**Exercises 1**

Create a word cloud for Minnesota. Use “Minnesota” as a term and retrieve 500 tweets. Your word cloud should be similar to the following image. Explain your word cloud. What insights (if any) do you find in your word cloud?

****

**Exercise 2**

Remove the specific or obvious words from the previous word cloud. For example, removing “Minnesota” from word cloud may reveal better insights.



Are there better insights?

Hint for removing specific words:

# Clean the corpus by removing punctuation, numbers, and white spaces

bd\_clean <- tm\_map(bd\_corpus, removePunctuation)

bd\_clean <- tm\_map(bd\_clean, tolower)

bd\_clean <- tm\_map(bd\_clean, removeNumbers)

# Remove stopwords and specific words

bd\_clean <- tm\_map(bd\_clean, removeWords, stopwords("english"))

bd\_clean <- tm\_map(bd\_clean, stripWhitespace)

bd\_clean <- tm\_map(bd\_clean, removeWords, c("Minnesota", "mn"))

**Exercise 3**

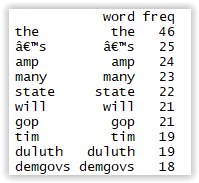
Extend exercise 2. This time retrieve the tweets posted from a geographical area. You can retrieve tweets posted from a given location using the **geocode** parameter specified with the template ‘latitude,longitude,radius’ in twitteR package, for example, “37.781157,-122.398720,1mi”.

The following word cloud was created for the “Minnesota” keyword from tweets posted within 80

miles from the New York City.

**Exercise 4**

Create word frequencies in your corpus. Use the head() to find top 10 words.



HINT: You will need to create a term-document matrix (a table that contains the frequency of words).

# Build a term document matrix

bd\_clean2 <- TermDocumentMatrix(bd\_clean)

bd\_clean2 <- as.matrix(bd\_clean2)

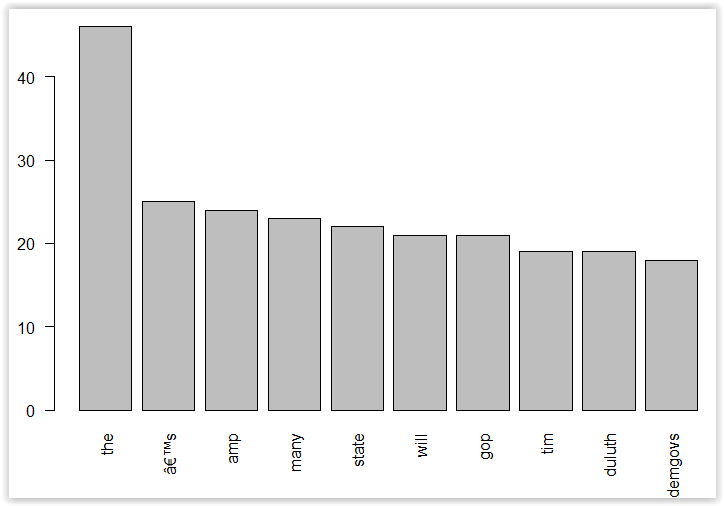
bd\_clean2 <- sort(rowSums(bd\_clean2),decreasing=TRUE)

bd\_clean2 <- data.frame(word = names(bd\_clean2),freq=bd\_clean2)

head (bd\_clean2, 10)

**Exercise 5**

Create a bar chart of 10 most frequent words in the corpus.



HINT: You can use the barplot().

barplot(bd\_clean2[1:10,]$freq, las = 2, names.arg = bd\_clean2[1:10,]$word)