Data Analytics
Assignment -7

Name : R Siva Girish SRN : PES1201700159

Dataset : Hilary Clinton and Donald Trump Tweets

The dataset is based on tweets from the 2016 US Presidential Elections between the two majority party candidates Donald Trump and Hillary Clinton.The dataset provides around 3000 tweets each from Donald Trump and Hillary Clinton.Our Goal is to build a word cloud from the given data for both Trump and Hillary.

	id	handle	text	is_retweet	original_author	time	in_reply_to_screen_name	in_reply_to_status_id	in_reply_to_user_id	is_quote_status	lang	retweet_count
1	7.809256e+17	HillaryClinton	The question in this election: Who can put the plans into act...	False		2016-09-28T00:22:34		NA	NA	False	en	218
2	7.809162e+17	HillaryClinton	Last night, Donald Trump said not paying taxes was "smart."...	True	timkaine	2016-09-27T23:45:00		NA	NA	False	en	2445
3	7.809116e+17	HillaryClinton	Couldn't be more proud of @HillaryClinton. Her vision and ...	True	POTUS	2016-09-27T23:26:40		NA	NA	False	en	7834
4	7.809070e+17	HillaryClinton	If we stand together, there's nothing we can't do. Make sur...	False		2016-09-27T23:08:41		NA	NA	False	en	916
5	7.808974e+17	HillaryClinton	Both candidates were asked about how they'd confront raci...	False		2016-09-27T22:30:27		NA	NA	False	en	859
6	7.808931e+17	realDonaldTrump	Join me for a 3pm rally - tomorrow at the Mid-America Cent...	False		2016-09-27T22:13:24		NA	NA	False	en	2181
7	7.808836e+17	HillaryClinton	This election is too important to sit out. Go to https://t.co/tT...	False		2016-09-27T21:35:28		NA	NA	False	en	1303
8	7.808811e+17	HillaryClinton	When Donald Trump goes low...register to vote: https://t.co/...	False		2016-09-27T21:25:31		NA	NA	False	en	1833
9	7.808768e+17	realDonaldTrump	Once again, we will have a government of, by and for the pe...	False		2016-09-27T21:08:22		NA	NA	False	en	4132
10	7.808747e+17	HillaryClinton	3) Has Trump offered a single proposal to reduce the frictio...	True	mcuban	2016-09-27T21:00:13		NA	NA	False	en	1087

Data Preprocessing

- The dataset in its native form cannot be used to create a word cloud.
- Due to the fact that there are a lot of stopwords,punctuations,whitespaces etc that in general cannot be avoided.

- Introducing these words in our word cloud would defeat the purpose and make it meaningless.
- Therefore we Preprocess all the tweets.
- Steps in Preprocessing -:
  - Remove &amp,&gt which appear a lot in the tweets.

```
unique_words <- str_split(tweet.text, "(&amp;|&gt;|\\s)")[[1]]
```

  - Remove all URLS.(Urls would not be useful)

```
unique_words <- grep("^http", unique_words , value = TRUE, invert = TRUE)
```

  - Remove all punctuation except ‘,# and @ since we are on twitter these punctuations are important.(eg : #2019).

```
tweet.words<-unique_words%>%gsub(pattern="[^[:alnum:][:space:]'#@]",replacement="")
```

  - Remove all trailing whitespaces

```
tweet.words <- str_trim(tweet.words)
```

  - Remove all empty strings.

```
tweet.words <- tweet.words[-which(tweet.words == "")]
```

  - Remove stopwords (Regex of Stopwords)

```
stopwords.regex <-
"^(a|about|above|after|again|against|all|am|an|and|any|are|aren't|as|at|be|because|been|before|being|below|between|both|but|by|can't|cannot|could|couldn't|did|didn't|do|does|doesn't|doing|don't|down|during|each|few|for|from|further|had|hadn't|has|hasn't|have|haven't|having|he|he'd|he'll|he's|her|here|here's|hers|herself|him|himself|his|how|how's|i|i'd|i'll|i'm|i've|if|in|into|is|isn't|it|it's|its|itself|let's|me|more|most|mustn't|my|myself|no|nor|not|of|off|on|once|only|or|other|ought|our|ours|ourselves|out|over|own|same|shan't|she|she'd|she'll|she's|should|shouldn't|so|some|such|than|that|that's|the|their|theirs|them|themselves|then|there|there's|these|they|they'd|they'll|they're|they've|this|those|through|to|too|under|until|up|very|was|wasn't|we|we'd|we'll|we're|we've|were|weren't|what|what's|when|when's|where|where's|which|while|who|who's|whom|why|why's|with|won't|would|wouldn't|you|you'd|you'll|you're|you've|your|yours|yourself|yourselves)$"
tweet.words <- grep(stopwords.regex, tweet.words, value = TRUE, ignore.case = TRUE, invert = TRUE)
```

Word Cloud

- Split all the tweets according to Clinton and Trump based on their handles.

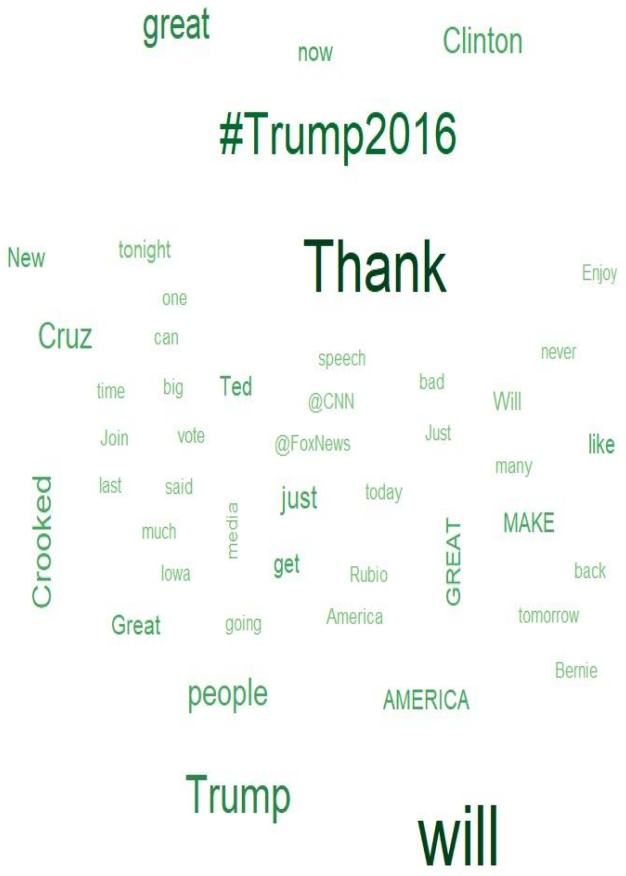
```
trump <- filter(tweets, handle == "realDonaldTrump")
clinton <- filter(tweets, handle == "HillaryClinton")
```

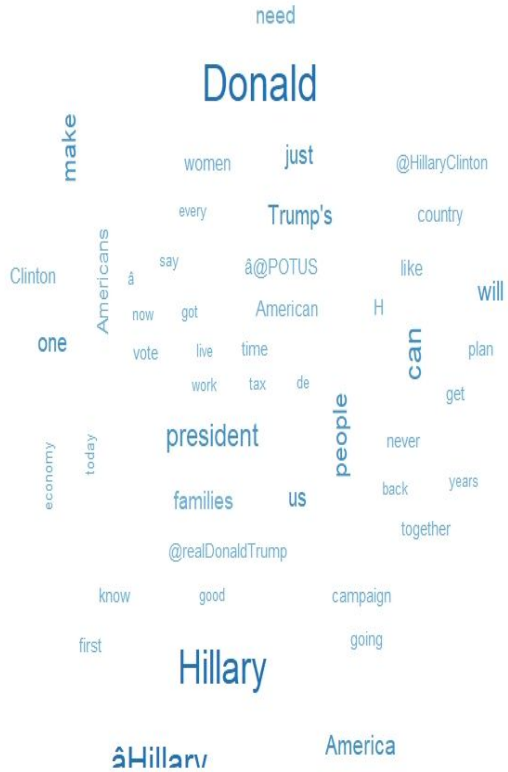
- Green- Donald Trump
  Blue - Hilary Clinton
- Design the word cloud and use a huge enough plotting device to plot both the word clouds next to each other.\

```
dev.new(width = 1000, height = 1000, unit = "px")
par(mfrow = c(1,2))
pal <- brewer.pal(n = 9, name = "Greens")
pal <- pal[-(1:4)]
wordcloud(trump.words$tweet.words, trump.words$n,
      scale = c(4, .5), max.words = 50, colors = pal)
pal <- brewer.pal(n = 9, name = "Blues")
pal <- pal[-(1:4)]
wordcloud(clinton.words$tweet.words, clinton.words$n,
      scale = c(4, .5), max.words = 50, colors = pal)
```

- The above snippet of code when executed will generate the word cloud for both Donald Trump tweets as well as Hillary Clinton tweets side by side on a plotting device of appropriate dimensions.

- Based on these word clouds we can easily determine some of the most common words used by the two candidates during the election.





Inferences

- From the word cloud we notice that some of the most commonly used words by trump were Trump,willm and Thank.
- The most common words used by Hillary Clinton were Donald,President and Hillary.