Code and Output

```r
##Load packages for Natural Language Processing
library('tm')
library('SnowballC')
library('wordcloud')
library('textstem')

##Load packages for visualization
library('RColorBrewer')
library('igraph')
library ('shiny')
```

```r
#set working directory
setwd('C:/Users/Sushant/Desktop/Data_Science_R/Constitution_Comparison/consts')

#Create a list of some words that you want to remove
other_words = c("article", "subarticle")

#Get the document and clean it
constitution = read.csv('Constitutions.csv', stringsAsFactors = FALSE)
corpus = Corpus(VectorSource(constitution$Constitution))
corpus = tm_map(corpus, stripWhitespace)
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, removeNumbers)
corpus = tm_map (corpus, content_transformer(tolower))
corpus = tm_map(corpus, removeWords, stopwords('en'))
corpus = tm_map(corpus, removeWords, other_words)
```

```r
#Lemmatize the corpus and remove the name of the country
corpus = tm_map(corpus, lemmatize_strings)
corpus = tm_map(corpus, removeWords, as.array(constitution$Country))
```

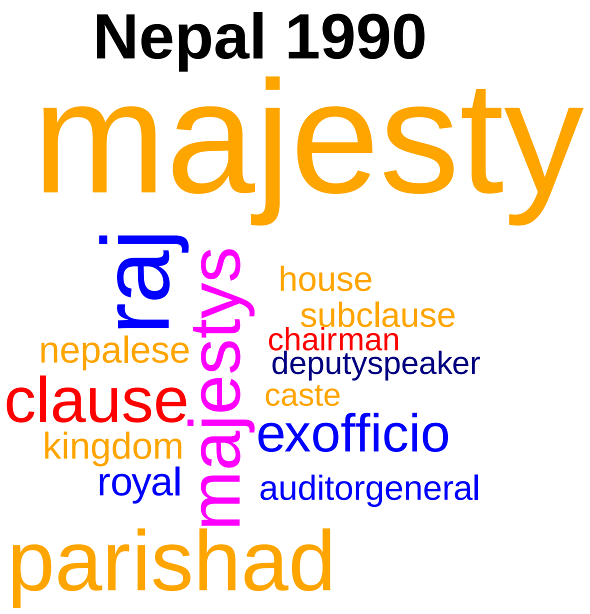
```r
#Create a document term matrix and thenconver to regular R matrix
term_mat = DocumentTermMatrix(corpus)
R_mat = as.matrix(term_mat)
#Attach country names as rowname of the matrix and compute the frequency of each words
rownames(R_mat)=constitution$Country
corp_freq = colSums(R_mat)
```

```r
#create wordcloud
cols = c('red', 'blue', 'orange', 'navy',   '#006400', 'magenta')
dev.new(width = 10000, height = 10000, unit = "px")
wordcloud(colnames(R_mat), R_mat['nepal old',],  max.words = 15, scale=c(5,0.5), colors = cols, random.c
title(main = "Nepal 1990", cex.main = 2)
wordcloud(colnames(R_mat), R_mat['nepal',],  max.words = 15, scale=c(4,0.5), colors = cols, random.colo
title(main = "Nepal 2015", cex.main = 2)
wordcloud(colnames(R_mat), R_mat['bhutan',],  max.words = 15, scale=c(4,0.5), colors = cols, random.colo
title(main = "Bhutan 2008", cex.main = 2)
```

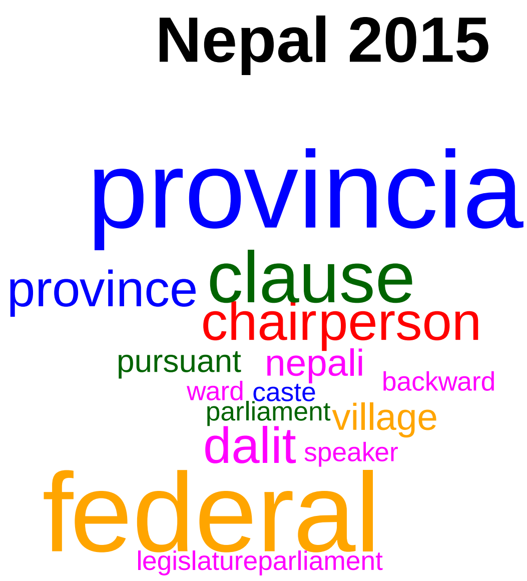
Above, we can see that auxiliary verb, ‘shall’ is the most common word in all three constitutions. However, such words are common in all litigation documents and do not capture the unique aspects of any country’s constitution. Therefore we can reduce the weight of these words by using Inverse Document Frequency, which normalizes the word frequency within a document based on the word frequency within the entire corpus.

```
#Weight the matrix
R_mat_weighted = as.matrix(weightTfIdf(term_mat, normalize = FALSE))
rownames(R_mat_weighted) = constitution$Country
```

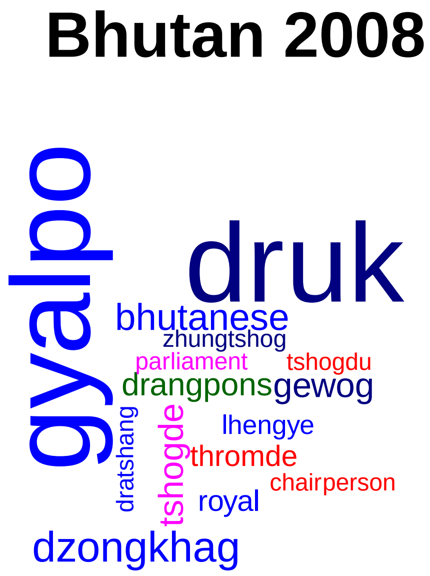
```
#create wordcloud
wordcloud(colnames(R_mat_weighted), R_mat_weighted['nepal old',],  max.words = 15, scale=c(5,0.5), colo
title(main = "Nepal 1990", cex.main = 2)
```



```
wordcloud(colnames(R_mat_weighted), R_mat_weighted['nepal',],  max.words = 15, scale=c(3.5,0.5), colors
title(main = "Nepal 2015", cex.main = 2)
```



```
wordcloud(colnames(R_mat_weighted), R_mat_weighted['bhutan',],  max.words = 15, scale=c(3.5,0.5), color
title(main = "Bhutan 2008", cex.main = 2)
```


Bhutan 2008

In the above word cloud, we can see that auxiliary verbs such as 'shall' and 'may' are absent. Now, before computing K-means clusters of the documents, we normalize the document using L2 normalization. This is a within-document normalization process to ensure that the word frequencies of longer and shorter documents are comparable.

```
#Normalize by length of each document, so that frequency is unaffected by length of preamble
R_mat_normalized = R_mat_weighted/sqrt(rowSums(R_mat_weighted^2))

K_means = kmeans(R_mat_normalized, centers = 10)

for (c in 1:10) {
  cat("Cluster ", c, ":\n")
  print(constitution$Country[K_means$cluster == c])
  cat("\n\n")
}
```

```
## Cluster  1 :
##  [1] "belgium"
##  [2] "benin"
##  [3] "burundi"
##  [4] "cameroon"
##  [5] "congo (democratic republic of the)"
##  [6] "cÃ´te d'ivoire"
##  [7] "djibouti"
##  [8] "equatorial guinea"
##  [9] "gabon"
## [10] "guinea"
## [11] "madagascar"
## [12] "mali"
## [13] "mauritania"
## [14] "niger"
## [15] "rwanda"
## [16] "senegal"
## [17] "togo"
##
##
## Cluster  2 :
##  [1] "austria"              "ethiopia"             "germany"
##  [4] "iraq"                 "malaysia"             "mexico"
##  [7] "nepal"                "nigeria"              "pakistan"
## [10] "somalia"              "switzerland"          "united arab emirates"
## [13] "yemen"
##
##
## Cluster  3 :
##  [1] "argentina"                        "chile"
##  [3] "china (peopleâ\200\231s republic of)"   "colombia"
##  [5] "dominican republic"               "guatemala"
##  [7] "honduras"                         "libya"
##  [9] "micronesia (federated states of)" "paraguay"
## [11] "peru"                             "united states of america"
##
##
## Cluster  4 :
##  [1] "albania"
##  [2] "algeria"
##  [3] "angola"
##  [4] "armenia"
##  [5] "belarus"
##  [6] "bolivia (plurinational state of)"
##  [7] "bulgaria"
##  [8] "cape verde"
##  [9] "costa rica"
## [10] "cuba"
## [11] "cuba old"
## [12] "cyprus"
## [13] "czech republic"
## [14] "ecuador"
## [15] "egypt"
## [16] "el salvador"
## [17] "france"
## [18] "guinea-bissau"
## [19] "haiti"
## [20] "hungary"
## [21] "italy"
## [22] "kazakhstan"
## [23] "korea (republic of)"
## [24] "kosovo"
## [25] "lebanon"
## [26] "libya"
## [27] "mozambique"
## [28] "nicaragua"
## [29] "panama"
## [30] "portugal"
## [31] "serbia"
## [32] "slovenia"
## [33] "suriname"
## [34] "syrian arab republic"
## [35] "syrian arab republic"
## [36] "tunisia"
## [37] "turkey"
## [38] "uruguay"
## [39] "venezuela (bolivarian republic of)"
## [40] "yemen old"
##
##
## Cluster  5 :
##  [1] "bangladesh"          "botswana"            "dominica"
##  [4] "fiji"                "finland"             "georgia"
##  [7] "ghana"               "greece"              "guyana"
## [10] "malawi"              "malta"               "mauritius"
## [13] "montenegro"          "nauru"               "sierra leone"
## [16] "south africa"        "sri lanka"           "trinidad and tobago"
## [19] "vanuatu"             "zambia"              "zimbabwe"
##
##
## Cluster  6 :
##  [1] "antigua and barbuda"              "australia"
##  [3] "bahamas (the)"                    "barbados"
##  [5] "belize"                           "grenada"
##  [7] "jamaica"                          "saint kitts and nevis"
##  [9] "saint lucia"                      "saint vincent and the grenadines"
## [11] "solomon islands"
##
##
## Cluster  7 :
## [1] "azerbaijan"    "liechtenstein" "maldives"      "monaco"
## [5] "oman"          "qatar"         "uzbekistan"
##
##
## Cluster  8 :
##  [1] "afghanistan"
##  [2] "andorra"
##  [3] "bhutan"
##  [4] "bosnia and herzegovina"
##  [5] "brunei darussalam"
##  [6] "burkina faso"
##  [7] "canada"
##  [8] "comoros"
##  [9] "croatia"
## [10] "denmark"
## [11] "estonia"
## [12] "gambia (the)"
## [13] "iceland"
## [14] "iceland old"
## [15] "indonesia"
## [16] "iran (islamic republic of)"
## [17] "israel"
## [18] "japan"
## [19] "kiribati"
## [20] "korea (democratic people's republic of)"
## [21] "kuwait"
## [22] "kyrgyzstan"
## [23] "lao people's democratic republic"
## [24] "latvia"
## [25] "lithuania"
## [26] "luxembourg"
## [27] "macedonia (republic of)"
## [28] "marshall islands"
## [29] "moldova (republic of)"
## [30] "mongolia"
## [31] "myanmar"
## [32] "norway"
## [33] "palau"
## [34] "palestine"
## [35] "papua new guinea"
## [36] "philippines"
## [37] "poland"
## [38] "sao tome and principe"
## [39] "slovakia"
## [40] "sweden"
## [41] "taiwan (republic of china)"
## [42] "tajikistan"
## [43] "tanzania (united republic of)"
## [44] "timor-leste"
## [45] "turkmenistan"
## [46] "tuvalu"
## [47] "ukraine"
## [48] "viet nam"
## 
## 
## Cluster  9 :
##  [1] "bahrain"      "cambodia"     "jordan"       "lesotho"     
##  [5] "morocco"      "netherlands"  "saudi arabia" "swaziland"   
##  [9] "thailand"     "tonga"       
## 
## 
## Cluster  10 :
##  [1] "eritrea"     "india"       "kenya"       "liberia"     "namibia"    
##  [6] "nepal old"   "samoa"       "seychelles"  "south sudan" "sudan"      
## [11] "uganda"
```

Some aspects of the above results are interesting, while others are less so. First and foremost, the two constitutions of Nepal lie in different clusters, which is a good sign for us to continue with further analysis. Additionally, Cluster 1 comprises of Caribbean nations, except for Australia. The inclusion of Australia in the cluster can be justified by the fact that it is a British colony and still constitutionally accepts British Monarchy. Additionally, Cluster 2 is a cluster of Monarchies–absolute or otherwise. Cluster 5 is a group of republics with a dominant Muslim population except for Oman, which is a Monarchy. Finally, Cluster 3 is Belgium and African countries with Belgian or French influence.

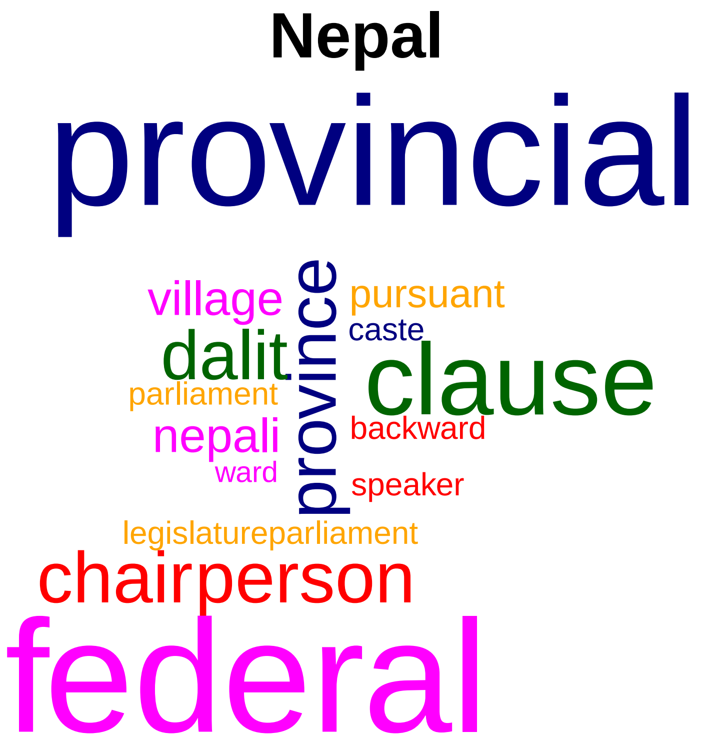
While there are some faint patterns in other clusters as well, they can be considered insignificant. The method of clustering also seems to be unstable, as its result depends on the initial choice of centres, which takes place randomly. Furthermore, cluster analysis doesn't give a measure of difference or similarity. Therefore, in the next step I measure cosine similarity of the constitution, which mathematically, is the cosine of the angle between the two vectors.

```r
#Write a cosine similarity function
#Nothe that the function computes similarity between one vector and one matrix
cosine = function(vec_a, mat_b) {
dot_product = apply(vec_a * t(mat_b), 2, sum)
mag_product = sqrt(sum(vec_a^2)) * sqrt(apply(mat_b^2, 1, sum))
return(dot_product / mag_product)
}
```

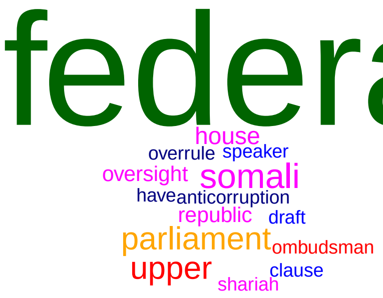
```r
#Compute the cosine similarity between new constitution of Nepal and rest of the countries
# Get Nepal's index
nep_row = (rownames(R_mat_weighted) == "nepal")

# Get similarity with US using `term_mat_tfidf`
nep_similarity = cosine(R_mat_weighted[nep_row, ], R_mat_weighted[!nep_row, ])
sort(nep_similarity, decreasing = TRUE)[1:35]
```

```
##                            somalia                           pakistan
##                          0.5576909                          0.5419547
##                             mexico                       south africa
##                          0.4725055                          0.4328148
##                              yemen                           ethiopia
##                          0.4117622                          0.3784356
##                            austria                          sri lanka
##                          0.3597555                          0.3356079
##                            nigeria                            germany
##                          0.3284383                          0.3137844
##                             zambia                              ghana
##                          0.2489883                          0.2475293
##                        switzerland                          argentina
##                          0.2474898                          0.2390377
## congo (democratic republic of the)                           malaysia
##                          0.2292146                          0.2227758
##                         bangladesh                               iraq
##                          0.2209522                          0.1990639
##                               cuba                              india
##                          0.1971446                          0.1932926
##                              samoa                             guyana
##                          0.1924133                          0.1782770
##               united arab emirates                             serbia
##                          0.1752263                          0.1748455
##                            vanuatu                              nauru
##                          0.1680620                          0.1642954
##                        south sudan                         seychelles
##                          0.1601544                          0.1523413
##                          nepal old                       saudi arabia
##                          0.1522859                          0.1494171
##                              kenya                             uganda
##                          0.1326147                          0.1268564
##                        netherlands                               fiji
##                          0.1212769                          0.1185905
##                             ecuador
##                           0.1112882
```

```
wordcloud(colnames(R_mat_weighted), R_mat_weighted['nepal',],  max.words = 15, scale=c(5,0.5), colors =
title(main = "Nepal", cex.main = 2)
```



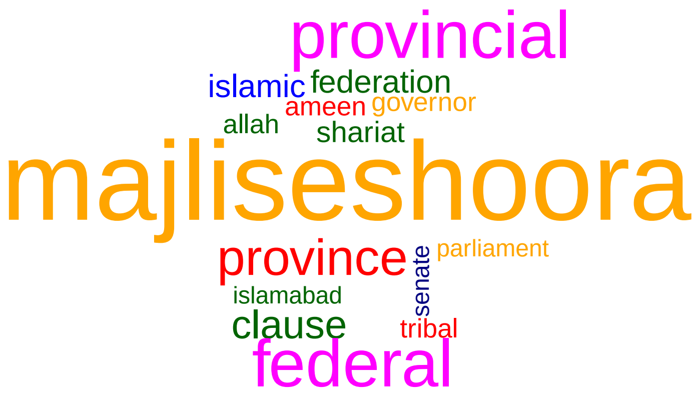
```
wordcloud(colnames(R_mat_weighted), R_mat_weighted['somalia',],  max.words = 15, scale=c(5, 0.5), color
title(main = "Somalia", cex.main = 2)
```



Somalia

```
wordcloud(colnames(R_mat_weighted), R_mat_weighted['pakistan',],  max.words = 15, scale=c(3.5,0.5), col
title(main = "Pakistan", cex.main = 2)
```

Pakistan



Finally, we see that the old constitution of Nepal ranks 28th in the similarity with Nepal's 2015 constitution. Interestingly, Somalia's constitution is the most similar to Nepal's. Initially, it can be difficult to reconcile given no cultural, political, or geographic connection between Nepal and Somalia. However, if we see the word cloud of the two countries that have the most similar constitutions to Nepal's, it is evident that all three nations focus on the structure of governance and administrative system in the constitution. Especially given the fact that the Somali constitution was adopted in 2012, which states the provision for the Federal Parliamentary Republic, the similarity with Nepal's constitution is plausible.