue went to Vienna to study briefly with von Beethoven.
Bonn is the birthplace of Ludwig van Beethoven (born 1770).
"Johann van Beethoven joined the court, primarily as a singer, in 1764."
Camper van Beethoven were inactive between late 1990 and 1999.
"Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round."
Beetehoven hit theaters in april 1992.
KA000019:Downloads hsa$
```

- would also include “Beeteethoven”, “Beeteeethoven”, “Beeteeeeethoven”, etc.

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions



/Beete*hoven/g

Flags ▾

11 matches (0.3ms)

The Andante favori is a work for piano solo by Ludwig van Beethoven. The other great passion of the young Mirabeau was the music of van Beethoven. L.V. Beethoven spent the better part of his life in Vienna. Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962. Among the few composers writing for the orchestra was Ludwig van Beethoven. Beethoven, too, used this key extensively in his second piano concerto. Naue went to Vienna to study briefly with von Beethoven. Bonn is the birthplace of Ludwig van Beethoven (born 1770). Johann van Beethoven joined the court, primarily as a singer, in 1764. Camper van Beethoven were inactive between late 1990 and 1999. Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round. Beethoven hit theaters in April 1992.

<http://regexr.com/>

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Anchor** example with `grep`:

- `grep ^Beethoven Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep ^Beethoven Beethovencorpus.csv
KA000019:Downloads hsa$
```

- `grep ^.Beethoven Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep ^.Beethoven Beethovencorpus.csv
"Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round."
KA000019:Downloads hsa$
```

- `grep Beethoven.$ Beethovencorpus.csv`

```
KA000019:Downloads hsa$ grep Beethoven.$ Beethovencorpus.csv
The Andante favori is a work for piano solo by Ludwig van Beethoven.
The other great passion of the young Mirabehn was the music of van Beethoven.
Naue went to Vienna to study briefly with von Beethoven.
KA000019:Downloads hsa$
```

- **^** matches word only at **start of the line**
\$ matches word only at the **end of the line**

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **String disjunction** example with `grep`:

- `grep Vienna|Bonn Beethovencorpus.csv`

Mac OSX: `grep "Vienna\\|Bonn" Beethovencorpus.csv`

```
[KA000019:Downloads hsa$ grep "Vienna\\|Bonn" Beethovencorpus.csv
L.V. Beethoven spent the better part of his life in Vienna.
Naue went to Vienna to study briefly with von Beethoven.
Bonn is the birthplace of Ludwig van Beethoven (born 1770).
KA000019:Downloads hsa$ █
```

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Repetition** example with `grep`:

- `grep [0-9]{4} Beethovencorpus.csv`

Mac OSX: `grep "[0-9]\{4\}" Beethovencorpus.csv`

```
[KA000019:Downloads hsa$ grep "[0-9]\{4\}" Beethovencorpus.csv
```

```
Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
```

```
Bonn is the birthplace of Ludwig van Beethoven (born 1770).
```

```
"Johann van Beethoven joined the court, primarily as a singer, in 1764."
```

```
Camper van Beethoven were inactive between late 1990 and 1999.
```

```
Beethoven hit theaters in april 1992.
```

```
KA000019:Downloads hsa$ █
```

- **Operator precedence**

- Parenthesis `()`
- Counters `* + ? {}`
- Sequences and anchors `the ^my end$`
- Disjunction `|`

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- Some characters need to be **backslashed**:

RE	Match
*	An asterisk
\.	A period
\?	A question mark
\n	An newline
\t	A tab
\,	A comma

- All functional characters that are to be used as '**characters only**' in a pattern must be backslashed

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- Advanced operators:

RE	Expansion	Match
\d	[0-9]	Any digit
\D	[^0-9]	Any non digit
\w	[a-zA-Z0-9_]	Any alphanumeric + underscore
\W	[^\w]	Any non-alphanumeric
\s	[\r\t\n\f]	whitespace
\S	[^\s]	non-whitespace

Regular Expressions

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- Numeric ranges:

RE	Match
*	Zero or more occurrences of previous character or expression
+	One or more occurrences of previous character or expression
?	Exactly zero or one occurrence of previous character or expression
{ n }	n occurrences of previous character or expression
{ n , m }	From n to m occurrences of previous character or expression
{ n , }	At least n occurrences of previous character or expression

Synonyms and Variations

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

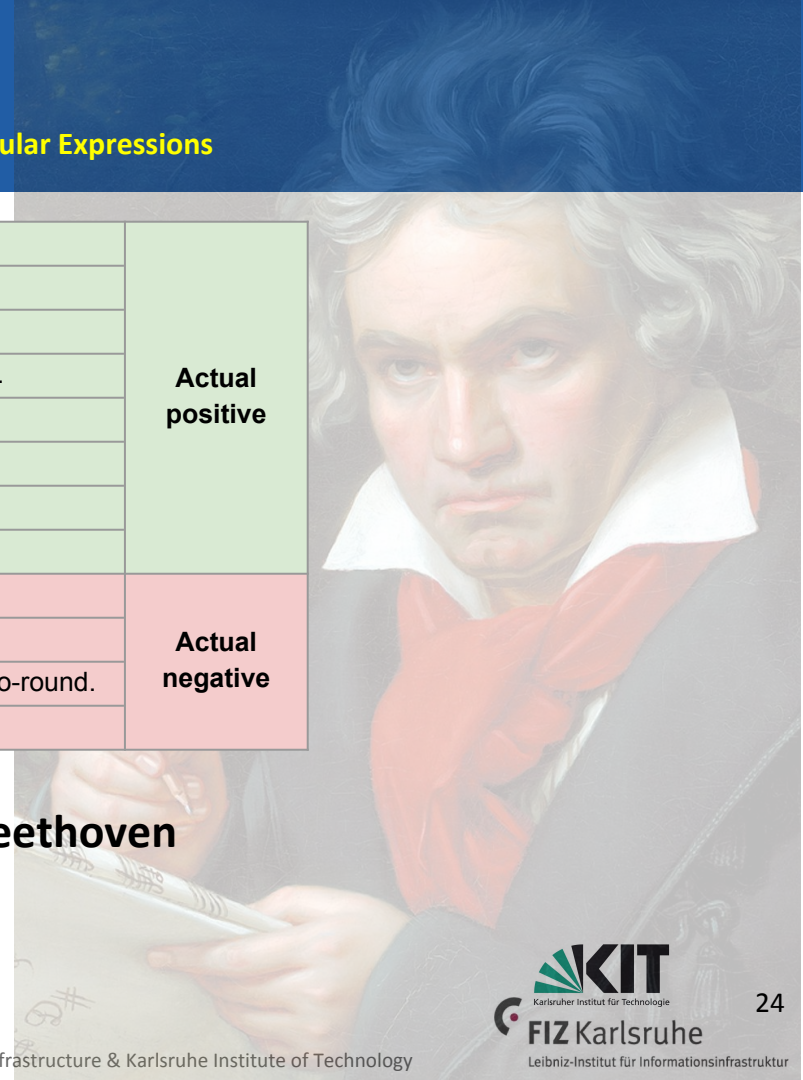
- If we are searching for **all occurrences of an entity in a text**, we have to consider **synonyms** and **variations** of its name
 - **Real synonyms** e.g. mobile phone -> cell phone, cellular telephone
 - **Quasi synonyms** e.g. mobile phone -> flip phone, mobile
 - **Upper case variations** e.g. cell phone and Cell phone
 - **Orthographic variations** e.g. cell phone and cell-phone
 - **Plural forms** e.g. cell phone and cell phones
 - **Typographic errors** e.g. cellular phone
 - **Related topics** e.g. cellphone video, cellular radio, phone carrier

Experiment

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.	Actual positive
2	The other great passion of the young Mirabehn was the music of van Beethoven.	
3	L.V. Beethoven spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.	
5	Among the few composers writing for the orphica was Ludvig von Beethoven	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von Beethoven.	
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).	
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.	Actual negative
10	Camper van Beethoven were inactive between late 1990 and 1999.	
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task:** Identify sentences that refer to **Ludwig van Beethoven**
- **Another Algorithm:** *RE Match* with “**Bee*t+hoven**”



Experiment

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- **Another Algorithm: RE Match with “*Bee*t+hoven*”**
 - Identified **11 lines** (1,2,3,4,5,6,7,8,9,10,11) as **positive**,
 - **8 lines** of it (1,2,3,4,5,7,8) are **actual positive** (**true positive**),
 - **3 lines** of it (9,10,11) are **actual negative** (**false positive**)
 - Identified **1 line** (12) as **negative**,
 - **1 line** of it (12) is **actual negative** (**true negative**)

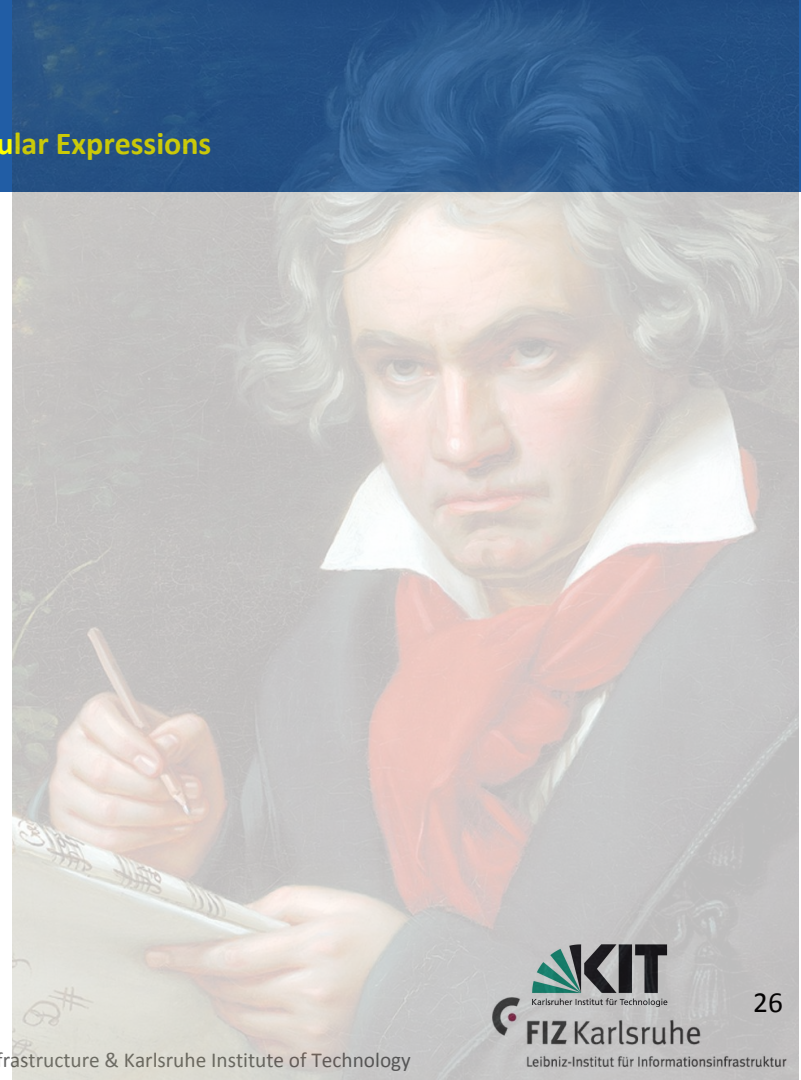
		predicted	
		true	false
actual	true	8	0
	false	3	1

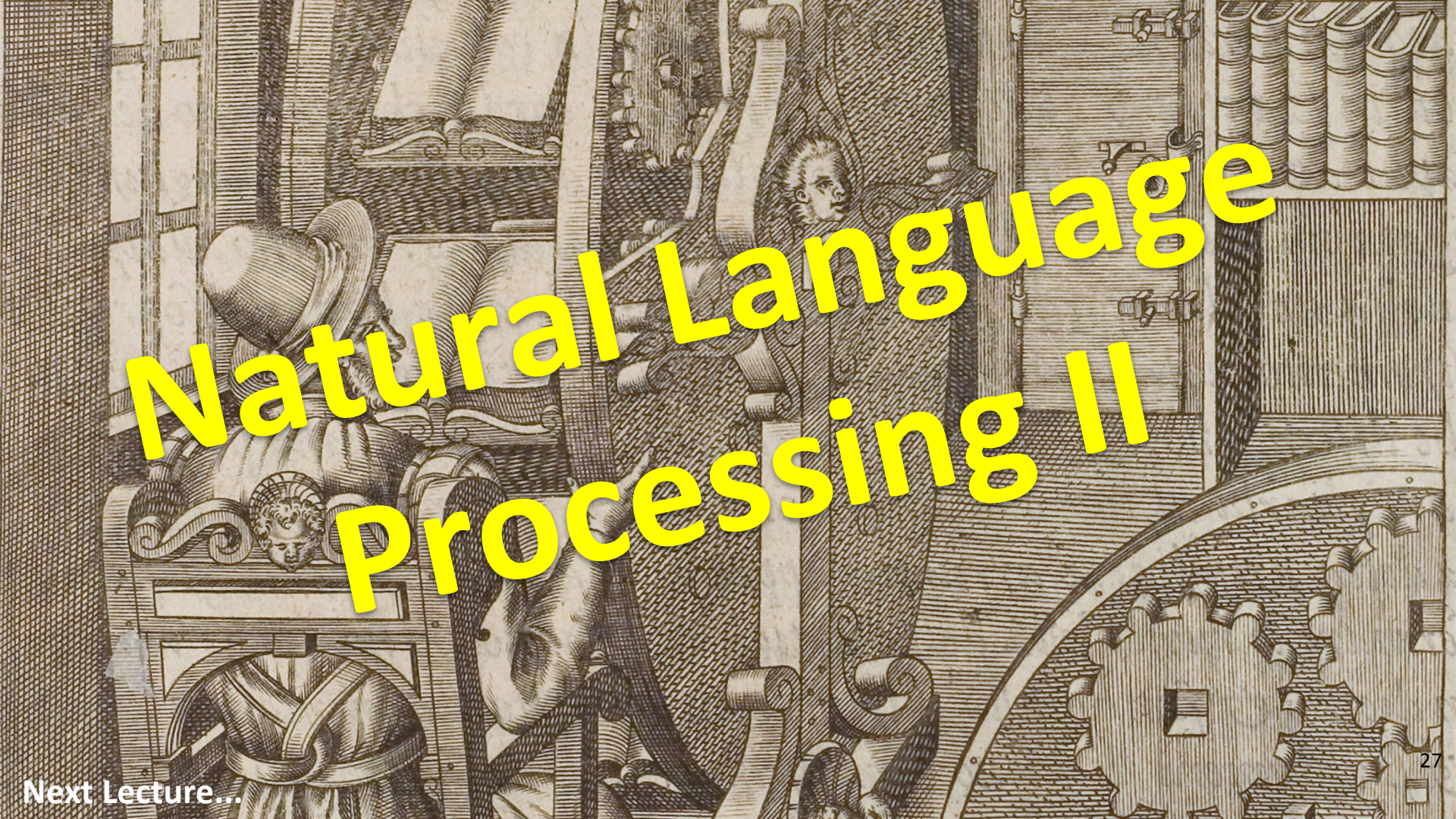
- **Precision** = $8/11$ = **72,7%**
- **Recall** = $8/8$ = **100%**
- F_1 = $2 (0.727*1)/(0.727+1)$
= 0.842 = **84.2%**

Experiment

Information Service Engineering - Lecture 1 Natural Language Processing - 1.7 Regular Expressions

- Can you obtain $F_1 = 100\%$?
- If so,
will this be the “perfect” Beethoven Identifier?





Natural Language Processing II

Next Lecture...

Picture References:

- P.23/25: Ludwig van Beethoven, Portrait by Joseph Karl Stieler (1820), <https://commons.wikimedia.org/wiki/File:Beethoven.jpg>
- P.26: Le diverse et artificiose machine del Capitano Agostino Ramelli, https://commons.wikimedia.org/wiki/File:Le_diverse_et_artificiose_machine_del_Capitano_Agostino_Ramelli_Figure_CLXXXVIII.jpg