Fuzzy Merge in R

(v. 1.0)

Oscar Torres-Reyna

otorres@princeton.edu

August 2015

http://www.princeton.edu/~otorres/

The focus here is on using the `agrep()` function.

The main goal is to get a *key file* to merge the data files.

Keep in mind that string merging/matching is not exact and require constant checking and some trial-and-error.

The example presented here will try to merge two files which only common variable is company name.

Getting the data (example)

```
# Reading the data, two files 'sp500' and 'nyse'

sp500 <- read.csv(("http://www.princeton.edu/~otorres/sandp500.csv")
head(sp500)
```

	Name	Volume
1	Agilent Technologies	1,723,233
2	Alcoa Inc	25,876,128
3	American Airlines Group Inc	9,930,273
4	Advance Auto Parts Inc	783,826
5	Apple Inc	34,553,089
6	Abbvie Inc. Common Stock	6,169,855

```
nyse <- read.csv(("http://www.princeton.edu/~otorres/nyse.csv")
head(nyse)
```

	Symbol	Name	IPOyear	Sector	Industry
1	DDD	3D Systems Corporation	n/a	Technology	Computer Software: Prepackaged Software
2	MMM	3M Company	n/a	Health Care	Medical/Dental Instruments
3	WBAI	500.com Limited	2013	Consumer Services	Services-Misc. Amusement & Recreation
4	WUBA	58.com Inc.	2013	Technology	Computer Software: Programming, Data Processing
5	AHC	A.H. Belo Corporation	n/a	Consumer Services	Newspapers/Magazines
6	ATEN	A10 Networks, Inc.	2014	Technology	Computer Communications Equipment

Separating string variables

```
# Separating the string variable from each dataset


sp500.name = data.frame(sp500$Name)
names(sp500.name)[names(sp500.name)=="sp500.Name"] = "name.sp"
sp500.name$name.sp = as.character(sp500.name$name.sp)
sp500.name = unique(sp500.name) # Removing duplicates
head(sp500.name)
```

```
                     name.sp
1        Agilent Technologies
2                   Alcoa Inc
3 American Airlines Group Inc
4      Advance Auto Parts Inc
5                   Apple Inc
6    Abbvie Inc. Common Stock
```

Separating string variables

```
# Separating the string variable from each dataset


nyse.name = data.frame(nyse$Name)
names(nyse.name)[names(nyse.name)=="nyse.Name"] = "name.nyse"
nyse.name$name.nyse = as.character(nyse.name$name.nyse)
nyse.name = unique(nyse.name) # Removing duplicates
head(nyse.name)

                name.nyse
1 3D Systems Corporation
2              3M Company
3         500.com Limited
4             58.com Inc.
5  A.H. Belo Corporation
6      A10 Networks, Inc.
```

Matching strings

```
# Matching string variables from sp500 to nyse data

sp500.name$name.nyse <- ""   # Creating an empty column

for(i in 1:dim(sp500.name)[1]) {
  x <- agrep(sp500.name$name.sp[i], nyse.name$name.nyse,
             ignore.case=TRUE, value=TRUE,
             max.distance = 0.05, useBytes = TRUE)
  x <- paste0(x,"")
  sp500.name$name.nyse[i] <- x
}
```

[See next slide for the resulting file]

For more info/details, type ?agrep

NOTE: The following warning may pop-up, : "There were 28 warnings (use warnings() to see them)"

Matching strings

 # First column has the original names in the file sp500; second column has the corresponding matched names from the nyse file. This file is the *key file* to merge the full datasets (make sure to check it first)

```
head(sp500.name, 13)
```

	name.sp	name.nyse
1	Agilent Technologies	Agilent Technologies, Inc.
2	Alcoa Inc	Alcoa Inc.
3	American Airlines Group Inc	
4	Advance Auto Parts Inc	Advance Auto Parts Inc
5	Apple Inc	
6	Abbvie Inc. Common Stock	
7	Amerisourcebergen Corp	AmerisourceBergen Corporation (Holding Co)
8	Abbott Laboratories	Abbott Laboratories
9	Ace Ltd	Third Point Reinsurance Ltd.
10	Accenture Plc	Accenture plc.
11	Adobe Systems Incorporated	
12	Analog Devices	
13	Archer Daniels Midland Company	Archer-Daniels-Midland Company

Notice line 9 has the wrong match. Further cleaning and data inspection is needed when performing fuzzy matching.

Merging the *key file*

```
# Merging the key file sp500.name to the original dataset sp500

sp500 = merge(sp500, sp500.name, by.x=c("Name"), by.y=c("name.sp"), all= TRUE)

head(sp500)
                      Name    Volume           name.nyse
1 21St Century Fox Class A 6,206,568
2 21St Century Fox Class B 1,756,450
3               3M Company 1,606,555          3M Company
4      Abbott Laboratories 4,747,144 Abbott Laboratories
5 Abbvie Inc. Common Stock 6,169,855
6            Accenture Plc 1,497,234      Accenture plc.


# Renaming the original variable for company name in dataset sp500 (this is for
better tracking when merging the nyse file).

names(sp500)[names(sp500)=="Name"] = "name.sp"
```

Merging the two files

```
# Merging the two original data files (keeping all data from both)

companies = merge(sp500, nyse, by.x=c("name.nyse"), by.y=c("Name"), all = TRUE)

companies[sample(nrow(companies), 30), ]
```

	name.nyse	name.sp	Volume	Symbol	IPOyear	Sector	Industry
3356	Winnebago Industries, Inc.	<NA>	<NA>	WGO	n/a	Consumer Non-Durables	Homebuilding
1297	F.N.B. Corporation	<NA>	<NA>	FNB^E	n/a	n/a	n/a
894	Colgate-Palmolive Company	Colgate-Palmolive Company	1,703,477	CL	n/a	Consumer Non-Durables	Package Goods/Cosmetics
1185	Embraer-Empresa Brasileira de Aeronautica	<NA>	<NA>	ERJ	2000	Capital Goods	Aerospace
2080	Mobileye N.V.	<NA>	<NA>	MBLY	2014	Technology	Computer Software: Prepackaged Software
1956	Macerich Company (The)	Macerich Company	464,872	MAC	1994	Consumer Services	Real Estate Investment Trusts
1308	Federal Agricultural Mortgage Corporation	<NA>	<NA>	AGM^A	n/a	n/a	n/a
405	Ashford Hospitality Trust Inc	<NA>	<NA>	AHT^E	n/a	n/a	n/a
413	Aspen Insurance Holdings Limited	<NA>	<NA>	AHL^C	n/a	n/a	n/a
1307	Federal Agricultural Mortgage Corporation	<NA>	<NA>	AGM.A	n/a	n/a	n/a
1412	Furmanite Corporation	<NA>	<NA>	FRM	n/a	Basic Industries	Engineering & Construction
2786	Ship Finance International Limited	<NA>	<NA>	SFL	n/a	Transportation	Marine Transportation
2994	TELUS Corporation	<NA>	<NA>	TU	n/a	Public Utilities	Telecommunications Equipment
1914	Level 3 Communications, Inc.	Level 3 Communications	1,346,416	LVLT	2011	Public Utilities	Telecommunications Equipment
239	Alibaba Group Holding Limited	<NA>	<NA>	BABA	2014	Miscellaneous	Business Services
559	Blackrock California Municipal 2018 Term Trust	Blackrock	748,524	BJZ	2001	n/a	n/a
2933	Sysco Corporation	Sysco Corp	2,034,966	SYY	n/a	Consumer Non-Durables	Food Distributors
2154	Nautilus Group, Inc. (The)	<NA>	<NA>	NLS	n/a	Consumer Non-Durables	Recreational Products/Toys
2301	Nuveen Municipal Opportunity Fund, Inc.	<NA>	<NA>	NIO	1991	n/a	n/a
15		Qualcomm Inc	8,266,265	<NA>	<NA>	<NA>	<NA>
14		Qorvo Inc	1,436,749	<NA>	<NA>	<NA>	<NA>
2071	Miller/Howard High Income Equity Fund	<NA>	<NA>	HIE	2014	n/a	n/a
1935	Lockheed Martin Corporation	Lockheed Martin Corp	1,573,547	LMT	n/a	Capital Goods	Military/Government/Technical
649	Brandywine Realty Trust	<NA>	<NA>	BDN	n/a	Consumer Services	Real Estate Investment Trusts
1586	Healthcare Realty Trust Incorporated	<NA>	<NA>	HR	1993	Consumer Services	Real Estate Investment Trusts
1728	Invesco Trust for Investment Grade New York Municipal	<NA>	<NA>	VTN	n/a	n/a	n/a
3352	Willbros Group, Inc.	<NA>	<NA>	WG	1996	Energy	Oilfield Services/Equipment
2636	Reinsurance Group of America, Incorporated	<NA>	<NA>	RZA	n/a	Finance	Life Insurance
1945	Luxottica Group, S.p.A.	<NA>	<NA>	LUX	1990	Health Care	Ophthalmic Goods
2257	Nu Skin Enterprises, Inc.	<NA>	<NA>	NUS	1996	Health Care	Other Pharmaceuticals

NOTE: Since the key file was not cleaned or corrected, when merging the files we still have some wrong matches

Merging the two files

```
# Merging the two original data files (keeping only perfect matches)

companies1 = merge(sp500, nyse, by.x=c("name.nyse"), by.y=c("Name"))

companies1[sample(nrow(companies1), 30), ]
```

	name.nyse	name.sp	Volume	Symbol	IPOyear	Sector	Industry
183	Emerson Electric Company	Emerson Electric Company	2,296,936	EMR	n/a	Energy	Consumer Electronics/Appliances
237	Helmerich & Payne, Inc.	Helmerich & Payne	1,858,952	HP	n/a	Energy	Oil & Gas Production
437	TEGNA Inc.	Tegna Inc	1,080,754	GCI	2015	Consumer Services	Newspapers/Magazines
126	CIT Group Inc (DEL)	Citigroup Inc	16,896,288	CIT	n/a	Finance	Finance: Consumer Services
38	American International Group, Inc.	American International Group	5,653,643	AIG	n/a	Finance	Property-Casualty Insurers
162	Dominion Resources, Inc.	Dominion Resources	2,426,675	DCUA	n/a	Public Utilities	Electric Utilities: Central
463	U.S. Bancorp	U.S. Bancorp	3,984,404	USB	n/a	Finance	Major Banks
351	PerkinElmer, Inc.	Perkinelmer	793,740	PKI	n/a	Capital Goods	Biotechnology: Laboratory Analytical Instruments
457	Tyson Foods, Inc.	Tyson Foods	2,631,171	TSN	n/a	Consumer Non-Durables	Meat/Poultry/Fish
130	Coach, Inc.	Coach Inc	2,225,233	COH	n/a	Consumer Non-Durables	Apparel
133	Colgate-Palmolive Company	Colgate-Palmolive Company	1,703,477	CL	n/a	Consumer Non-Durables	Package Goods/Cosmetics
338	Noble Energy Inc.	Noble Energy Inc	3,706,337	NBL	n/a	Energy	Oil & Gas Production
404	Sempra Energy	Sempra Energy	1,330,101	SRE	n/a	Public Utilities	Natural Gas Distribution
249	International Paper Company	International Paper Company	3,673,875	IP	n/a	Basic Industries	Paper
85	BB&T Corporation	BB&T Corp	4,117,963	BBT^G	n/a	n/a	n/a
6	Acuity Brands Inc	L Brands Inc	823,995	AYI	n/a	Consumer Durables	Building Products
301	Marsh & McLennan Companies, Inc.	Marsh & Mclennan Companies	2,114,387	MMC	n/a	Finance	Specialty Insurers
167	Dr Pepper Snapple Group, Inc	Dr Pepper Snapple Group Inc	742,071	DPS	n/a	Consumer Non-Durables	Beverages (Production/Distribution)
80	Bank of America Corporation	Bank of America Corp	66,656,351	BAC^L	n/a	n/a	n/a
274	Kimco Realty Corporation	Kimco Realty Corp	2,141,077	KIM^J	n/a	n/a	n/a
141	CONSOL Energy Inc.	Consol Energy Inc	9,348,755	CNX	1999	Energy	Coal Mining
168	DTE Energy Company	Dte Energy Company	1,195,981	DTQ	n/a	Public Utilities	Electric Utilities: Central
90	Best Buy Co., Inc.	Best Buy Co	3,229,841	BBY	n/a	Consumer Services	Consumer Electronics/Video Chains
250	Interpublic Group of Companies, Inc. (The)	Interpublic Group of Companies	1,451,810	IPG	n/a	Technology	Advertising
482	Vornado Realty Trust	Vornado Realty Trust	470,069	VNO^G	n/a	n/a	n/a
369	Public Storage	Public Storage	436,269	PSA^T	n/a	Consumer Services	Real Estate Investment Trusts
289	Lincoln National Corporation	Lincoln National Corp	2,784,853	LNC	n/a	Finance	Life Insurance
412	Southern Company (The)	Southern Company	3,797,694	SOJA	2015	Public Utilities	Electric Utilities: Central
307	McKesson Corporation	Mckesson Corp	6,463,091	MCK	n/a	Health Care	Other Pharmaceuticals
107	Capital One Financial Corporation	Capital One Financial Corp	2,035,595	COF^C	n/a	Finance	Major Banks

NOTE: Since the key file was not cleaned or corrected, when merging the files we still have some wrong matches

Sources

http://stats.stackexchange.com/questions/3425/how-to-quasi-match-two-vectors-of-strings-in-r

http://stackoverflow.com/questions/8273313/random-rows-in-dataframe-in-r

http://www.barchart.com/stocks/sp500.php

http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NYSE