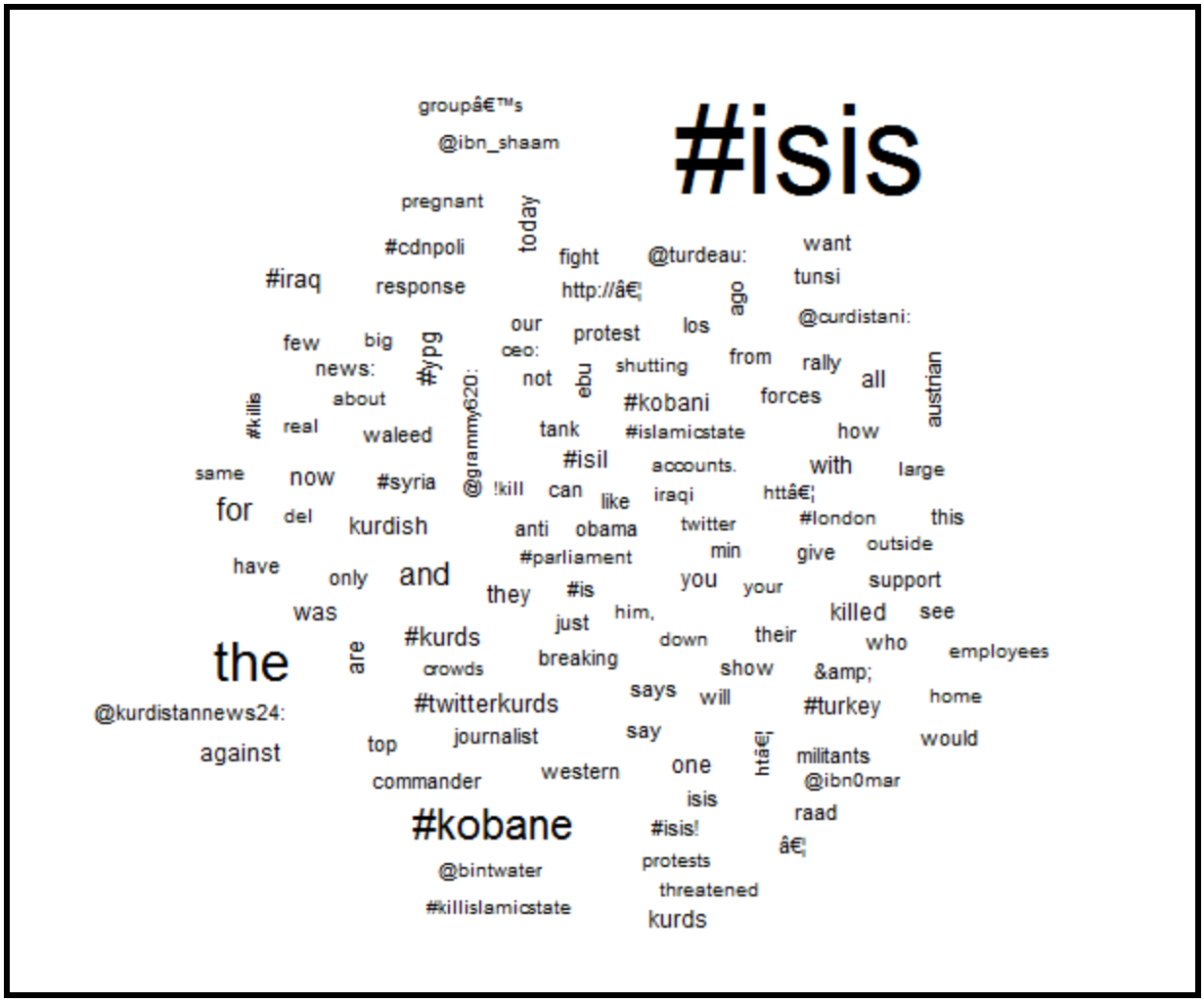
CleaningDataInR

Exercise 3: Cleaning Tweets

1. Open up Rstudio program and ex3.R

2. This exercise requires 2 external library
    - install.packages('tm')
        - text mining package
    - install.packages('wordcloud')
        - to visualize the words
3. Load the libraries
4. Read the tweets
    - ***df <- read.csv("C:\\Users\\benjit\\Google Drive\\CleaningDataWithR\\ex3\\tweets.csv",stringsAsFactors = FALSE)***
5. Read in the tweets as a corpus
    - ***tweet_corpus <- Corpus(VectorSource(df[,1]))***
        - VectorSource will create a list of tweets as a list of document
        - Corpus is a collection of documents
6. Create the document matrix
    - ***tweet_dtm <- TermDocumentMatrix(tweet_corpus)***
        - This will convert the corpus into a matrix
        - ***inspect(tweet_dtm)*** to see the matrix which is just a frequency of words
7. We want to visualize the word cloud
    - Convert back to matrix
        - ***tweet_m<- as.matrix(tweet_dtm)***
    - Sort the words by decreasing order
        - ***tweet_sort<- sort(rowSums(tweet_m), decreasing=TRUE)***
    - Get the data frame consisting of words and its frequency
        - ***tweets_df <- data.frame(word=names(tweet_sort), freq=tweet_sort)***
    - Show the word cloud
        - ***wordcloud(tweets_df$word, tweets_df$freq, min.freq=3)***

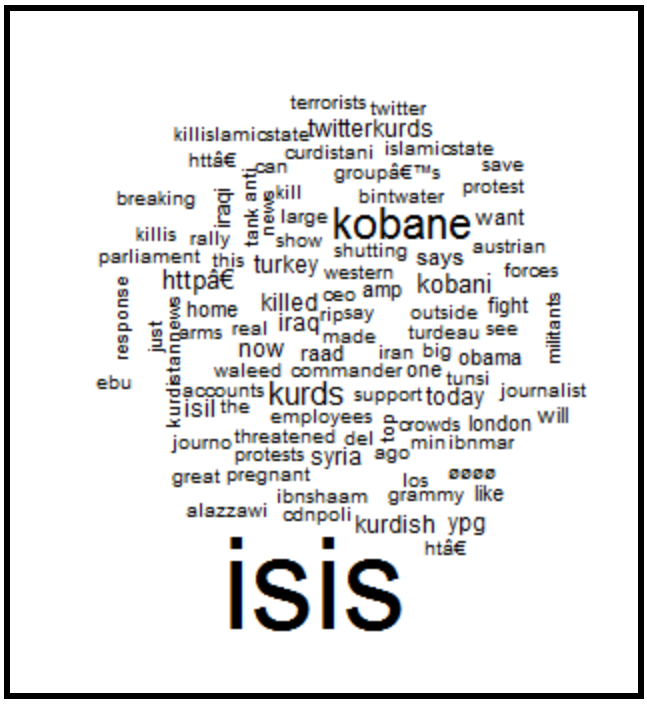


8. We need to remove some of the commons words from the list of tweets to visualize them more correctly. We shall use tm_map to do some transformation
    - Remove punctuations (hashtags, commas, question mark)
        - ***tweet_corpus<- tm_map(tweet_corpus, removePunctuation)***
    - Remove dictionary words
        - ***tweet_corpus<- tm_map(tweet_corpus, removeWords, stopwords('english'))***

```
> stopwords('english')
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"
 [10] "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"       "their"
 [28] "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"
```

    - Remove numbers
        - ***tweet_corpus<- tm_map(tweet_corpus, removeNumbers)***

9. Once u have done transforming the corpus just run Step 6 to 7 again
    - Much Cleaner !



10. You can read up on how to stem words after doing some basic cleaning of your tweets
    - http://www.rdatamining.com/examples/text-mining