Pattern Recognition

Table of contents:

1. Pattern recognition using regular expression
2. Pattern recognition using open source libraries
3. Pattern recognition using third party tools

# Pattern Recognition using regular expression

The following regular expression is provided by Stanford NLP Tokenizer, the original files provides various pattern recognition regular expression, here we only presented some regular expression that would be most useful for us. The detail can be found in this [link](https://github.com/stanfordnlp/CoreNLP/blob/master/src/edu/stanford/nlp/process/PTBLexer.flex).

EMAIL = (&lt;|<)?[a-zA-Z0-9][^ \t\n\f\r\"<>|()\u00A0{}]\*@([^ \t\n\f\r\"<>|(){}.\u00A0]+\.)\*([^ \t\n\f\r\"<>|(){}.\u00A0]+)(&gt;|>)?

PHONE = (\([0-9]{2,3}\)[ \u00A0]?|(\+\+?)?([0-9]{2,4}[\- \u00A0])?[0-9]{2,4}[\- \u00A0])[0-9]{3,4}[\- \u00A0]?[0-9]{3,5}|((\+\+?)?[0-9]{2,4}\.)?[0-9]{2,4}\.[0-9]{3,4}\.[0-9]{3,5}

DATE = \d{1,2}[\-\/]\d{1,2}[\-\/]\d{2,4}

FULLURL = https?:\/\/[^ \t\n\f\r\"<>|(){}]+[^ \t\n\f\r\"<>|.!?(){},-]

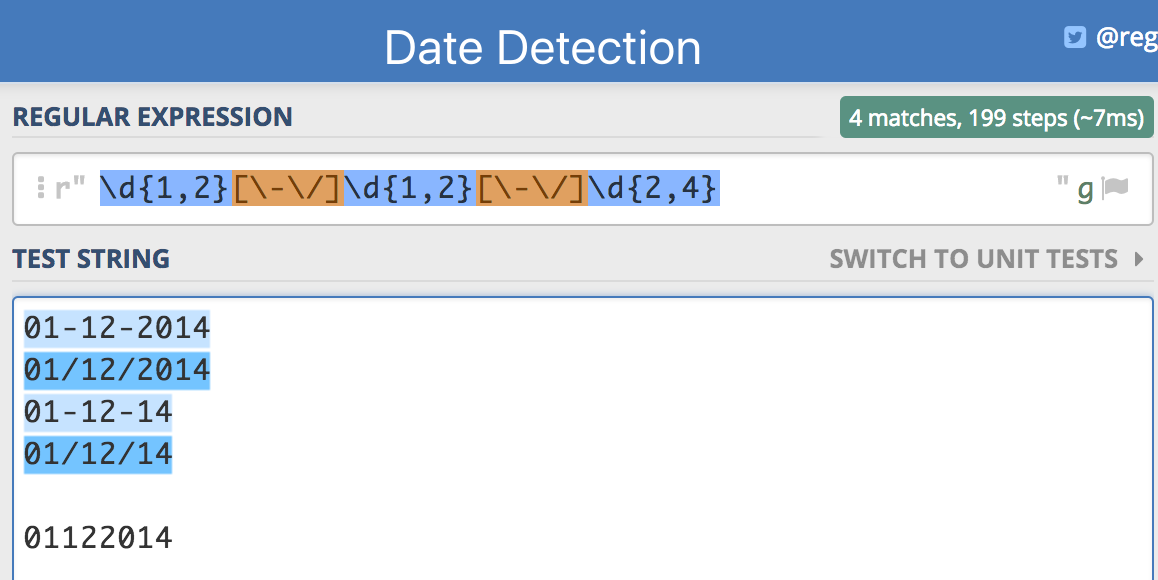
LIKELYURL = ((www\.([^ \t\n\f\r\"<>|.!?(){},]+\.)+[a-zA-Z]{2,4})|(([^ \t\n\f\r\"`'<>|.!?(){},-\_$]+\.)+(com|net|org|edu)))(\/[^ \t\n\f\r\"<>|()]+[^ \t\n\f\r\"<>|.!?(){},-])?

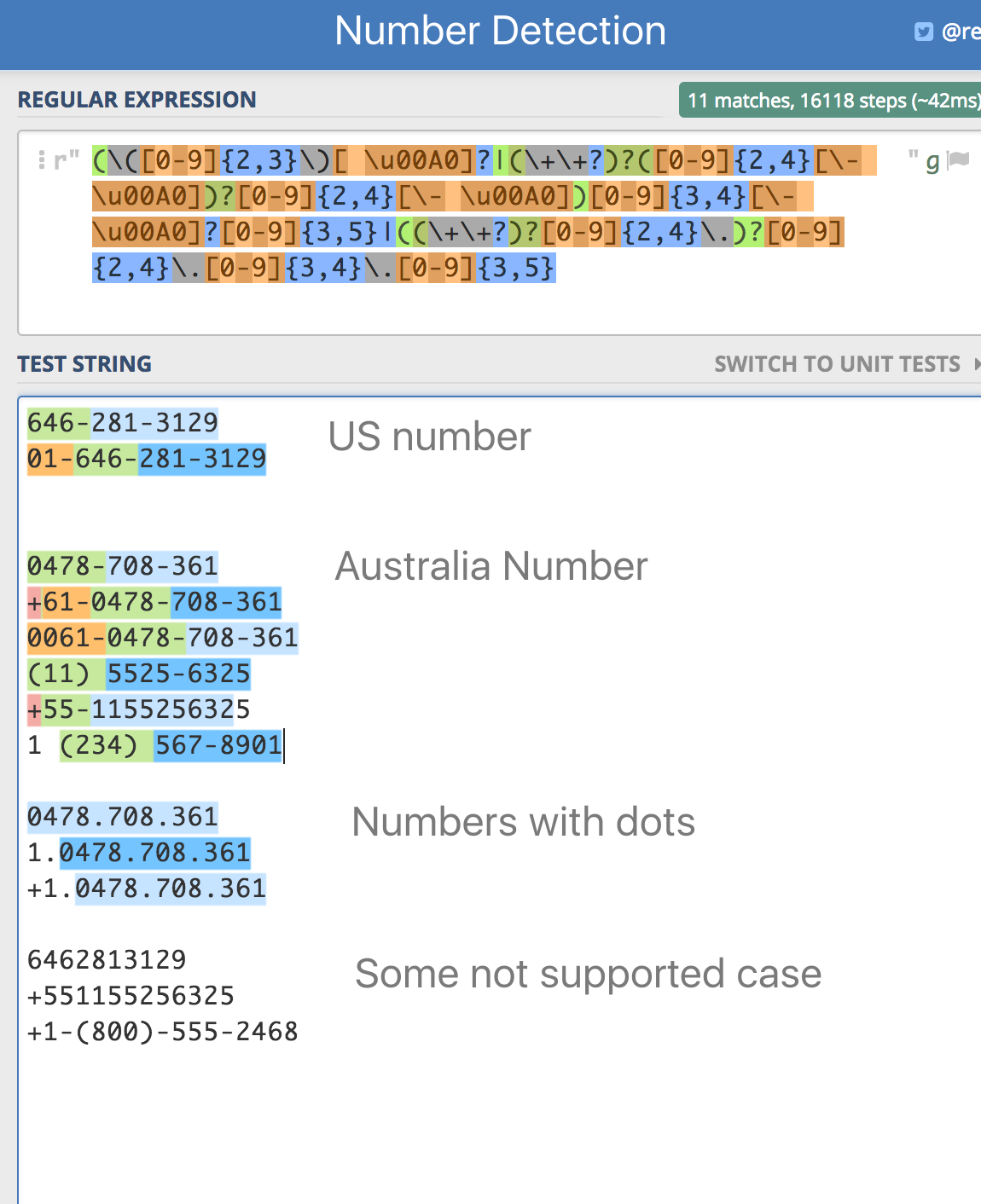
ABMONTH = Jan|Feb|Mar|Apr|Jun|Jul|Aug|Sep|Sept|Oct|Nov|Dec

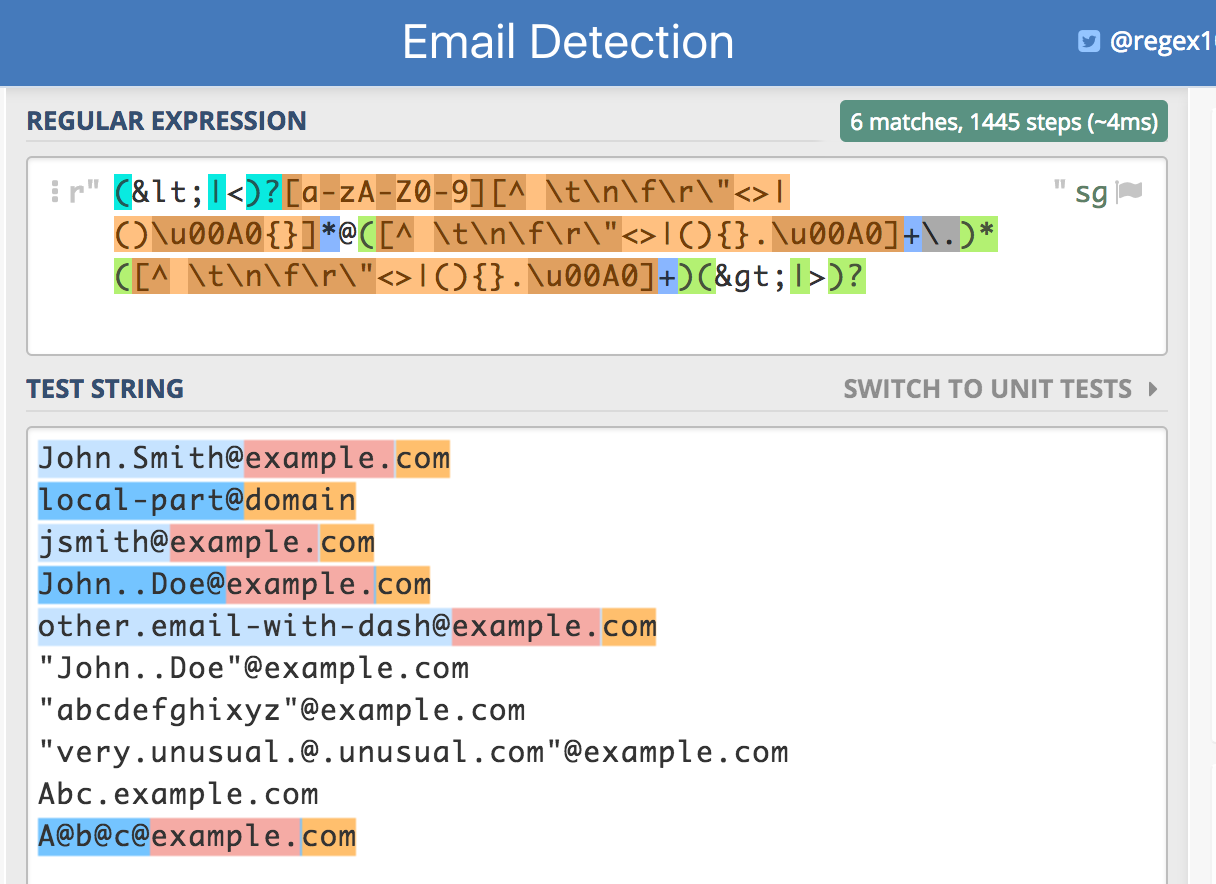
ABDAYS = Mon|Tue|Tues|Wed|Thu|Thurs|Fri

ABSTATE = Ala|Ariz|[A]z|[A]rk|Calif|Colo|Conn|Ct|Dak|[D]el|Fla|Ga|[I]ll|Ind|Kans?|Ky|[L]a|[M]ass|Md|Mich|Minn|[M]iss|Mo|Mont|Neb|Nev|Okla|[O]re|[P]a|Penn|Tenn|[T]ex|Va|Vt|[W]ash|Wisc?|Wyo

Some Demo Result of using regular expression:







# Pattern Recognition using open source libraries

Google's provides Java, C++ and JavaScript library for parsing, formatting, and validating international phone numbers. <https://github.com/googlei18n/libphonenumber>

This library can validate phone number, extract country code and national number, reformat the number, etc.

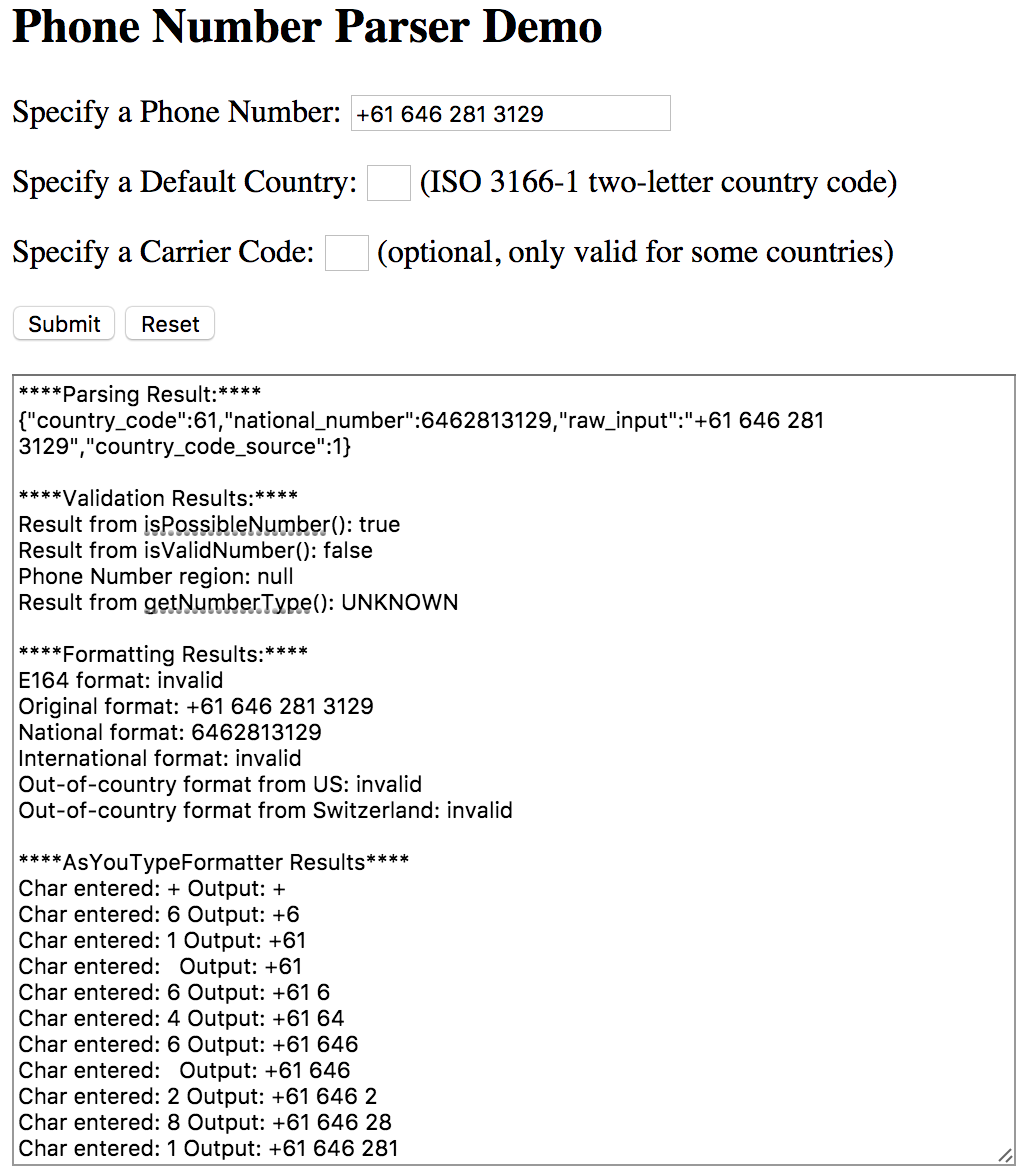
A quick example can be found here:

<https://github.com/googlei18n/libphonenumber#quick-examples>

A JavaScript Demo can be found here:

<https://rawgit.com/googlei18n/libphonenumber/master/javascript/i18n/phonenumbers/demo-compiled.html>

You can try it with your own number. For example, the demo output looks like this:



# Pattern Recognition using third party software

Generally, open source libraries already provide satisfying functionality for daily uses. But if more advanced functionality is required, we’ve found some third party software to provide more sophisticated functionalities.