The Basics of Regular Expressions in Tableau

Inger Bergom & Christina Butler

Tufts University Office of Institutional Research

2019 Northeast Association for Institutional Research Annual Conference

Why use regular expressions in Tableau?

- One of the five **Duties and Functions of Institutional Research** stated by AIR is collecting, analyzing, interpreting and reporting data and information (airweb.org, 2019)

- In order to achieve this in a rapidly-changing data environment, IR professionals need to continually **adapt to new technologies and data querying tools**

- Tableau is **increasingly used** in IR for data management, analysis, and visualization

What are regular expressions?

- **Not new or unique to Tableau**; originated in the 1950s and standard in many computer languages and tools, such as Python and MySQL
  - Released in Tableau 9.0

- Help execute **complex queries of text data**—such as addresses or birthdates

- **Extract patterns of text**, similar to wildcard notations, but can be more complex and offer additional functionality

What do regular expressions look like?

Find five consecutive digits:

[0-9]{5}

Look for digits

5 in a row

What do regular expressions look like?

Find two consecutive capital letters:

[A-Z]{2}

Look for capital letters only

2 in a row

Regular Expression

```
/[A-Z]{2}/g
```

Test String

```
63 Pendergast Lane
Plainfield, NJ 07060
194 Division Dr.
Cranston, RI 02920
28 Westminster Street
Hanover, PA 17331
7351 North La Sierra Ave.
Staunton, VA 24401
97 Bald Hill Street
Shirley, NY 11967
458 Plumb Branch St.
West Deptford, NJ 08096
```

Regular expressions in Tableau

REGEXP_REPLACE(string, pattern, replacement): Returns a copy of the given string where the regular expression pattern is replaced by the replacement string.

REGEXP_MATCH(string, pattern): Returns true if a substring matches the regex pattern.

REGEXP_EXTRACT(string, pattern): Returns the portion of the string matching the regular expression pattern.

REGEXP_EXTRACT_NTH(string, pattern, index): Returns the portion of the string that matches the regular expression pattern. The substring is matched to the nth capturing group, where n is the given index.

Source: https://www.tableau.com/about/blog/2015/6/become-regular-regular-expressions-39802

Use Case #1:

Reporting Student Addresses

Reporting student addresses

- Bi-annual reporting of student addresses to comply with city ordinances
- Student-entered data (messy!)
- Request for specific format:

Street Number	Street Name	Suffix	Unit	Zip Code
61	ADAMS	ST	2	02144
27R	ALBION	ST	1	02144
27R	ALBION	ST	1	02144
27R	ALBION	ST	2	02144
9	ALDERSEY	ST	3	02144
21	ALDERSEY	ST	2	02144
297	ALEWIFE	PKWY	2	02144
26	APPLETON	ST	2	02144

- Old way in Excel = slow, tedious, reliant on manual checks
- New way in Tableau = more automated, faster, built-in checks

Reporting student addresses

Zip code: Return 5 consecutive digits

Address	Zip
1 Amherst Way, Princeton Junction, NJ 08550	08550
1 Austin Road West, Cullinan Ocean Sky 22C, Kowloon HKG	Null
1 Bellevue St, Saugus, MA 01906-2223	01906
1 Bond Str., Apt. 6D, Ny, NY 10012	10012

Returns Null if no match

Data we have to work with

Field we want to create

Find digits

5 in a row (the "quantifier")

Calculation:

```
REGEXP_EXTRACT([Address],"([0-9]{5})")
```

Tableau function

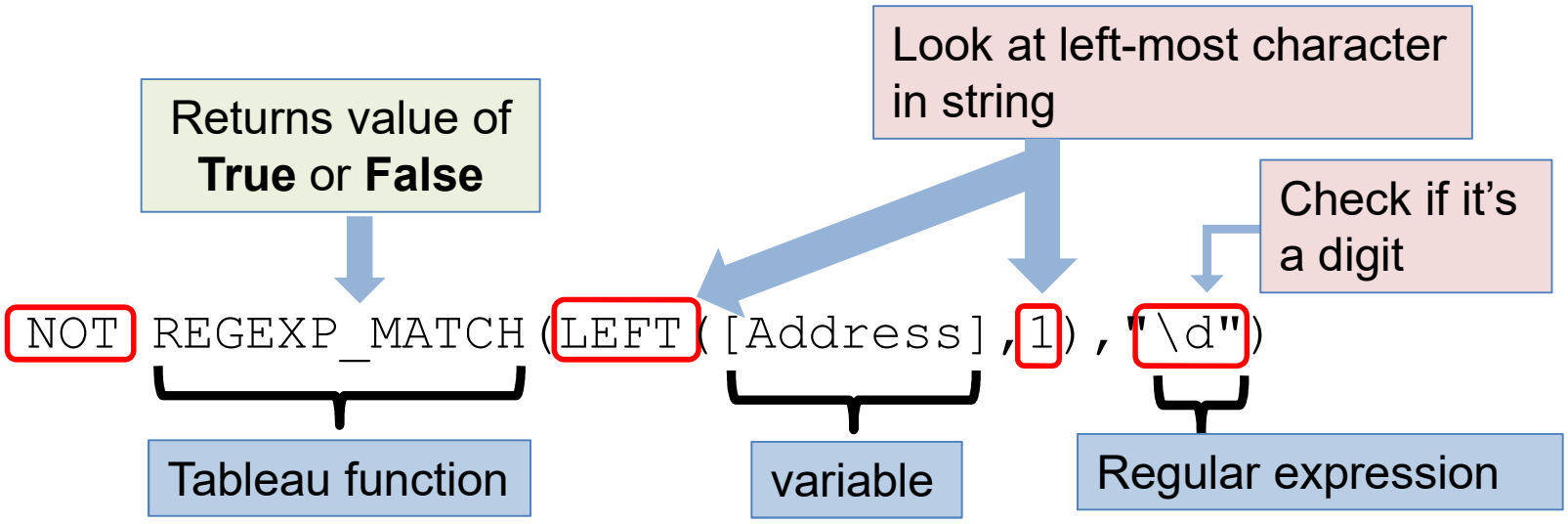
variable

Regular expression
(needs to be in quotes for Tableau)

Reporting student addresses

Identify addresses not starting with number

Address	Starts with non-number
Bush Hall 316	True
PO Box 64, Quincy, MA 02170	True
44241 Topaz Way, Fremont, CA 94539	False
The Fletcher School, 160 Packard Ave Box 76, Medford, MA 02155-5815	True
39917 Chalon Ct, Temecula, CA 92591	False



Reporting student addresses

Street suffix: Identify any of a specific set of strings (returns first match)

Address	Street suffix
44 Gordon Street, Somerville, MA 02144	Street
94 Bartlett St., Apt. 1, Somerville, MA 02145	St.
112 Thurston St Apt 6, Somerville, MA 02145	St
129 Powder House Blvd Apt 1, Somerville, MA 02144-1613	Blvd
197 Mount Vernon Street, Malden, MA 02148	Street
201 E 87th St Apt 19F, Cameron King, New York, NY 10128-3217	St
No. 7, 9th Alley, Ayatollah Khamanei St., Ardabil 5613635786 IRN	St.

Data we have to work with

Extract any of these strings
(abbreviated list for demo)

Field we want to create

Calculation:

```
REGEXP_EXTRACT ([Address],'(Street|St |St.|Blvd)')
```

Tableau function

variable

Regular expression

Regex vs. String Calculation

This **regular expression…**

```
REGEXP_EXTRACT (Address],'(Street|St |St.|Blvd)')
```

returns the same result as this **string calculation…**

```
IF CONTAINS([Address],"Street") THEN "Street"
ELSEIF CONTAINS([Address],"St ") THEN "St "
ELSEIF CONTAINS([Address],"St.") THEN "St."
ELSEIF CONTAINS([Address],"Blvd") THEN "Blvd"
END
```

(but the regular expression is more concise).

Use Case #2:

Removing special characters for naming PDFs

Removing special characters for naming PDFs

- Every year, over 200 advising reports are generated from Tufts' Senior Survey
- This process was moved to Tableau and using Tableau's tabcmd command line utility, PDF reports can be created from a single dashboard
- Each PDF is given a custom file name based on **Major** and **Faculty Name**
- In order for this process to run successfully, certain characters must be removed from the **Major** and **Faculty Name** strings

Removing special characters for naming PDFs

Remove non-alpha characters for a field used to name PDFs (PDFs cannot accept special characters)

Find any non-alpha characters

```
REGEX_REPLACE([Major], '[^a-zA-Z]',"")
```

Caret ("^") means "is not" when inside the brackets

Replace with nothing

Major	Major_for_PDF
Child Study and Human Development	ChildStudyandHumanDevelopment
Computer Science (EN)	ComputerScienceEN
Economics (Quantitative)	EconomicsQuantitative
Engineering - Biomedical	EngineeringBiomedical
Music, Sound, and Culture	MusicSoundandCulture
Women's, Gender, and Sexuality Studies	WomensGenderandSexualityStudies

Use Case #3:

Cleaning admissions data

Cleaning admissions data

- OIR received a request to report on admissions data by **Term**
- However, it was discovered that the **Term** field in the data source was unreliable and did not always align with when an application was actually being reported
- The solution was to derive a more accurate **Term** field by considering the year of the corresponding **Admissions Round**

Cleaning admissions data

Fixing data anomalies:
Override term field based on admissions round string field

Find 4 consecutive digits in Term field

Replace with 4 consecutive digits found in Round field

```
REGEXP_REPLACE([Term],'(\d{4})',REGEXP_EXTRACT([Round],'(\d{4})'))
```

Term	Round	Term (modified)
Fall 2015	2015 Doctor of Philosophy (PhD)	Fall 2015
	Fall 2015	Fall 2015
Fall 2016	2016 DVM First Year Student	Fall 2016
	2019 Sackler Non Degree Coursework, Programs, and Visiting..	Fall 2019
Fall 2017	Certificate - Fall 2017	Fall 2017
Fall 2018	2018 Exchange	Fall 2018
	2019 Friedman School of Nutrition Science and Policy	Fall 2019
	2019 Sackler Non Degree Coursework, Programs, and Visiting..	Fall 2019
	Fall 2020	Fall 2020
Fall 2021	2020 Sackler PhD & MS Programs	Fall 2020

Resources

Regex testers:
https://www.regextester.com/
https://www.regexpal.com/
https://regex101.com/

Regular expressions information, library of expressions:
https://github.com/ziishaned/learn-regex
https://www.regular-expressions.info/

Regular Expressions Quick Start Guide:
https://www.regular-expressions.info/quickstart.html

Regular Expressions and Tableau
https://www.tableau.com/about/blog/2015/6/become-regular-regular-expressions-39802
https://code.tutsplus.com/tutorials/8-regular-expressions-you-should-know--net-6149

Tips for improving performance

Materialize Calculations in Your Extracts
https://help.tableau.com/current/pro/desktop/en-us/extracting_optimize.htm

Other lists of tips

- https://www.tableau.com/about/blog/2016/1/5-tips-make-your-dashboards-more-performant-48574
- https://vizualintelligenceconsulting.com/blog-tableau-performance-tips/