# Regular Expressions

reg(ular expressions? | ex(p | es)?)

#### Inhaltsübersicht

- Überblick
- Regular Expression Engine
- Zeichenklassen
- Quantifier
- Modifier
- ... alle weiteren Funktionen

### Regular Expressions

#### Was?

Beschreibung von Mustern in Zeichenketten

#### Warum?

- Suche Textstellen in Dokumenten
- Validierung von Eingaben
- Textstellen aus Dokumenten extrahieren
- Textstellen einfügen oder ersetzen

### Regex Engine

#### Eingabe

- Subjekt: Ein Text (lange Zeichenkette) auf den ein Muster (Regular Expression) angewandt wird
- Muster: Zeichenkette welche eine Musterbeschreibung enthält

#### Ausgabe

- Match: ja oder nein
- Capturings: gefundene Teststellen

Abcd e aeaeafa fef a faefaefafich theth lakefeafealfjlaefflaf afa



### Regex Engine

- Das Muster wird positionsweise auf das Subjekt angewandt.
- Stimmt eine Position des Musters mit einer Position der Subjekts überein, wird die nächste Position des Musters geprüft.
- Stimmt eine Position des Musters nicht mit der des Subjekts überein, kommt es zum Backtracking. Alle Permutationen werden geprüft.
- Sind alle Position des Musters erfolgreich geprüft gilt dies als Match.

### Regex Engine Beispiel

Subjekt: abc dde abcd aabb bb aa bb ab9d ab?d abdd abkkkd

Muster: ab[a-e]d?

Matches: ?

### Regex Engine Beispiel

Subjekt: <u>abc</u> dde <u>abcd</u> aabb bb aa bb ab9d ab?d <u>abdd</u> abkkkd

Muster: ab[a-e]d?

Matches: abc, abcd, abdd

### ASCII – Das ursprüngliche Regex Alphabet

0 0 000 NUL (null)	Dec	H	Oct	Cha	r	Dec	Нх	Oct	Html	Chr	Dec	Нх	Oct	Html	Chr	Dec	Нх	Oct	Html Ch	<u>ır</u>
1 1 001 SOH (start of heading) 2 2 002 STX (start of text) 3 3 21 041 ! ! 6 4 4 101 A A 9 7 61 141 a a 8 2 2 002 STX (start of text) 3 3 20 043 # # 6 4 4 102 B B 9 8 62 142 b b 3 3 003 ETX (end of transmission) 5 5 005 ENQ (enquiry) 6 6 006 ACK (acknowledge) 3 8 26 046 & 6 7 7 007 BEL (bell) 3 9 27 047 ' ' 8 8 010 BS (backspace) 9 9 011 TAB (horizontal tab) 10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 D1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 2) 20 14 024 DC4 (device control 4) 21 15 025 MAK (negative acknowledge) 22 16 026 SYM (synchronous idle) 23 17 027 ETB (end of trans. block) 25 19 031 EM (end of medium) 25 30 10 036 FS (record separator) 26 10 036 FS (record separator) 27 10 036 FS (record separator) 28 10 036 FS (record separator) 29 10 035 FS (record separator) 20 10 036 FS (record separator) 20 10 036 FS (record separator) 20 10 04 04 (48 #35; # 66 42 102 04 #36; D 66 42 102 04 #66; D 70 40 106 06 #70; F 10 00 64 144 06 #101; e 10 04 06 06 #70; F 10	0	0	000	NUL	(null)	32	20	040		Space	64	40	100	@	0	96	60	140	`	
3 3 003 ETX (end of text) 4 4 004 ECT (end of transmission) 5 5 005 ENO (enquiry) 37 25 045 6#33; \$ 68 44 104 6#68; D 100 64 144 6#100; d 68 006 ACK (acknowledge) 38 26 046 6#38; 6 7 7 007 BEL (bell) 39 27 047 6#33; † 7 007 BEL (bell) 39 27 047 6#33; † 70 46 106 6#70; F 102 66 146 6#102; f 7 1 47 107 6#71; G 103 67 147 6#103; g 8 8 010 BS (backspace) 40 28 050 6#40; ( 72 44 110 6#72; H 104 68 150 6#104; h 9 9 011 TAB (horizontal tab) 41 29 051 6#41; ) 42 2A 052 6#42; † 43 2B 053 6#43; + 75 4B 113 6#75; K 107 6B 153 6#106; j 43 2B 053 6#45; - 74 4A 112 6#76; L 107 6B 153 6#100; k 45 2D 055 6#45; - 75 4B 113 6#75; K 107 6B 153 6#100; k 46 2E 056 6#46; . 76 45 114 6#76; L 107 6B 153 6#100; k 47 2F 057 6#47; / 74D 115 6#77; M 109 6D 155 6#100; k 48 30 060 6#46; . 78 4F 116 6#78; N 110 6E 156 6#110; p 17 11 021 DC1 (device control 1) 49 31 061 6#49; l 48 30 060 6#45; - 80 50 120 6#80; p 112 70 160 6#112; p 17 11 021 DC1 (device control 2) 18 12 022 DC2 (device control 3) 19 13 023 DC3 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 25 19 031 EM (end of medium) 25 19 031 EM (end of medium) 25 19 032 EM (end of medium) 25 19 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 45 20 076 6#86; V 40 28 070 6#35; # 36 24 044 6#36; \$ 68 44 104 6#69; E 101 65 145 6#102; f 66 44 104 6#69; E 101 65 145 6#102; f 70 46 106 6#70; F 102 66 146 6#102; f 71 47 107 6#71; H 103 6#51; G 103 67 147 6#103; g 72 44 112 6#72; H 104 68 150 6#104; h 73 49 111 6#73; H 105 69 151 6#105; i 74 4A 112 6#72; H 104 68 150 6#104; h 75 4B 113 6#72; H 104 68 150 6#104; h 75 4B 113 6#75; K 107 6B 153 6#105; i 104 68 154 6#103; g 75 4B 113 6#75; K 107 6B 153 6#105; i 105 69 151 6#105; i 107 44 11 6#76; L 107 6#112; p 108 60 155 6#101; p 109 60 155 6#101; p 110 115 6#77; M 109 6D 155 6#101; p 110 115 6#77; M 109 6D 155 6#101; p 111 6F 157 6#111; o 111 6F 157 6#1	1					33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
4 4 004 EOT (end of transmission) 5 5 005 ENQ (enquiry) 6 6 006 ACK (acknowledge) 7 7 007 BEL (bell) 8 8 010 BS (backspace) 9 9 011 TAB (horizontal tab) 10 A 012 LF (ML line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 10 202 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 4) 19 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 25 19 031 EM (end of medium) 25 10 036 RS (group separator) 26 16 036 RS (frecord separator) 27 18 023 GS (group separator) 28 10 036 RS (group separator) 29 10 035 GS (group separator) 20 14 Cecord separator) 20 15 CR (sing wearator) 20 17 18 032 GS (group separator) 20 18 030 GS (group separator) 20 18 030 GS (group separator) 20 19 03 GS (group separator) 20 10 03 CS (device control 4) 21 10 03 GS (group separator) 21 10 03 GS (group separator) 22 16 036 RS (group separator) 23 17 03 07 GS (group separator) 24 18 030 GR (tecord separator) 25 19 031 EM (tecord separator) 26 18 036 RS (tecord separator) 27 18 036 RS (tecord separator) 28 18 030 CR (tecord separator) 28 18 030 CR (tecord separator) 29 10 035 GS (group separator) 20 10 03 CR (tecord separator) 21 10 03 CR (tecord separator) 22 10 03 CR (tecord separator) 23 10 03 CR (tecord separator) 24 18 030 CR (tecord separator) 25 19 031 EM (tecord separator) 26 18 032 CR (tecord separator) 27 18 033 CR (tecord separator) 28 18 030 CR (tecord separator) 29 18 033 CR (tecord separator) 29 18 033 CR (tecord separator) 29 18 033 CR (tecord separator) 29 18 034 CR (tecord separator) 29 18 035 CR (	2	2	002	STX	(start of text)	34	22	042	 <b>4</b> ;	"	66	42	102	B	В	98	62	142	b	b
5 5 005 ENQ (enquiry) 6 6 006 ACK (acknowledge) 7 7 007 BEL (bell) 8 8 010 BS (backspace) 9 9 011 TAB (horizontal tab) 10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SyN (synchronous idle) 23 17 027 ETB (end of trans. block) 25 19 031 EM (end of medium) 26 16 036 CR (group separator) 27 18 036 CR (group separator) 28 25 045 Ƶ % 70 44 6106 F F 70 446 106 F F 70 446 106 F F 70 447 107 G G 103 67 147 g g 70 47 ' ' 71 47 107 G G 103 67 147 g g 71 44 112 I I 105 69 151 i i 106 6A 152 j j 106 6A 152 j j 106 6A 152 k k 107 6B 153 k k 10	3	3	003	ETX	(end of text)	35	23	043	#	#	67	43	103	a#67;	C					
8 6 6 006 ACK (acknowledge) 7 7 007 BEL (bell) 8 8 010 BS (backspace) 9 9 011 TAB (horizontal tab) 10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 25 19 031 EM (end of medium) 26 16 036 CRS (group separator) 27 18 036 CRS (group separator) 28 10 046 046 4#38; 6 38 26 046 4#38; 6 39 27 047 4#39; 7 44 107 4#71; 6 103 67 147 4#103; g 71 44 110 4#71; H 72 48 110 4#72; H 73 49 111 4#72; H 74 41 112 4#72; H 75 48 111 4#72; H 75 48 113 4#75; K 107 68 153 4#105; I 107 69 151 4#105; I 108 60 154 4#105; I 109 60 152 4#106; J 109 60 155 4#105; I 109	4	4	004	EOT	(end of transmission)				-											
7 7 007 BEL (bell) 39 27 047 6#39; 71 47 107 6#71; 6 103 67 147 6#103; 9 8 8 010 BS (backspace) 40 28 050 6#40; 72 48 110 6#72; H 104 68 150 6#104; h 105 64 151 6#105; i 10 A 012 LF (NL line feed, new line) 42 2A 052 6#42; 7 74 4A 112 6#74; J 106 6A 152 6#106; j 11 B 013 VT (vertical tab) 43 2B 053 6#44; 75 4B 113 6#75; K 107 6B 153 6#107; k 12 C 014 FF (NP form feed, new page) 44 2C 054 6#44; 75 4B 113 6#75; K 107 6B 153 6#107; k 12 C 014 FF (NF form feed, new page) 44 2C 054 6#44; 75 4B 113 6#75; M 109 6D 155 6#109; m 14 E 016 S0 (shift out) 46 2E 056 6#46; - 77 4D 115 6#77; M 109 6D 155 6#109; m 15 F 017 SI (shift in) 47 2F 057 6#47; / 79 4F 117 6#79; O 111 6F 157 6#111; O 16 10 020 DLE (data link escape) 48 30 060 6#48; O 18 120 6#80; P 112 70 160 6#112; P 17 1021 DC1 (device control 1) 49 31 061 6#49; 1 18 12 12 6#81; O 113 71 161 6#113; Q 18 12 022 DC2 (device control 3) 51 33 063 6#51; 3 83 53 123 6#83; S 115 73 163 6#115; S 20 14 024 DC4 (device control 4) 52 34 064 6#52; 4 8 55 125 6#85; U 117 75 165 6#117; U 23 17 027 ETB (end of trans. block) 55 37 067 6#55; 7 87 57 127 6#87; U 119 77 167 6#119; W 24 18 030 CAN (cancel) 55 38 070 6#55; 8 38 070 6#55; 8 10 034 ESC (escape) 59 3B 073 6#59; 9 10 035 CS (group separator) 61 30 075 6#61; 9 35 D156 6#94; 1 126 7E 176 6#125; 3 126 6#10; 1 126 6#126; 1 126 7E 176 6#125; 3 126 6#94; 1 126 7E 176 6#125; 3 126 6#10; 1 126 6#126; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 126 6#10; 1 127 0 126 6#10; 1 127 0 127	5													-		4000	100000			
8 8 010 BS (backspace) 9 9 011 TAB (horizontal tab) 10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NF form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 25 19 032 GS (group separator) 26 16 036 RS (record separator) 26 16 036 RS (record separator) 27 18 030 CR (group separator) 28 0 014 024 CC4 (file separator) 29 10 035 GS (group separator) 30 16 036 RS (record separator) 40 28 050 4#40; ( 41 29 051 4#41; ) 42 20 052 4#42; 7 44 41 110 4#73; I 100 66 A 152 6#106; j 44 22 054 6#41; ) 74 44 111 6#73; I 100 66 A 152 6#106; j 75 48 113 6#77; M 100 66 153 6#107; k 76 4C 114 6#76; L 108 6C 154 6#108; l 77 4D 115 6#77; M 109 6D 155 6#109; m 78 4E 116 6#78; N 110 6E 155 6#109; m 78 4E 116 6#78; N 110 6E 155 6#110; n 79 4F 117 6#79; O 111 6F 157 6#111; O 80 50 120 6#48; O 80 50 120 6#30; P 112 70 160 6#112; P 113 023 DC3 (device control 2) 81 50 120 6#86; O 112 6#81; O 113 71 161 6#113; O 116 6#112; P 117 71 161 6#113; O 116 6#112; P 117 71 161 6#113; O 116 6#112; P 117 71 161 6#112; O 116	6					100														
9 9 011 TAB (horizontal tab) 10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 23 17 027 ETB (end of medium) 25 19 031 EM (end of medium) 25 19 032 SUB (substitute) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FF (file separator) 30 1E 036 RS (record separator) 30 1E 036 RS (record separator) 41 29 051 «#41; ) 42 2A 052 «#42; * 43 2B 053 «#43; + 74 4A 112 «#74; J 76 4C 114 «#76; L 77 4D 115 «#77; M 78 4E 116 «#78; N 79 4F 117 «#79; O 100 6A 152 «#106; j 77 4D 115 «#77; M 79 4F 117 «#79; O 110 6 6A 152 «#106; j 77 4D 115 «#77; M 79 4F 117 «#79; O 110 6 6A 152 «#106; j 77 4D 115 «#77; M 79 4F 117 «#79; O 110 6 6A 152 «#106; j 77 4D 115 «#77; M 79 4F 117 «#79; O 110 6 6A 152 «#106; j 77 4D 115 «#77; M 79 4F 117 «#79; O 111 6F 157 «#111; O 78 4E 116 «#78; N 79 4F 117 «#79; O 111 6F 157 «#111; O 78 4E 116 «#78; N 79 4F 117 «#79; O 111 6F 157 «#111; O 78 4E 116 «#78; N 79 4F 117 «#79; O 111 6F 157 «#111; O 78 4E 116 «#78; N 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O 79 4F 117 «#79; O 111 6F 157 «#111; O	7				(bell)	10000					alle Co					The state of the s				
10 A 012 LF (NL line feed, new line) 11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 25 10 035 GS (group separator) 26 16 036 RS (record separator) 27 10 035 GS (group separator) 28 10 036 RS (record separator) 30 1E 036 RS (record separator) 43 2B 053 + + 44 2C 054 , , 45 2B 055 ƽ - 47 4A 112 J J 106 6A 152 j J 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K K 107 6B 153 k k 75 4B 113 K M 108 6C 154 l l 75 4B 113 M M 109 6D 155 m m 76 4C 114 L L 108 6C 154 l	8									•	A Control									
11 B 013 VT (vertical tab) 12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 43 2B 053 «#43; + 44 2C 054 «#44; , 45 2D 055 «#45; - 45 2D 055 «#45; - 46 2E 056 «#46; . 47 2F 057 «#47; / 48 2E 056 «#46; . 48 2D 055 «#47; / 49 115 «#77; M 109 6D 155 «#109; m 48 2D 055 «#47; / 49 115 «#77; M 109 6D 155 «#109; m 49 3D 060 «#48; O 80 50 120 «#80; P 111 07 160 «#112; P 81 51 121 «#81; Q 111 371 161 «#113; Q 81 51 121 «#81; Q 113 71 161 «#113; Q 81 51 121 «#81; Q 113 71 161 «#113; Q 81 51 121 «#81; Q 113 71 161 «#113; Q 81 52 122 «#82; R 114 72 162 «#114; r 81 51 121 «#81; Q 113 71 161 «#113; Q 81 52 122 «#82; R 114 72 162 «#114; r 81 51 121 «#81; Q 113 71 161 «#113; Q 81 51 121 «#81; Q 113 71 161 «#113; Q 81 52 122 «#82; R 114 72 162 «#114; r 81 51 121 «#81; Q 113 71 161 «#113; Q 81 52 122 «#82; R 114 72 162 «#114; r 81 52 124 «#84; T 116 74 164 «#116; t 82 52 122 «#82; R 114 72 162 «#114; r 83 53 123 «#83; S 115 73 163 «#115; S 83 53 123 «#83; S 115 73 163 «#115; S 84 54 124 «#84; T 116 74 164 «#116; t 85 55 125 «#86; U 117 75 165 «#117; U 118 76 166 «#118; V 118 76 166 «#112; P 118 77 167 «#120; X 118 77 167 «#120;	_									•										
12 C 014 FF (NP form feed, new page) 13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 23 17 027 ETB (end of medium) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 38 1C 034 FS (file separator) 39 1E 036 RS (record separator) 44 2C 054 «#44; ,									4000000	100	100		2000000							
13 D 015 CR (carriage return) 14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 23 17 027 ETB (end of medium) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 30 1E 036 RS (record separator) 45 2D 055 «#45; - 46 2E 056 «#46; . 47 4F 17 «#79; 0 48 4E 116 «#78; N 79 4F 117 «#79; 0 110 11 6F 157 «#111; 0 110 6E 156 «#10; n 79 4F 117 «#79; 0 111 6F 157 «#111; 0 11 16 F 157 «#111; 0 11 17 1 16 «#112; p 11 7 7 1 66 «#145; E 11 7 7 1 66 «#115; F 11 7 7 1 66 «#115;	1000	5.55								+	F - 100-10	1000	No.							
14 E 016 SO (shift out) 15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 23 17 027 ETB (end of medium) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 46 2E 056 «#46; .  78 4E 116 «#78; N   110 6E 156 «#110; n   79 4F 117 «#79; 0   111 6F 157 «#111; 0   111 6F 157 «#112; D   111 6F 157 «#111; 0   111 6F 157 «#112; D   113 71 161 «#113; 0   111 6F 157 «#113; 0   111 6F 157 «#125; D   113 71 161 «#113; 0   111 6F 157 «#125; D   113 161 «#114; E   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113; 0   113 71 161 «#113;		- 7			, , , , , , , , , , , , , , , , , , , ,															
15 F 017 SI (shift in) 16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 47 2F 057 / / 48 30 060 0 0 48 30 060 0 1 48 121 y p &#						1000					The same of the									
16 10 020 DLE (data link escape) 17 11 021 DC1 (device control 1) 18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 30 1E 036 RS (record separator) 30 1E 036 RS (record separator) 48 30 060 0 0 49 31 061 1 1 50 32 062 2 2 82 52 122 R R 114 72 162 r L 81 51 121 Q Q 113 71 161 q Q 113 71 161 q Q 114 024 C4 (device control 4) 83 53 123 S S 115 73 163 s S 84 55 125 U U 117 75 165 u U 117 75 166 v V 118 76 166 v V 119 77 167 w W 119 77 167 y Y 110 120 020 020 020 020 020 020 020 020							0000	100000	1000		1000000									
17 11 021 DC1 (device control 1)								1000		-										
18 12 022 DC2 (device control 2) 19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 25 19 032 ESC (escape) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 30 1E 036 RS (rec							100													
19 13 023 DC3 (device control 3) 20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 30 1E 036 RS (record separator) 31 30 03 4#51; 3 33 063 4#51; 3 34 064 4#52; 4 34 064 4#52; 4 35 34 064 4#52; 4 36 4 54 124 4#84; T 36 4 54 124 4#84; T 37 16 4 4 4 4 116; t 38 55 125 4#85; U 38 57 127 4#87; U 38 58 59 125 4#85; U 38 58 120 4#84; T 38 59 123 4#84; T 38 59 123 4#84; T 38 59 123 4#84; T 38 59 124 4#84; T 38 59 125 4#86; U 38 59 126 4#86; V 38 59 127 4#87; U 38 59 128 4#89; V 38 59 129 4#86; U 38 5						100000	5000 aug	0.000												
20 14 024 DC4 (device control 4) 21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 30 1E 036 RS (record separator) 52 34 064 4 4 84 54 124 T T 116 74 164 t t 85 34 064 4 4 85 55 125 U U 117 75 165 u u 118 76 166 v v 119 77 167 w w 119 77 167 x x 120 78 170 x x 120 78 170 y y 121 79 171 y y 122 7A 172 z z 123 7B 173 { { 124 7C 174     125 7D 175 } } 126 7E 176 ~ ~						100													-	
21 15 025 NAK (negative acknowledge) 22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 4 36 066 4#54; 6 54 36 066 4#54; 6 55 37 067 4#55; 7 87 57 127 4#87; W 119 77 167 4#119; W 120 78 170 4#120; X 120 78 170 4#121; Y 121 79 171 4#121; Y 122 7A 172 4#122; Z 123 7B 173 4#123; ( 124 7C 174 4#124; I 125 7D 175 4#125; } 126 7E 176 4#126; ~						1000														
22 16 026 SYN (synchronous idle) 23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 4 36 066 6 6 54 36 066 6 6 55 37 067 7 7 87 57 127 W W 119 77 167 w W 120 78 170 x X 120 78 170 y Y 121 79 171 y Y 122 7A 172 z Z 123 7B 173 { ( 124 7C 174   I 125 7D 175 } } 126 7E 176 ~ ~	20	14	024	DC4	(device control 4)															
23 17 027 ETB (end of trans. block) 24 18 030 CAN (cancel) 25 19 031 EM (end of medium) 26 1A 032 SUB (substitute) 27 1B 033 ESC (escape) 28 1C 034 FS (file separator) 29 1D 035 GS (group separator) 30 1E 036 RS (record separator) 55 37 067 4#55; 7 87 57 127 4#87; W 119 77 167 4#119; W 120 78 170 4#120; X 121 79 171 4#121; Y 122 7A 172 4#122; Z 123 7B 173 4#123; { 124 7C 174 4#124;   125 7D 175 4#125; } 126 7E 176 4#126; ~					The state of the s	100000														
24 18 030 CAN (cancel) 56 38 070 8 8 88 58 130 X X 120 78 170 x X 25 19 031 EM (end of medium) 57 39 071 9 9 89 59 131 Y Y 121 79 171 y Y 26 1A 032 SUB (substitute) 58 3A 072 : 90 5A 132 Z Z 122 7A 172 z Z 27 1B 033 ESC (escape) 59 3B 073 ; 91 5B 133 [ [ 123 7B 173 { { 28 1C 034 FS (file separator) 60 3C 074 < < 92 5C 134 \ \ 124 7C 174     29 1D 035 GS (group separator) 61 3D 075 = = 93 5D 135 ] ] 125 7D 175 } } 30 1E 036 RS (record separator) 62 3E 076 > > 94 5E 136 ^ \ 126 7E 176 ~ \(\div \)					II. CALLEY CONTROL AND ADMINISTRATION OF THE AMERICAN CONTROL OF THE AMERICAN															
25 19 031 EM (end of medium) 57 39 071 9 9 89 59 131 Y Y 121 79 171 y Y 26 1A 032 SUB (substitute) 58 3A 072 : : 90 5A 132 Z Z 122 7A 172 z Z 27 1B 033 ESC (escape) 59 3B 073 ; ; 91 5B 133 [ [ 123 7B 173 { { 28 1C 034 FS (file separator) 60 3C 074 < < 92 5C 134 \ \ 124 7C 174     29 1D 035 GS (group separator) 61 3D 075 = = 93 5D 135 ] ] 125 7D 175 } } 30 1E 036 RS (record separator) 62 3E 076 > > 94 5E 136 ^ \ 126 7E 176 ~ \(\times\)																				
26 1A 032 SUB (substitute) 58 3A 072 :: 90 5A 132 Z Z 122 7A 172 z Z 27 1B 033 ESC (escape) 59 3B 073 ; 91 5B 133 [ [ 123 7B 173 { { 28 1C 034 FS (file separator) 60 3C 074 < < 92 5C 134 \ \ 124 7C 174     29 1D 035 GS (group separator) 61 3D 075 = = 93 5D 135 ] ] 125 7D 175 } } 30 1E 036 RS (record separator) 62 3E 076 > > 94 5E 136 ^ \ 126 7E 176 ~ \(\div \)	24	18	030	CAN	(cancel)	100			-					-						
27 1B 033 ESC (escape)	25	19	031	EM	(end of medium)														-	_
28 1C 034 FS (file separator) 60 3C 074 < < 92 5C 134 \ \ 124 7C 174     29 1D 035 GS (group separator) 61 3D 075 = = 93 5D 135 ] ] 125 7D 175 } } 30 1E 036 RS (record separator) 62 3E 076 > > 94 5E 136 ^ ^ 126 7E 176 ~ ~	26	1A	032	SUB	(substitute)	58	ЗA	072	<b>%#58;</b>	:	90	5A	132	Z	Z	122	7A	172	z	Z
29 1D 035 GS (group separator) 61 3D 075 = = 93 5D 135 ] ] 125 7D 175 } } 30 1E 036 RS (record separator) 62 3E 076 > > 94 5E 136 ^ ^ 126 7E 176 ~ ~	27	1B	033	ESC	(escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
30 1E 036 RS (record separator) 62 3E 076 6#62; > 94 5E 136 6#94; ^ 126 7E 176 6#126; ~	28	10	034	FS	(file separator)	60	30	074	<	<	92	5C	134	\	1					
	29	1D	035	GS	(group separator)	61	ЗD	075	=	=	93	5D	135	]	]					
31 1F 037 US (unit separator)   63 3F 077 6#63; ?   95 5F 137 6#95; _   127 7F 177 6#127; DEL	30	1E	036	RS	(record separator)	62	3E	076	>	>	94	5E	136	<b>4</b> ;	^					
	31	1F	037	US	(unit separator)	63	3F	077	?	2	95	5F	137	_	_	127	7F	177		DEL

#### Zeichenklassen

Bezeichnung	Bedeutung	Beispiele
Positive Zeichenklasse	Definition einer Klasse von Zeichen	[a-z]: Alle Kleinbuchstaben [!-~]: Alle ASCII Zeichen
Negative Zeichenklasse	Alles außer dem Spezifizierten Zeichen	[^a-z]: Alles außer Kleinbuchstaben

Innerhalb der []-Klammern müssen nur [, ], -, \ als Metazeichen escaped werden.

## Spezielle Zeichenklassen

Zeichen	Bedeutung	Alternative
\d	Eine Zahl	[0-9]
\s	Einen "Weißraum"	[ \f\n\r\t\v] und alle Unicodevarianten zB \u00a0
\w	Ein alphanumerisches Zeichen inklusive _	[A-Za-z0-9_]
\D	Alles außer Zahlen	[^0-9], [^\d]
\S	Alles außer "Weißraum"	[^\s]
\W	Alles außer alphanumerische Zeichen	[^A-Za-z0-9_]
\b	Eine Wortgrenze	$(^ w \ws \w \w \w)$
	Alle Zeichen außer Zeilenumbrüche	[^\n]

#### Anker

Zeichen	Bedeutung
^	Anfang der Zeichenkette (kann durch Modifier verändert werden)
\$	Ende der Zeichenkette (kann durch Modifier verändert werden)
\A	Anfang der Zeichenkette
\z	Ende der Zeichenkette
١Z	Ende der Zeichenkette plus ein oder kein Zeilenumbruch

## Quantifier

Regex	Bedeutung	Beispiele
{x}	Zeichenkette der Länge x	[a-z]{2}
{x,y}	Zeichenkette der Länge x bis y	[a-z]{2,4}
{x,}	Zeichenkette von mindestens Länge x	[a-z]{2,}
+	{1,}	[a-z]+
*	{0,}	[a-z]*
?	{0,1}	[a-z]?

- 1. Quantifier sind generell "greedy". Durch das Anhängen eines ? werden sie "lazy".
- 2. Der possessiv (besitznehmende) Quantifier + erlaubt "strenges" Gruppieren

#### Lookarounds

Zeichen	Bedeutung	Beispiele
(?=)	Positive Lookahead	ab(?=c)
(?!)	Negative Lookahead	ab(?!z)
(?<=)	Positive Lookbehind	(?<=a)bc
(? )</td <td>Negative Lookbehind</td> <td>(?<!--x)bc</td--></td>	Negative Lookbehind	(? x)bc</td

Die meisten Implementierungen erlauben keine Quantifier in Lookbehind Ausdrücken.

# Capturing Groups

Bezeichnung	Bedeutung	Beispiel
Capturing Group	Klammern werden verwendet für Capturing Groups, mittels \1 können die Gruppen referenziert werden (Backreference)	(abc) (def) \1 \2
Non-Capturing Group	?: wird verwendet damit der Match nicht referenziert werden kann	(?:abc)
Named Capturing Group	Anstatt einer numerierten Backreference kann auch eine Bezeichnung verwendet werden und mit \k'name' referenziert werden	(?'name'abc) \k'name'

#### Modifier

Zeichen	Bedeutung
i	Case Insensitive
S	. Zeichen inkludiert auch Zeilenumbrüche
m	Multiline
X	Free Spacing mode, # sind Kommentare
g	Globaler Match

#### Unicode

- Unicode kennt einen Unterschied zwischen Zeichen und Zeichencode, deshalb würde ein ü aus zwei Unicode Zeichen bestehen können, der .-Operation kann deshalb zu Problemen führen
- X für alle Zeichen inklusive Zeilenumbrüche in Unicode
- Unicode ist in 156 Blöcke aufgeteilt. zB Kyrillisch (U+0400...U+04FF)
- Mit \p{Cyrillic} kann ein Zeichen des kyrillischen Alphabet referenziert werden

#### Weitere Funktionen

IF/ELSE Konstrukt	Falls die referenzierte Capture Group exisitiert wird der if-Teil gematcht ansonsten der else-Teil	(regex)?(?(1)foo  bar)
Rekursion	Rekursion (~Kopie) des gesamten Musters	L(?R)R
Atomic Grouping	Hauptsächlich aus Performanzgründen. Backtracking wird verhindert	a(?>bc b)c

### Check beim Erstellen von Regular Expression

- A (Anker): Können Anker oder Wortgrenzen gesetzt werden?
- G (Greedy): Greedy vs. Lazy Quantifier
- R (Repeat): Gibt es Teile des Musters die sich wiederholen und präziser beschrieben werden können?
- A (Atomic): Atomare oder possesive Quantifier verwenden?

### Erstellung von Regular Expressions

- Verwenden eines Emulators zum testen
  - https://regex101.com/
- So spezifisch wie möglich vs. so Allgemein wie möglich
- Wartbarkeit/Lesbarkeit vs Präzision
- Quellcodeformatierung (ja/nein)
- Nach Ideen/Vorschlägen bei Google suchen

### Wichtige Fragen

- Wer ist für die Wartung der Regular Expressions zuständig?
- Wann kann eine Regular Expression aufgeteilt werden?
- Wann sind die Grenzen von Regular Expressions erreicht?
- Gibt es Alternativen zu Regular Expressions?

## Quellen / Literaturtipps

- Link: <a href="http://www.regular-expressions.info">http://www.regular-expressions.info</a>
- Link: <a href="http://www.rexegg.com">http://www.rexegg.com</a>
- Buch (Einführung): Introducing Regular Expressions – Michael Fitzgerald
- Buch (Fortgeschrittene): Reguläre Ausdrücke Jeffrey E. F. Friedel



