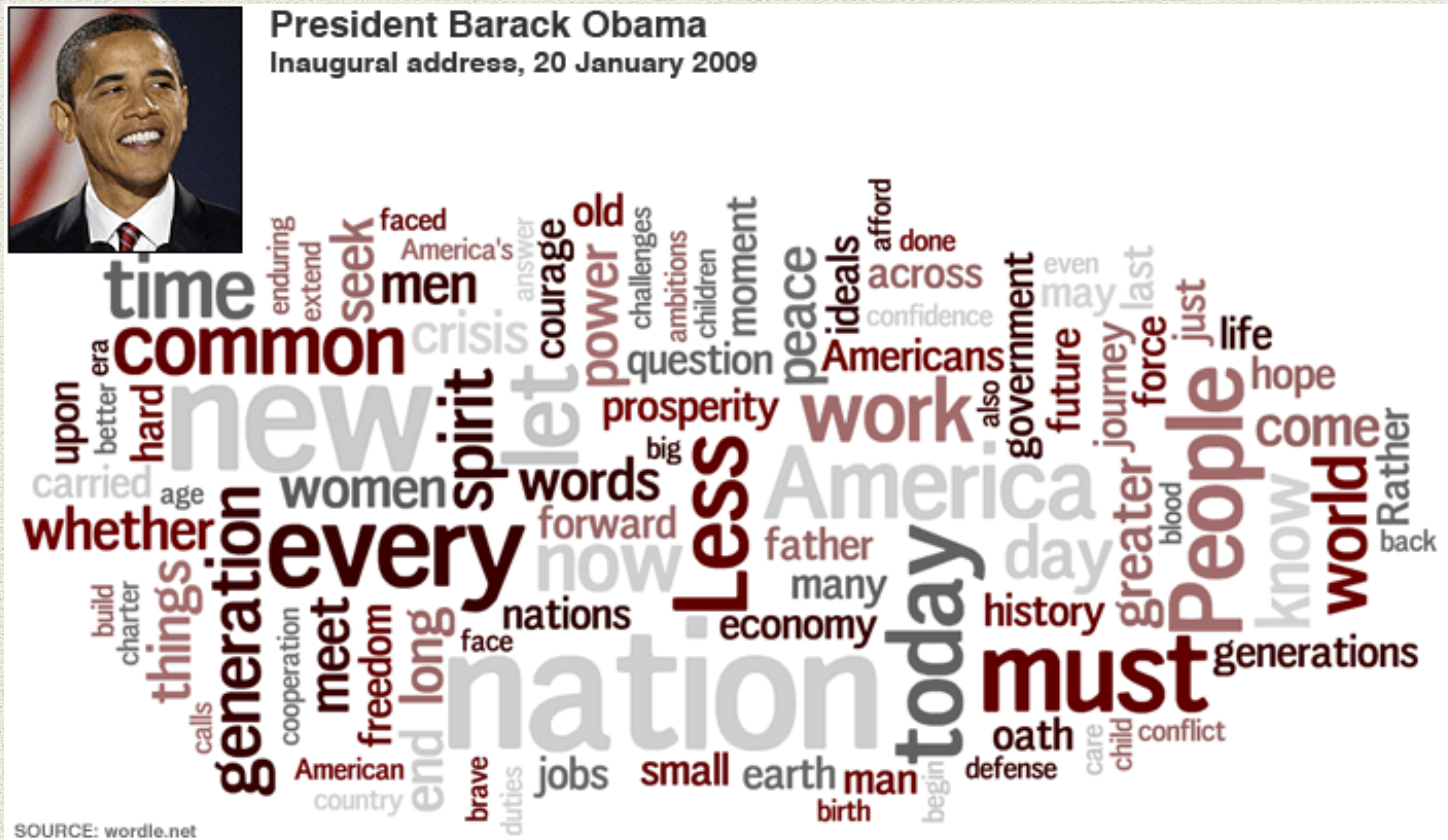
Using Simple Frequencies

Practical 3: Simple Frequencies

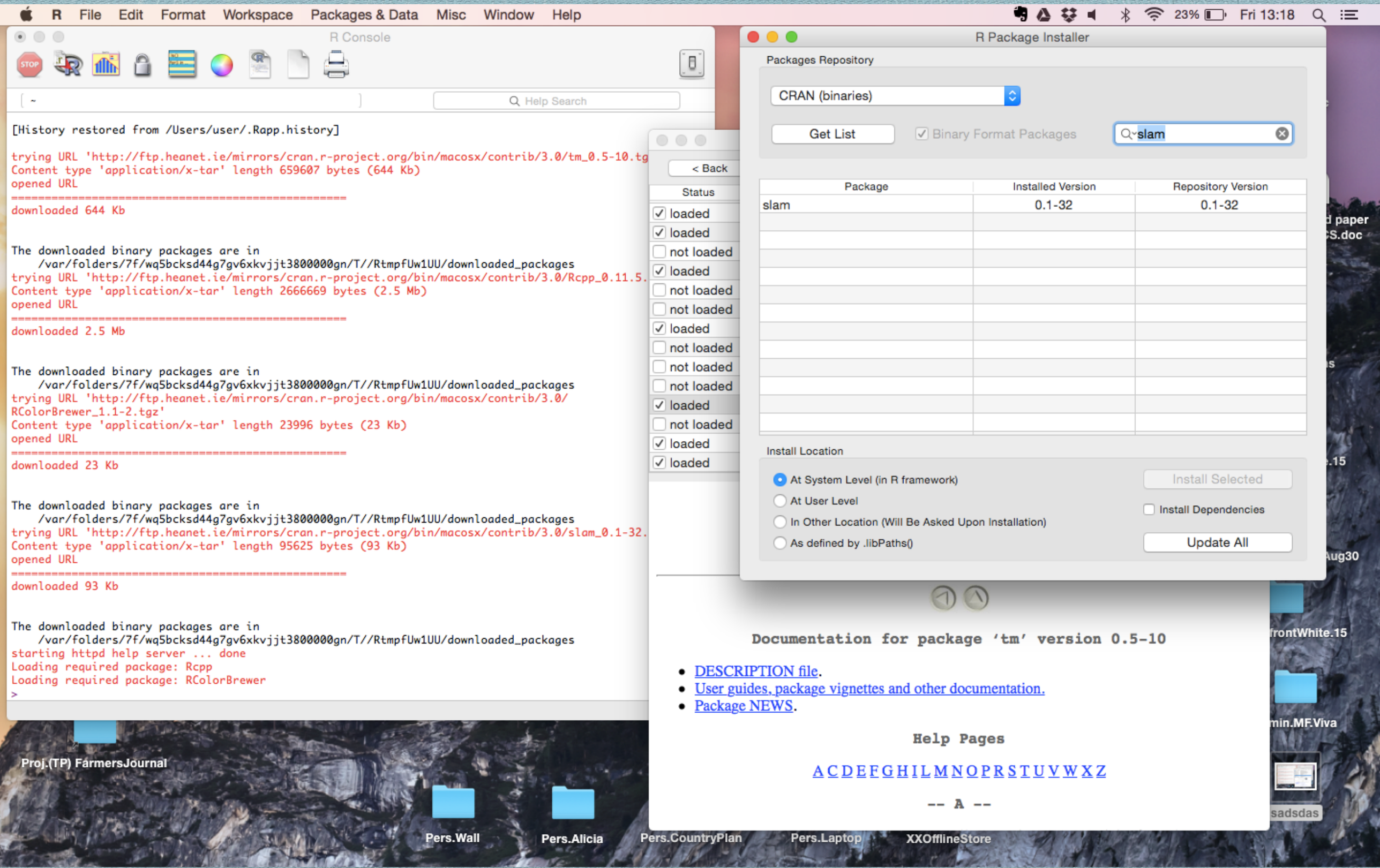
Word Clouds
Q1


President Barack Obama
Inaugural address, 20 January 2009
faced
old
afford
done
America's
enduring
extend
seek
answer
courage
challenges
ambitions
children
moment
peace
ideals
across
even
time
men
power
confidence
government
may
last
just
common
crisis
question
Americans
life
era
upon
better
hard
new
spirit
let
prosperity
work
also
future
journey
force
hope
People
come
big
carried
age
women
words
Less
America
Rather
whether
forward
father
day
greater
blood
know
world
back
build
charter
things
generation
every
now
many
history
cooperation
meet
freedom
long
face
nations
economy
today
must
generations
calls
nation
oath
care
child
conflict
American
end
brave
duties
jobs
small
earth
man
begin
defense
country
birth
SOURCE: wordle.net

Using Wordcloud in R

- Install wordcloud package
- Install
  - tm package (for text mining)
  - Rcpp package
  - RColourBrewer package
  - slam

Installing Packages



Loading Packages

R File Edit Format Workspace Packages & Data Misc Window Help

24% Fri 13:15

R Console

```
[History restored from /Users/user/.Rapp.history]

trying URL 'http://ftp.heanet.ie/mirrors/cran.r-project.org/bin/macosx/contrib/3.0/tm_0.5-10.tg
Content type 'application/x-tar' length 659607 bytes (644 Kb)
opened URL
==================================================
downloaded 644 Kb


The downloaded binary packages are in
    /var/folders/7f/wq5bcksd44g7gv6xkvjjt3800000gn/T//RtmpfUw1UU/downloaded_packages
trying URL 'http://ftp.heanet.ie/mirrors/cran.r-project.org/bin/macosx/contrib/3.0/Rcpp_0.11.5.
Content type 'application/x-tar' length 2666669 bytes (2.5 Mb)
opened URL
==================================================
downloaded 2.5 Mb


The downloaded binary packages are in
    /var/folders/7f/wq5bcksd44g7gv6xkvjjt3800000gn/T//RtmpfUw1UU/downloaded_packages
trying URL 'http://ftp.heanet.ie/mirrors/cran.r-project.org/bin/macosx/contrib/3.0/
RColorBrewer_1.1-2.tgz'
Content type 'application/x-tar' length 23996 bytes (23 Kb)
opened URL
==================================================
downloaded 23 Kb


The downloaded binary packages are in
    /var/folders/7f/wq5bcksd44g7gv6xkvjjt3800000gn/T//RtmpfUw1UU/downloaded_packages
trying URL 'http://ftp.heanet.ie/mirrors/cran.r-project.org/bin/macosx/contrib/3.0/slam_0.1-32.
Content type 'application/x-tar' length 95625 bytes (93 Kb)
opened URL
==================================================
downloaded 93 Kb


The downloaded binary packages are in
    /var/folders/7f/wq5bcksd44g7gv6xkvjjt3800000gn/T//RtmpfUw1UU/downloaded_packages
starting httpd help server ... done
Loading required package: Rcpp
Loading required package: RColorBrewer
>
```

R Package Installer

Packages Repository

CRAN (binaries)

R Package Manager

< Back | Fwd > | Refresh List

Status	Package	Description
loaded	RColorBrewer	ColorBrewer palettes
loaded	Rcpp	Seamless R and C++ Integration
not loaded	rpart	Recursive Partitioning and Regression Trees
loaded	slam	Sparse Lightweight Arrays and Matrices
not loaded	spatial	Functions for Kriging and Point Pattern Analysis
not loaded	splines	Regression Spline Functions and Classes
loaded	stats	The R Stats Package
not loaded	stats4	Statistical Functions using S4 Classes
not loaded	survival	Survival Analysis
not loaded	tcltk	Tcl/Tk Interface
loaded	tm	Text Mining Package
not loaded	tools	Tools for Package Development
loaded	utils	The R Utils Package
loaded	wordcloud	Word Clouds

Text Mining Package

Documentation for package 'tm' version 0.5-10

- DESCRIPTION file.
- User guides, package vignettes and other documentation.
- Package NEWS.

Help Pages

A C D E F G H I L M N O P R S T U V W X Z

-- A --

Proj.(TP) FarmersJournal

Pers.Wall

Pers.Alicia

Pers.CountryPlan

Pers.Laptop

XXOfflineStore

Using Wordcloud

- In the R console window type:

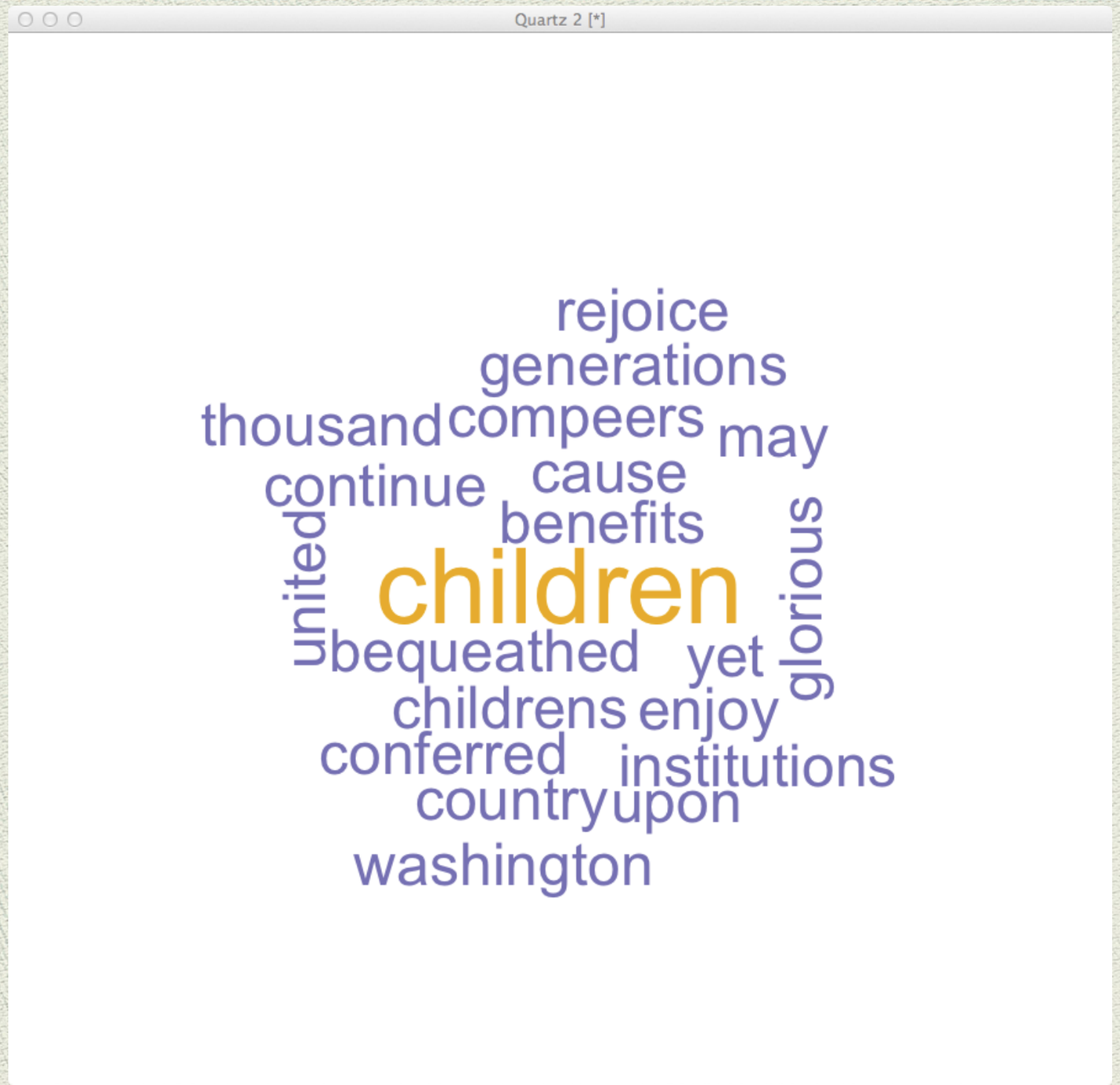
> library(wordcloud) [cr]

> library(tm) [cr]

> wordcloud("May our children and our childrens children to a thousand generations continue to enjoy the benefits conferred upon us by a united country and have cause yet to rejoice under those glorious institutions bequeathed us by Washington and his compeers.”, colors=brewer.pal(6,"Dark2"),random.order=FALSE)

Do not cut and paste this text, as it will not work work with smart-quotes

Using Wordcloud: Result



Using Wordcloud: Errors

- If you get errors about unloaded packages then load them using the package-loader window
- If you get an error like this:

```
> wordcloud( “a long list of repeated words
”,colors=brewer.pal(6,”Dark2"), random.order=FALSE)

Error: unexpected input in "wordcloud(,"
```

then you have used the wrong type of quotes

OR…from file

- Open the file with the script/program
- Use the menu-command:

  “Edit” ->“Source Document”

- Or in the R Console type:

```
> source('filename.R')
```

Wordcloud Practical

- Now, put your own set of words into this, about 30-50
- Only some words are included ?
- Change the number of repeated words and see what happens?
- Hint: Note the prompt give in the console

```
> wordcloud("list word word word a washington list time time repeat repeat repeat the the the cloud
cloud cloud",min.freq=1,colors=brewer.pal(6,"Dark2"),random.order=TRUE)
  wordcloud(words, freq, scale = c(4, 0.5), min.freq = 3, max.words = Inf, random.order = TRUE, random.color = FALSE, rot.per = 0.1, colors = "black",
  ordered.colors = FALSE, use.r.layout = FALSE, fixed.asp = TRUE, ...)
```

Google Ngram Viewer Q2

Corpora & Word Frequency

a. Put in “Mark Keane” as a search term and explain the peaks that appear in the graph over time.

b. Put your own name in and describe what happens, explaining where the hits are coming from.

c. Pick a word that you think is a recent introduction into the English language (like “exit strategy”) and plot its emergence, showing the graphs. If it actually emerges before you thought, explain why?

d. Describe some of the effects of smoothening these graphs with different values?

e. Do a comparison between 3 or more related terms to see how their relative frequencies have changed over time*. Is there anything surprising about how these terms differ in their frequency and, if so, why? Why do you think the frequencies vary in the way they do.

f. Use syntactic tags in a search for two words that are the same but syntactically different (e.g., fish-verb, fish-noun; *do not use fish*) and report what you find.

g. Think of some major cultural change that has happened over the last 500 years and some words that could denote to this event/events. Check these words of the relevant time-period. Report what you find.

Normalisation
Q3

Too Easy…

Using a spreadsheet set up your own list of 15 words and give each a made-up frequency between 0 and 2000 for each of three years (2010, 2011, 2012). Now perform two different normalisations on them:

a. **Method1**: produce a normalised frequency for each word in each year, using the total N of words in all years

b. **Method2:** produce a normalised frequency for each word in each year, using the N of words in a given year

c. Does normalising by **method1** or **method2** make a big difference to the scores produced? Graph the difference and comment on it.

Google Trends
Q4

Too Hard?…

Find the article by Choi & Varian (2009/2011/2012) and find the R program they give for their Ford prediction model.

What would you have to do to get it to work ? Can you do this?