Nick Radcliffe
Stochastic Solutions Limited

pip install tdda
http://rexpy.herokuapp.com

Data analysis (lisp-like) interface

This is Miró version 2.14.09, running under Salvador 1.2.59.


```
Seed: 1565957785
Logs started at 2018/06/08 08:13:06 host godel.local.
Logging to /Users/njr/miro/log/2018/06/08/session002.

[1]>  . ~/edipydata/rex.miros

[2]> load rex

rex.miro: 10 records; 10 (100%) selected; 7 fields.

[3]> show
```

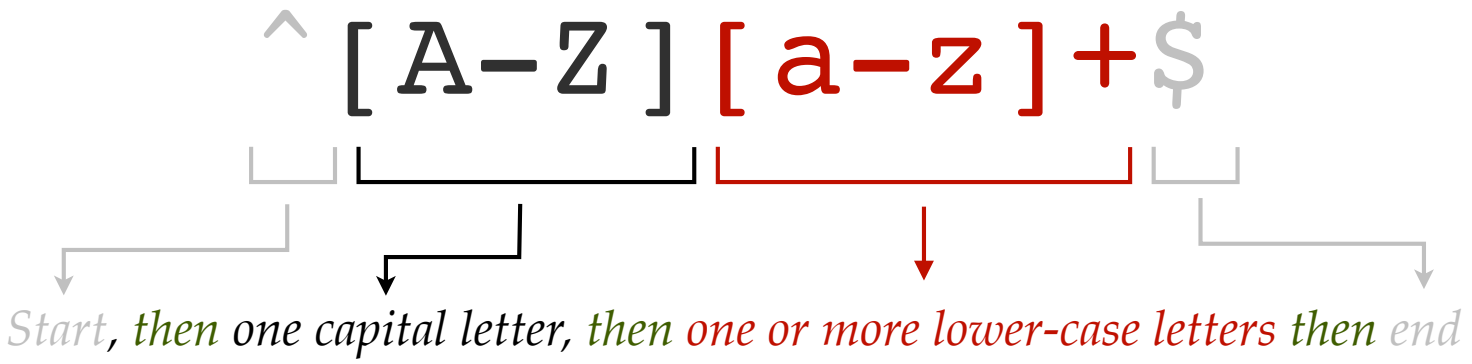
postcode	name	numberplate	greek	altname	fave_colour	day_born
M19 9PV	Jerry John-Joe	10 E	μν	Åmy	Red	Monday
AB2 2QT	Billy Beetle	3180 KQ	φχψω	Björn	Orange	Tuesday
W1 4QQ	Claire Cockroach	ABC 123W	αβγ	cl4ire	Yellow	Wednesday
B22 8AB	David Dragfly	SO65 UVB	φχψω	daviδ	Green	Thursday
RG12 1TT	Ellen Vannin	SO66 UVA	αβγ	Ell-n	Blue	Friday
N1 4YT	Freddie Freeloader	3128 NW	μν	Fredd-ee	Indigo	Saturday
G1 3ER	Geraldine Gnat	PPP 927	αβγ	Géráldïnè	Violet	Sunday
AC2 1LR	Harry Helicopter	UVA SO65	μν	happφ	Red-Orange	Monday
E4 5PH	Isla St-Clair	RAD 10	αβγ	Isla	Cr1ms0n	Tuesday
M19 9PV	Jorge John-Joe	10 E	μν	jøøøøørg	Blackest_Black	Wednesday

```
[3]> show
```

postcode	name	numberplate	greek	altname	fave_colour	day_born
M19 9PV	Jerry John-Joe	10 E	μν	Åmy	Red	Monday
AB2 2QT	Billy Beetle	3180 KQ	φχψω	Björn	Orange	Tuesday
W1 4QQ	Claire Cockroach	ABC 123W	αβγ	cl4ire	Yellow	Wednesday
B22 8AB	David Dragfly	SO65 UVB	φχψω	daviδ	Green	Thursday
RG12 1TT	Ellen Vannin	SO66 UVA	αβγ	Ell-n	Blue	Friday
N1 4YT	Freddie Freeloader	3128 NW	μν	Fredd-ee	Indigo	Saturday
G1 3ER	Geraldine Gnat	PPP 927	αβγ	Géráldïnè	Violet	Sunday
AC2 1LR	Harry Helicopter	UVA SO65	μν	happφ	Red-Orange	Monday
E4 5PH	Isla St-Clair	RAD 10	αβγ	Isla	Cr1ms0n	Tuesday
M19 9PV	Jorge John-Joe	10 E	μν	jøøøøørg	Blackest_Black	Wednesday

```
[4]> rex day_born
```

```
^[A-Z][a-z]+$
```



- Rexpy (rex command in Miró) generates one or more regular expressions that, between them, match all the example strings

```
[3]> show
```

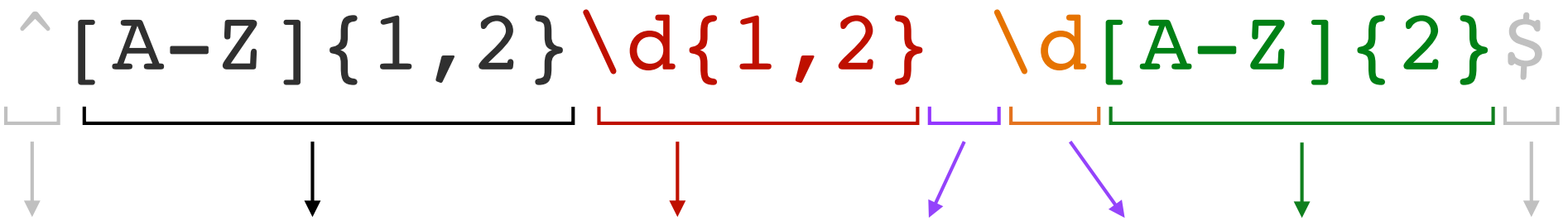
postcode	name	numberplate	greek	altname	fave_colour	day_born
M19 9PV	Jerry John-Joe	10 E	μν	Åmy	Red	Monday
AB2 2QT	Billy Beetle	3180 KQ	φχψω	Björn	Orange	Tuesday
W1 4QQ	Claire Cockroach	ABC 123W	αβγ	cl4ire	Yellow	Wednesday
B22 8AB	David Dragfly	SO65 UVB	φχψω	daviδ	Green	Thursday
RG12 1TT	Ellen Vannin	SO66 UVA	αβγ	Ell-n	Blue	Friday
N1 4YT	Freddie Freeloader	3128 NW	μν	Fredd-ee	Indigo	Saturday
G1 3ER	Geraldine Gnat	PPP 927	αβγ	Géräldïnè	Violet	Sunday
AC2 1LR	Harry Helicopter	UVA SO65	μν	haρρφ	Red-Orange	Monday
E4 5PH	Isla St-Clair	RAD 10	αβγ	Isla	Cr1ms0n	Tuesday
M19 9PV	Jorge John-Joe	10 E	μν	jøøøøørg	Blackest_Black	Wednesday

```
[4]> rex day_born
```

```
^[A-Z][a-z]+$
```

```
[5]> rex postcode
```

```
^[A-Z]{1,2}\d{1,2} \d[A-Z]{2}$
```



Start, then 1 or 2 capitals, then 1 or 2 digits then 1 space then 1 digit then two capitals end

Regular expressions can be used to search for/select matching strings in the Unix/Linux shell, with grep, sed, awk etc., in Python (using the `re` *module), and in Miró. As expected, all the postcodes shown match the regular expression Rexpy generated.*

```
[6]> select postcode =~ "^[A-Z]{1,2}\d{1,2} \d[A-Z]{2}$"

rex.miro: 10 records; 10 (100%) selected; 7 fields.
Selection:
(=~ postcode "^[A-Z]{1,2}\d{1,2} \d[A-Z]{2}$")

[7]> show postcode
```

postcode
M19 9PV
AB2 2QT
W1 4QQ
B22 8AB
RG12 1TT
N1 4YT
G1 3ER
AC2 1LR
E4 5PH
M19 9PV

```
[8]> select all

rex.miro: 10 records; 10 (100%) selected; 7 fields.
```

But what about alphabetical characters with accents or from other alphabets?
What would Rexpy generate for the field `greek` *or* `altname` *below?*

```
[11]> show
```

postcode	name	numberplate	greek	altname	fave_colour	day_born
M19 9PV	Jerry John-Joe	10 E	μν	Åmy	Red	Monday
AB2 2QT	Billy Beetle	3180 KQ	φχψω	Björn	Orange	Tuesday
W1 4QQ	Claire Cockroach	ABC 123W	αβγ	cl4ire	Yellow	Wednesday
B22 8AB	David Dragfly	SO65 UVB	φχψω	daviδ	Green	Thursday
RG12 1TT	Ellen Vannin	SO66 UVA	αβγ	Ell-n	Blue	Friday
N1 4YT	Freddie Freeloader	3128 NW	μν	Fredd-ee	Indigo	Saturday
G1 3ER	Geraldine Gnat	PPP 927	αβγ	Géráldïnè	Violet	Sunday
AC2 1LR	Harry Helicopter	UVA SO65	μν	happφ	Red-Orange	Monday
E4 5PH	Isla St-Clair	RAD 10	αβγ	Isla	Cr1ms0n	Tuesday
M19 9PV	Jorge John-Joe	10 E	μν	jøøøøørg	Blackest_Black	Wednesday

```
[12]> rex greek
```

```
^[^\W0-9_]{2,4}$
```

2–4 what???

Let's start with `\w` *(not* `\W`*):*
`\w` *matches any character that's OK in an identifier*
— letters, digits and underscores.

```
[13]> select fave_colour =~  "^\w+$"

rex.miro: 10 records; 9 (90%) selected; 7 fields.
Selection:
(=~ fave_colour "^\w+$")

[14]> show
```

postcode	name	numberplate	greek	altname	fave_colour	day_born
M19 9PV	Jerry John-Joe	10 E	μν	Åmy	Red	Monday
AB2 2QT	Billy Beetle	3180 KQ	φχψω	Björn	Orange	Tuesday
W1 4QQ	Claire Cockroach	ABC 123W	αβγ	cl4ire	Yellow	Wednesday
B22 8AB	David Dragfly	SO65 UVB	φχψω	daviδ	Green	Thursday
RG12 1TT	Ellen Vannin	SO66 UVA	αβγ	Ell-n	Blue	Friday
N1 4YT	Freddie Freeloader	3128 NW	μν	Fredd-ee	Indigo	Saturday
G1 3ER	Geraldine Gnat	PPP 927	αβγ	Géráldïnè	Violet	Sunday
E4 5PH	Isla St-Clair	RAD 10	αβγ	Isla	Cr1ms0n	Tuesday
M19 9PV	Jorge John-Joe	10 E	μν	jøøøøørg	Blackest_Black	Wednesday

```
[15]> flipselection

rex.miro: 10 records; 1 (10%) selected; 7 fields.
Selection:
(not (=~ fave_colour "^\w+$"))

[16]> show
```

postcode	name	numberplate	greek	altname	fave_colour	day_born
AC2 1LR	Harry Helicopter	UVA SO65	μν	happφ	Red-Orange	Monday

Only Red-Orange doesn't match (because of the hyphen)

But what's special about `\w` *is that the alphabetical characters it matches are alphabetical characters in* ***all languages, with or without accents****.*

So let's go back to the actual regular expression for **μν**, **φχψω** *and* **αβγ**

```
^[^\W0-9_]{2,4}$
```

The `^` *inside* `[]` *means "anything except"*

`\W` *is the inverse of* `\w`*, i.e. it's anything that isn't OK in an identifier*
i.e. anything except (multilingual) letters, digits and underscores

So this regular expression reads:

start, then any 2-4 characters that ***are*** *OK in identifiers,*
and are not digits, and are not underscores, then end

And if I hadn't been limited to 5 minutes(!), I'd have finished by showing the matching on accented characters too:

```
[21]>  select altname =~ /^[^\W0-9_]+$/

rex.miro: 10 records; 7 (70%) selected; 7 fields.
Selection:
(=~ altname "^[^\W0-9_]+$")

[22]>  show altname
```

altname
Åmy
Björn
davið
Géráldïnè
happφ
Isla
jøøøøørg

These all match

```
[23]>  flipselection

rex.miro: 10 records; 3 (30%) selected; 7 fields.
Selection:
(not (=~ altname "^[^\W0-9_]+$"))

[24]>  show altname
```

altname
cl4ire
Ell-n
Fredd-ee

cl4aire *doesn't because we've specifically excluded digits and* Ell-n *and* Fredd-ee *don't because of the hyphens*

And that's the power of `\W`

Rexpy is available in Miró (our own commercial data analysis suite)

and also in our open source test-driven data analysis library tdda

```
pip install tdda
```

```
git clone https://github.com/tdda/tdda.git
python setup.py install
```

and also online at

http://rexpy.herokuapp.com

The online version ~~might be~~ is out of date and ~~might~~ does not yet support the `\W` *goodness described here. But it will by the end of 2018-06-08!*

Nick Radcliffe
Stochastic Solutions Limited

pip install tdda
http://rexpy.herokuapp.com