Word Cloud

Kaushal Khatiwada

2024-04-23

```
library(rvest)
library(tm)
```

```
## Loading required package: NLP
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
urls<- "https://thehimalayantimes.com/opinion/navigating-nepals-digital-frontier-understanding-cybersecu
```

```
texts <- read_html(urls) %>%
  html_nodes("p") %>%   #Get <p> HTML tag
  html_text()
```

Combine all the text into a single corpus

```
myCorpus <- Corpus(VectorSource(texts))
```

```
inspect(myCorpus[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] The role of AI
## [2] The transformative advent of artificial intelligence (AI) in the digital space indicates a new e
## [3] AI's ability to meticulously analyze massive amounts of data in real-time can transform our curr
## [4] Collaboration between AI technologies and human expertise is equally important. Thus, prioritizi
```

Text Cleaning

1) Convert to a lowercase

```
text <- tm_map(myCorpus, content_transformer(tolower))
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] the role of ai
## [2] the transformative advent of artificial intelligence (ai) in the digital space indicates a new e
## [3] ai's ability to meticulously analyze massive amounts of data in real-time can transform our curr
## [4] collaboration between ai technologies and human expertise is equally important. thus, prioritizi
```

2) Remove Numbers

```
text <- tm_map(text, removeNumbers)
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] the role of ai
## [2] the transformative advent of artificial intelligence (ai) in the digital space indicates a new e
## [3] ai's ability to meticulously analyze massive amounts of data in real-time can transform our curr
## [4] collaboration between ai technologies and human expertise is equally important. thus, prioritizi
```

3) Remove Punctuation

```
text <- tm_map(text, removePunctuation)
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] the role of ai
## [2] the transformative advent of artificial intelligence ai in the digital space indicates a new era
## [3] ais ability to meticulously analyze massive amounts of data in realtime can transform our curren
## [4] collaboration between ai technologies and human expertise is equally important thus prioritizing
```

4) Remove URL

```
removeURL <- function(X) gsub("http[^[:space:]]*","",X)
text <- tm_map(text,removeURL)
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] the role of ai
## [2] the transformative advent of artificial intelligence ai in the digital space indicates a new era
## [3] ais ability to meticulously analyze massive amounts of data in realtime can transform our curren
## [4] collaboration between ai technologies and human expertise is equally important thus prioritizing
```

5) Remove Stopwords.

```
text <- tm_map(text, removeWords, stopwords("english"))
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1]  role  ai
## [2]  transformative advent  artificial intelligence ai   digital space indicates  new era  innovatio
## [3] ais ability  meticulously analyze massive amounts  data  realtime can transform  current approac
## [4] collaboration  ai technologies  human expertise  equally important thus prioritizing research  d
```

6) Remove Whitespaces.

```
text <- tm_map(text, stripWhitespace)
inspect(text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1]  role ai
## [2]  transformative advent artificial intelligence ai digital space indicates new era innovation sec
## [3] ais ability meticulously analyze massive amounts data realtime can transform current approach cy
## [4] collaboration ai technologies human expertise equally important thus prioritizing research devel
```

Stemming and lemmatization

Stemming is a text normalization technique used in natural language processing (NLP) that reduces words to their root or base form to enhance language understanding. Example {remove suffixes} : stem of these three words, “connections”, “connected”, “connects”, is “connect”.

Lemmatization analyzes the contexts of the sentences. Example: “Saw” would return as “see” or “saw” depending on the context of the word.

```
# Backup the text just in case
backup_text <- text
```

```
backup_text <- tm_map(backup_text, stemDocument)
inspect(backup_text[17:20])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 4
##
## [1] role ai
## [2] transform advent artifici intellig ai digit space indic new era innov secur often quot gamechang
## [3] ai abil meticul analyz massiv amount data realtim can transform current approach cyber secur can
## [4] collabor ai technolog human expertis equal import thus priorit research develop ai digit world i
```

It might not be good idea to use stemming in some cases because “artifici”,“intellig” is not the real words. So ww will not be using stemmed words.

Term-Document matrix

Text-Document Matrix represent the text data in the form of a matrix where row correspond to the term in the document and columns correspond to the document in the corpus and cells correspond to the weights of the term.

```
mytdm <- TermDocumentMatrix(text,control = list(wordLengths = c(1,Inf)))
mymatrix <- as.matrix(mytdm)
inspect(mytdm)
```

```
## <<TermDocumentMatrix (terms: 348, documents: 23)>>
## Non-/sparse entries: 484/7520
## Sparsity           : 94%
## Maximal term length: 15
## Weighting          : term frequency (tf)
## Sample             :
##            Docs
## Terms       11 14 16 19 20 3 5 7 8 9
##   can        2  0  2  2  0 0 0 0 1 0
##   cyber      1  0  0  1  1 0 0 0 1 0
##   data       1  0  0  1  0 2 0 1 0 0
##   digital    1  0  3  0  3 2 3 0 1 0
##   framework  0  0  0  0  0 0 1 0 1 1
##   security   0  1  0  1  0 1 1 0 2 1
##   smart      6  0  0  0  0 0 0 0 0 0
##   space      0  0  0  0  1 1 1 0 0 0
##   threats    1  1  0  1  0 0 0 0 1 0
##   world      0  0  1  0  2 1 0 0 0 0
```

Term Frequency represent the frequency of terms in the document.

```
freq.terms <- findFreqTerms(mytdm, lowfreq = 2)
freq.terms
```

```
##  [1] "digital"       "internet"      "nepal"         "nepali"
##  [5] "percent"       "population"    "space"         "authorities"
##  [9] "country"       "data"          "every"         "government"
## [13] "governments"   "important"     "including"     "individuals"
## [17] "information"   "life"          "organizations" "sector"
## [21] "security"      "world"         "building"      "framework"
## [25] "approach"      "build"         "business"      "cybersecurity"
## [29] "robust"        "significant"   "time"          "towards"
## [33] "along"         "assets"        "can"           "plans"
## [37] "policies"      "set"           "systems"       "technologies"
## [41] "threats"       "april"         "intelligence"  "microsoft"
## [45] "threat"        "citizens"      "cyber"         "economic"
## [49] "ensure"        "growth"        "investments"   "nation"
## [53] "system"        "trust"         "also"          "collaboration"
## [57] "detection"     "different"     "training"      "leadership"
## [61] "become"        "direction"     "handling"      "often"
## [65] "practices"     "protect"       "smart"         "strategic"
## [69] "technology"    "zero"          "develop"       "employees"
## [73] "potential"     "investment"    "journey"       "multiyear"
```

```
## [77] "development"   "global"        "prosperity"    "ai"
## [81] "ais"
```

To search for term associated with “ai” with correlation coefficient of at least 0.2

```
findAssocs(mytdm, "ai", 0.2)
```

```
## $ai
##   complexities      dominance        equally        evident      expertise
##           0.90           0.90           0.90           0.90           0.90
##       frontier          human   instrumental       integral        nations
##           0.90           0.90           0.90           0.90           0.90
##       navigate   prioritizing       research           safe           thus
##           0.90           0.90           0.90           0.90           0.90
##    development          world      important   technologies  collaboration
##           0.79           0.64           0.60           0.60           0.60
##         become     prosperity          space    investments        digital
##           0.60           0.60           0.47           0.46           0.44
##          cyber          nepal      strategic           role         advent
##           0.41           0.37           0.37           0.25           0.25
##     artificial            era   everevolving   gamechanging         impact
##           0.25           0.25           0.25           0.25           0.25
##      indicates     innovation            new         quoted         stands
##           0.25           0.25           0.25           0.25           0.25
## transformative
##           0.25
```

```
freq <- sort(rowSums(mymatrix),decreasing = TRUE)
```

```
library(RColorBrewer)
```

Word Cloud

```
set.seed(123)
wordcloud(words = names(freq), freq = freq, min.freq = 2, random.order = FALSE, colors = brewer.pal(8,"
```


technology
practices
employees
different
economic
collaboration
leadership
become
important
investments
citizens
ais
assets
population
set
systems
organizations
investment
develop
also
training
trust
data
robust
detection
internet
time
nepali
nepal
threats
world
life
growth
cyber
can
space
significant
including
every
digital
country
journey
potential
global
towards
building
strategic
protect
microsoft
security
smart
april
governments
ensure
sector
system
individuals
framework
ai
policies
nation
approach
zero
information
intelligence
build
multiyear
along
cybersecurity
percent
often
business
government
authorities
prosperity
development
threat
plans
technologies
handling
direction