TURNING A TWITTER HASHTAG INTO A WORD CLOUD

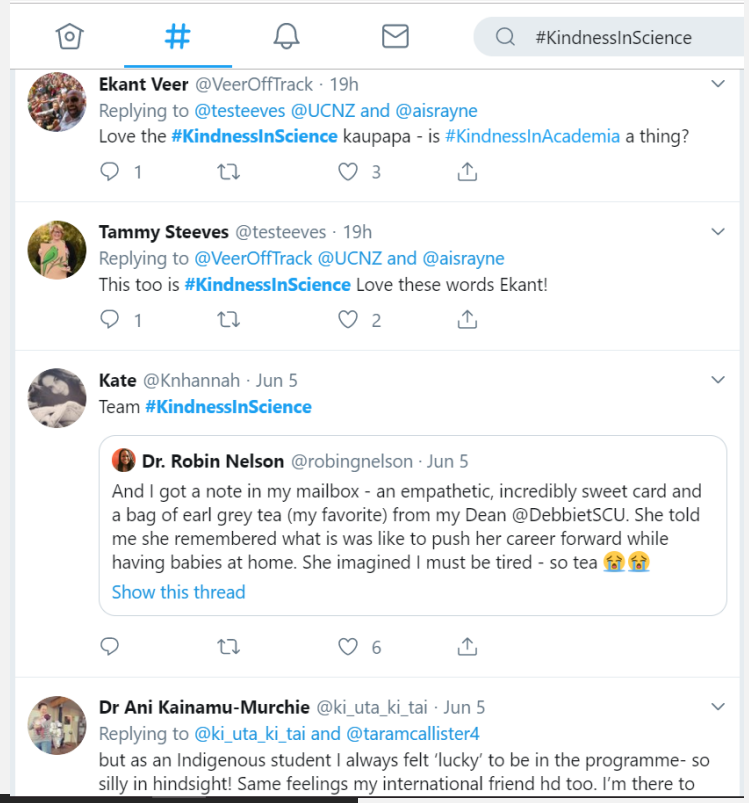
Jennifer Pannell

Ecology postdoc, AUT

@jennypannell

WAIT, WHAT?

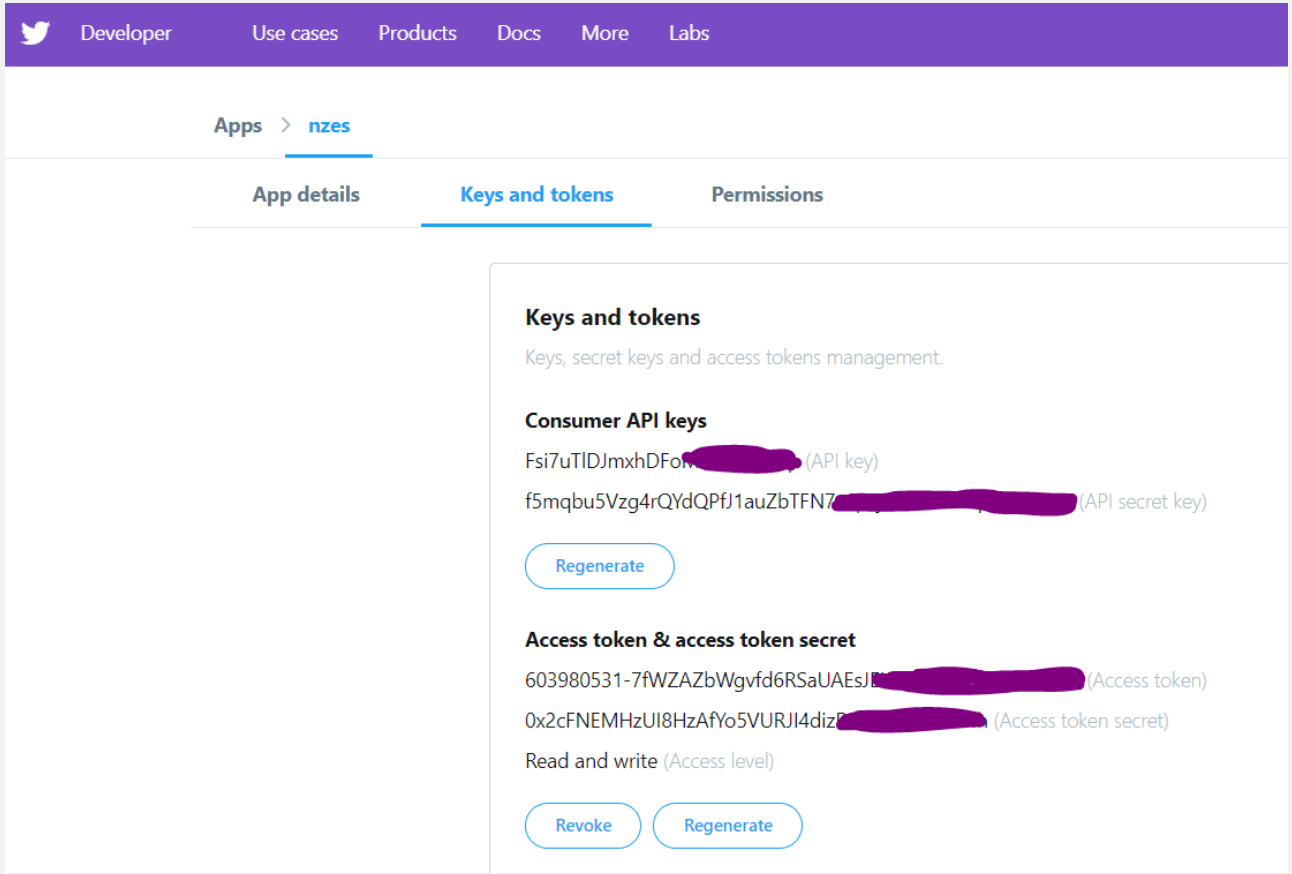
- Download all tweets from a hashtag



- Turn the tweets into a word cloud

STEP 1: GET TWITTER CREDENTIALS

- Go to https://developer.twitter.com/en/apps
- Create a blank app
- Copy consumer API keys, access token & secret

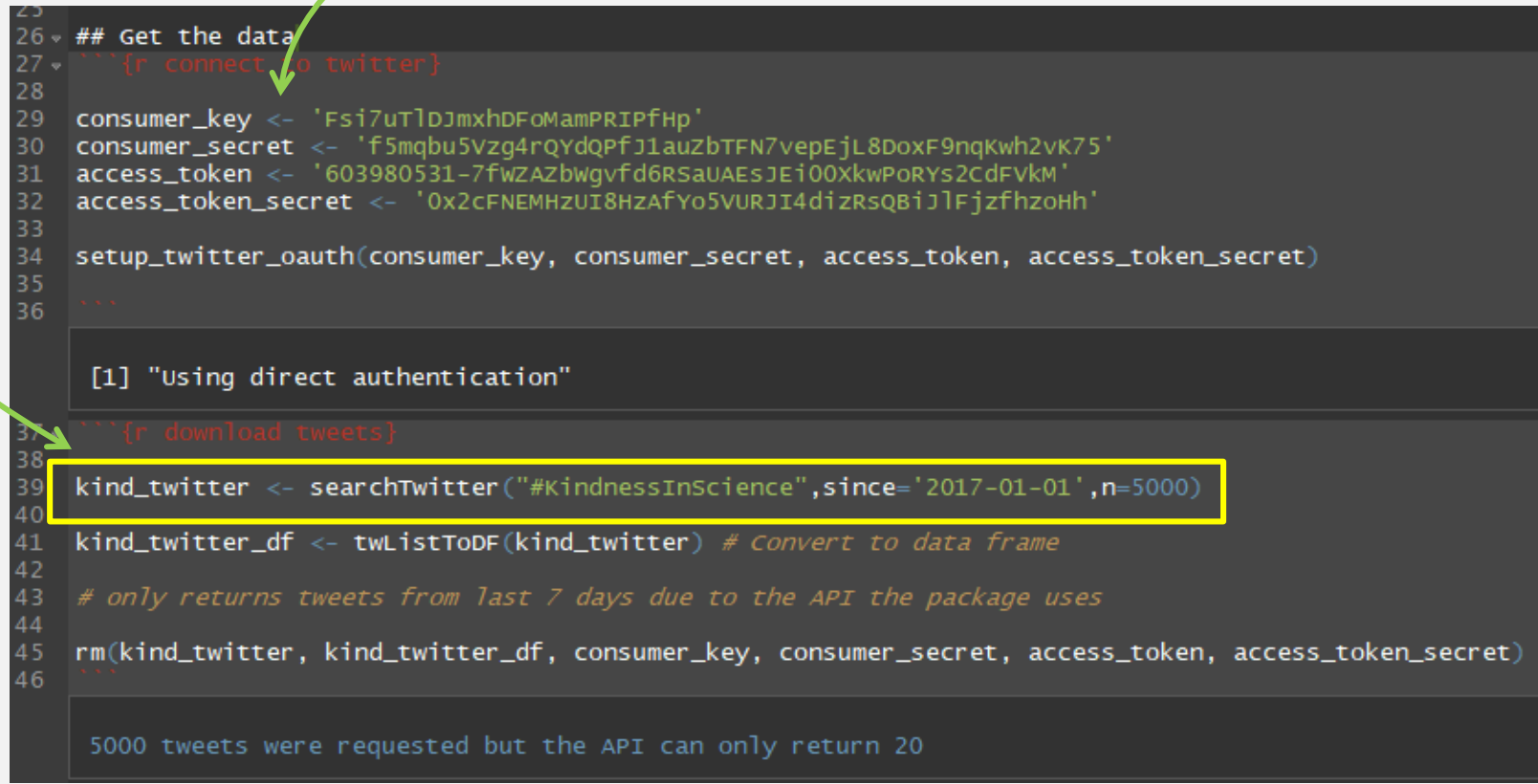


PACKAGES REQUIRED: twitteR

STEP 2: SCRAPE THE HASHTAG

Provide credentials and connect to twitter

Search parameters, and max no. tweets to return



````
## Get the data
```{r connect to twitter}

consumer_key <- 'Fsi7uTlDJmxhDFoMamPRIPfHp'
consumer_secret <- 'f5mqbu5Vzg4rQYdQPfJ1auZbTFN7vepEjL8DoxF9nqKwh2vK75'
access_token <- '603980531-7fWZAZbWgvfd6RSaUAEsJEiOOXkwPoRYs2CdFVkM'
access_token_secret <- '0x2cFNEMHzUI8HzAfYo5VURJI4dizRsQBiJlFjzfhzoHh'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_token_secret)

```

[1] "Using direct authentication"

```{r download tweets}

kind_twitter <- searchTwitter("#KindnessInScience",since='2017-01-01',n=5000)

kind_twitter_df <- twListToDF(kind_twitter) # Convert to data frame

# only returns tweets from last 7 days due to the API the package uses

rm(kind_twitter, kind_twitter_df, consumer_key, consumer_secret, access_token, access_token_secret)
```

5000 tweets were requested but the API can only return 20
````

@jennypannell

PACKAGES REQUIRED: twitteR

STEP 2: SCRAPE THE HASHTAG

Provide credentials and connect to twitter

Search parameters, and max no. tweets to return

```
## Get the data
```{r connect to twitter}

consumer_key <- 'Fsi7uTlDJmxhDFoMamPRIPfHp'
consumer_secret <- 'f5mqbu5Vzg4rQYdQPfJ1auZbTFN7vepEjL8DoxF9nqKwh2vK75'
access_token <- '603980531-7fWZAZbWgvfd6RSaUAEsJEiOOXkwPoRYs2CdFVkM'
access_token_secret <- '0x2cFNEMHzUI8HzAfYo5VURJI4dizRsQBiJlFjzfhzoHh'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_token_secret)

```

[1] "Using direct authentication"

```{r download tweets}

kind_twitter <- searchTwitter("#KindnessInScience",since='2017-01-01',n=5000)

kind_twitter_df <- twListToDF(kind_twitter) # Convert to data frame

# only returns tweets from last 7 days due to the API the package uses

rm(kind_twitter, kind_twitter_df, consumer_key, consumer_secret, access_token, access_token_secret)
```

5000 tweets were requested but the API can only return 20
```

- Issue: twitter only allows searching back as far as 7 days
- Python alternative using screen scraping: https://github.com/taspinar/twitterscraper

@jennypannell

PACKAGES REQUIRED: twitteR

STEP 2: SCRAPE THE HASHTAG

kind_twitter_df

Filter

	text	favorited	favoriteCount	replyToSN	created	truncate
1	@gringene_bio @WilliamPearman <U+0001F446><U...	FALSE	0	gringene_bio	2019-06-11 14:28:29	FALSE
12	@taramcallister4 but as an Indigenous student I alway...	FALSE	2	ki_uta_ki_tai	2019-06-04 21:44:10	TRUE
9	@testeeves @UCNZ @aisrayne Love the #KindnessInSc...	FALSE	3	testeeves	2019-06-10 03:39:03	FALSE
13	@TroyBaisden @sandrinha2021 @ChiefSciAdvisor Th...	FALSE	1	TroyBaisden	2019-06-03 20:13:53	TRUE
8	@VeerOffTrack @UCNZ @aisrayne It sure can be! We s...	FALSE	3	VeerOffTrack	2019-06-10 04:10:23	TRUE
10	@VeerOffTrack @UCNZ @aisrayne This too is #Kindne...	FALSE	2	VeerOffTrack	2019-06-10 03:37:11	FALSE
5	I especially like seeing inclusive & @PunahaMatati...	FALSE	11	NA	2019-06-10 18:34:15	TRUE
7	In preparation for this week's @RLadiesAKL I've been ...	FALSE	14	NA	2019-06-10 11:28:40	TRUE
16	Mentoring and equity- leaving a legacy (shout out to ...	FALSE	7	NA	2019-06-02 05:15:02	FALSE
14	RT @CillaWehi: I'm clear we need structural change of ...	FALSE	0	NA	2019-06-02 09:16:51	FALSE
4	RT @CillaWehi: I especially like seeing inclusive & ...	FALSE	0	NA	2019-06-10 19:23:42	FALSE
2	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 20:58:52	FALSE
3	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 19:34:01	FALSE
6	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 13:40:21	FALSE
11	Team #KindnessInScience https://t.co/Bjr9AjN24U	FALSE	6	NA	2019-06-05 04:18:26	FALSE
15	Welcome @itchyshin to twitter. Nice to see you here <...	FALSE	0	NA	2019-06-02 05:37:42	FALSE

PACKAGES REQUIRED: tm, tidytext, stringr

STEP 3: CLEAN UP TWEET DATA

kind_twitter_df

Filter

	text	favorited	favoriteCount	replyToSN	created	truncate
1	@gringene_bio @WilliamPearman <U+0001F446><U...	FALSE	0	gringene_bio	2019-06-11 14:28:29	FALSE
12	@taramcallister4 but as an Indigenous student I alway...	FALSE	2	ki_uta_ki_tai	2019-06-04 21:44:10	TRUE
9	@testeeves @UCNZ @aisrayne Love the #KindnessInSc...	FALSE	3	testeeves	2019-06-10 03:39:03	FALSE
13	@TroyBaisden @sandrinha2021 @ChiefSciAdvisor Th...	FALSE	1	TroyBaisden	2019-06-03 20:13:53	TRUE
8	@VeerOffTrack @UCNZ @aisrayne It sure can be! We s...	FALSE	3	VeerOffTrack	2019-06-10 04:10:23	TRUE
10	@VeerOffTrack @UCNZ @aisrayne This too is #Kindne...	FALSE	2	VeerOffTrack	2019-06-10 03:37:11	FALSE
5	I especially like seeing inclusive & @PunahaMatati...	FALSE	11	NA	2019-06-10 18:34:15	TRUE
7	In preparation for this week's @RLadiesAKL I've been ...	FALSE	14	NA	2019-06-10 11:28:40	TRUE
16	Mentoring and equity- leaving a legacy (shout out to ...	FALSE	7	NA	2019-06-02 05:15:02	FALSE
14	RT @CillaWehi: I'm clear we need structural change of ...	FALSE	0	NA	2019-06-02 09:16:51	FALSE
4	RT @CillaWehi: I especially like seeing inclusive & ...	FALSE	0	NA	2019-06-10 19:23:42	FALSE
2	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 20:58:52	FALSE
3	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 19:34:01	FALSE
6	RT @jennypannell: In preparation for this week's @RLa...	FALSE	0	NA	2019-06-10 13:40:21	FALSE
11	Team #KindnessInScience https://t.co/Bjr9AjN24U	FALSE	6	NA	2019-06-05 04:18:26	FALSE
15	Welcome @itchyshin to twitter. Nice to see you here <...	FALSE	0	NA	2019-06-02 05:37:42	FALSE

kindTable

Filter

	word	n
1	kindnessinscience	242
2	status	97
3	twittercom	97
4	https	93
5	science	45
6	testeeves	23
7	pictwittercom	20
8	conserteam	16
9	kindness	15
10	sgalla	15
11	http	12
12	love	12
13	change	11
14	inclusive	11
15	research	11
16	cillawehi	10
17	diversity	9

@jennypannell

PACKAGES REQUIRED: tm, tidytext, stringr

STEP 3: CLEAN UP TWEET DATA

Break up tweets into vector of single words

```
## Tidy up the text
Tweet text needs to be tidied by splitting into words, removing punctuation and numbers
and words like "the", "and", etc.
```{r wrangle data and tidy}
# make sure the tweets are stored as character data
kind$text<-as.character(kind$text)

# before we split up, make sure any words that need to stay together are replaced
kind$text<-gsub("te reo", "tereo", kind$text)
kind$text<-gsub("ka rawe", "karawe", kind$text)
kind$text<-gsub("kia ora", "kiaora", kind$text)
kind$text<-gsub("kia kaha", "kiakaha", kind$text)


#use the unnest_tokens function in tidytext to make the word frequency table
# make into data frame with 1 row per word

kindTable <- kind %>%
  unnest_tokens(word, text)
```

Remove unwanted words, punctuation, numbers, etc

```
#remove numbers and punctuation
kindTable$word<-removeNumbers(kindTable$word) # remove numbers & punctuation
kindTable$word<-removePunctuation(kindTable$word)
kindTable<-kindTable[!(is.na(kindTable$word) | kindTable$word==""), ] # remove
blank/whitespace words

#remove stop words - aka typically very common words such as "the", "of" etc
# now 7770 words
data(stop_words)

kindTable <- kindTable %>%
  anti_join(stop_words)

rm(stop_words)

#add word count to table
kindTable <- kindTable %>%
  dplyr::count(word, sort=TRUE)
```

Count how often each word occurs

@jennypannell
```

PACKAGES REQUIRED: wordcloud2 (OPTIONAL: htmlwidgets, webshot)

STEP 4: MAKE WORD CLOUD!


```
# Subset the data for the word cloud to a manageable size
```{r subset for word cloud}
# Too many for word cloud to cope with, choose only words that occur more than three times
clouddat<-subset(kindTable, n>2)
```

#Create the word cloud
Use a custom palette or built in, I used online palette generators
```{r create the word cloud}
# Create Palette
kindPalette <- c("#D5B0AC", "#CEA0EA", "#684551", "#402E2A", "#9CD08F")

mycloud<-wordcloud2(clouddat, color=rep_len(kindPalette, nrow(clouddat)), size=0.5)
mycloud
# save it in html
saveWidget(mycloud,"kis.html",selfcontained = F)
# save as a pdf
webshot::install_phantomjs()
webshot("kis.html","kis.pdf", delay =20, vwidth = 1500, vheight=1000)

```
```

@jennypannell

PACKAGES REQUIRED: wordcloud2 (OPTIONAL: htmlwidgets, webshot)

STEP 4: MAKE WORD CLOUD!

- I used a palette from https://coolors.co/app but you can use built-in colours
- Issues: wordcloud2 is glitchy - difficult to customise cloud and save output.

```
# Subset the data for the word cloud to a manageable size
```{r subset for word cloud}
# Too many for word cloud to cope with, choose only words that occur more than three times
clouddat<-subset(kindTable, n>2)
```

#Create the word cloud
Use a custom palette or built in, I used online palette generators
```{r create the word cloud}
# Create Palette
kindPalette <- c("#D5B0AC", "#CEA0EA", "#684551", "#402E2A", "#9CD08F")

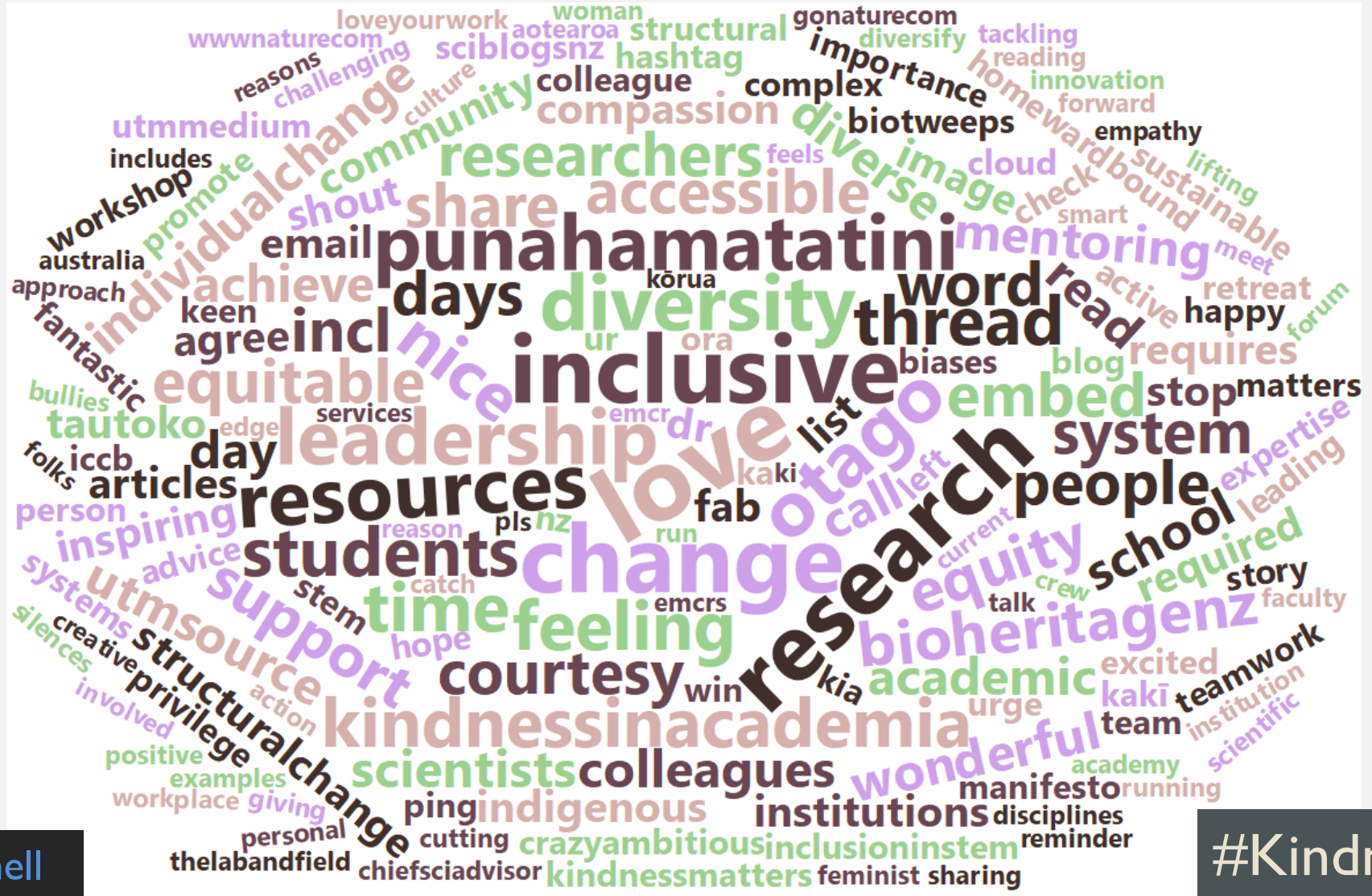
mycloud<-wordcloud2(clouddat, color=rep_len(kindPalette, nrow(clouddat)), size=0.5)
mycloud
# save it in html
saveWidget(mycloud,"kis.html",selfcontained = F)
# save as a pdf
webshot::install_phantomjs()
webshot("kis.html","kis.pdf", delay =20, vwidth = 1500, vheight=1000)

```
```

@jennypannell

TA-DA!

github.com/pannellj/twitter-clouds



@jennypannell

#KindnessInScience