GOLD STANDARD REGEXES

Listed here are the gold standard regexes for the six entity extraction tasks specified in Table 2. The procedure for generating the gold standard regexes is outlined in the Experiments section. Notice that some of the regexes contain positive and negative lookbehinds (`(?<=)` and `(?<!)`, respectively).

Table 1: DATE$_{\text{MIDEAST}}$

`\d{2}/\d{2}/\d{2}`
`\d/\d{2}/\d{2}`
`\d/\d/\d{2}`
`\d{2}-\d{2}-\d{2}`
`\d{2}\.\d{1}\.\d{2}`

Table 2: DATE$_{\text{WEBKB}}$

`\d{2}/\d{2}/\d{2}`
`\d/\d{2}/\d{2}`
`\d{2}/\d/\d{2}`
`\d/\d/\d{2}`
`\d{2}-\d{2}-\d{2}`
`1-Oct-96`
`25-May`
`4-November`
`06nov96`
`26aug96`
`9-September-96`
`08-Oct-96`
`12nov96`

Table 3: DATE$_{\text{ENRON}}$

`\d{2}/\d{2}/\d{2}`
`\d/\d{2}/\d{2}`
`\d{2}/\d/\d{2}`
`\d/\d/\d{2}`
`\d{2}\.\d{2}\.0\d`
`\d{2}\.03\.\d{2}`
`01-\d1-\d{2}`
`\d{2}-\d{2}-99`
`\d{2}-\d{2}-0\d`
`00-\d{2}-\d{2}`
`\d-\d-0\d`
`\d-\d-9\d`
`\d\.\d\.0\d`
`\d-\d{2}-\d{2}`
`\d\.\d{2}\.0\d`
`\d\.\d{2}\.9\d`
`\d{2}-\d-0\d`
`\d{2}-\d-9\d`
`\d{2}\.\d\.0\d`
`\d{2}\.1\.9\d`

Table 4: PHONE$_{\text{FORSALE}}$

\(\d{3}\)␣\d{3}-\d{4}
\d{3}-\d{3}-\d{4}
\d{3}/\d{3}-\d{4}
\d{3}␣\d{3}-\d{4}
\(\d{3}\)\d{3}-\d{4}
\(\d{3}\)-\d{3}-\d{4}
\(\d{3}\)/\d{3}-\d{4}
\(\d{3}\)␣\d{3}␣\d{4}
\d{3}\.\d{3}\.\d{4}
\d{3}␣\d{3}␣\d{4}
\d{3}-\d{3}␣\d{4}
\(\d{3}\)\d{7}
(?<!\d\)␣)(?<!\d␣)(?<=␣)\d{3}-\d{4}

Table 5: COURSE$_{\text{WEBKB}}$

CS\d{3}
cs\d{3}
EE\d{3}
ee\d{3}
CSE\d{3}
CS\d{3}[A-Z]
CS\d{4}
EE\d{3}[a-zA-Z]
CSE\d{3}[a-zA-Z]
PS\d{3}
MIS\d{3}
MA\d{3}
ma\d{3}
CS␣\d{2}
CS\d{2}
CSE\d{3}[A-Z]{2}

Table 6: PHONE$_{\text{ENRON}}$

\d{3}-\d{3}-\d{4}
\d{3}-␣\d{3}-\d{4}
\d{3}/\d{3}-\d{4}
\d{3}␣\d{3}-\d{4}
\d{3}␣␣\d{3}-\d{4}
\d{3}\.\d{3}\.\d{4}
\d{3}␣\d{3}␣\d{4}
\d{3}-\d{3}␣\d{4}
\d{4}-\d{3}-\d{4}
\d{3}\.\d{3}-\d{4}
\d{3}-\d{3}-␣\d{4}
\d{3}-\d{3}-[A-Z]{4}
\d{3}-[A-Z]{3}-\d{4}
\d{3}-\d{3}-\w{4}
\d{3}/\d{3}/\d{4}
\d{3}/\d{3}\.\d{4}
\d{2}-\d{4}-\d{4}
(?<!\d\)␣)(?<!\d␣␣)(?<!\d-␣)(?<!\d␣)(?<=␣)\d{3}-\d{4}
1-\d{3}-[A-Z]{6}\d
1-\d{3}-[A-Z]{7}
1-\d{3}-[A-Z]{3}-[A-Z]{4}
1-\d{3}-[A-Z]\d-[A-Z]{5}
1-\d{3}-[A-Z]\d-[A-Z][a-z]{4}