News Saturation DB Initial Exploration

## exploring database  
  
library(httr)  
library(jsonlite)  
library(lubridate)  
library(tidyverse)  
library(plyr)  
library(RODBC)  
library(RODBCext)  
library(ngram)  
library(stringr)  
library(tm)

get DB stuff

myServer <- odbcDriverConnect('driver={SQL Server};server=RIC007\\SEBISERVER;database=news;trusted\_connection=true')  
  
## git eet allllll outta thur  
  
queryGitEm <- "select \* from top\_news\_full"  
  
myDBframe <- sqlQuery(myServer,queryGitEm)

take a peek

glimpse(myDBframe)

## Observations: 6,779  
## Variables: 9  
## $ status <fctr> ok, ok, ok, ok, ok, ok, ok, ok, ok, ok, o...  
## $ source <fctr> al-jazeera-english, al-jazeera-english, a...  
## $ sortBy <fctr> top, top, top, top, top, top, top, top, t...  
## $ articlesauthor <fctr> , , , , , , , , , , , , , , Al Jazeera, A...  
## $ articlestitle <fctr> Iran's 2017 elections: All you need to kn...  
## $ articlesdescription <fctr> The May 19 vote will be the first stage, ...  
## $ articlesurl <fctr> http://www.aljazeera.com/indepth/interact...  
## $ articlesurlToImage <fctr> http://www.aljazeera.com/mritems/Images/2...  
## $ articlespublishedAt <fctr> list(), list(), 2017-05-18T13:53:00Z, 201...

most frequently used single words- no stop words removed

decontstructText <- function (vector){  
titleVector\_0<- as.vector(vector)   
titleVector<- str\_split(titleVector\_0," ")  
titleVector\_2<- unlist(titleVector)  
titleVector<- data.frame(table(titleVector\_2))  
  
TV <- split(titleVector, titleVector$Freq > 100)  
TV\_2 <- as.data.frame(TV[2])  
return (TV\_2)  
  
}  
  
  
  
humbug1 <- decontstructText(myDBframe$articlesdescription)  
titleVector\_0<- as.vector(myDBframe$articlesdescription)   
titleVector<- str\_split(titleVector\_0," ")  
titleVector\_2<- unlist(titleVector)  
titleVector<- data.frame(table(titleVector\_2))  
  
TV <- split(titleVector, titleVector$Freq > 100)  
TV\_2 <- as.data.frame(TV[2])  
  
  
glimpse(as.data.frame(humbug1[with(humbug1, order(-TRUE.Freq)), ]))

## Observations: 150  
## Variables: 2  
## $ TRUE.titleVector\_2 <fctr> the, to, a, of, in, and, on, The, for, tha...  
## $ TRUE.Freq <int> 7782, 4345, 4286, 3745, 3298, 2764, 2060, 1...

now lets look at most common 3 length ngrams,

helloWorld <- concatenate(titleVector\_0)  
  
helloWorld<- preprocess(helloWorld, case = "lower", remove.punct = TRUE,  
remove.numbers = FALSE, fix.spacing = TRUE)  
  
  
nope<- ngram(helloWorld, n = 3)  
  
  
  
phraseBaby <- get.phrasetable(nope)  
  
glimpse(as.data.frame(phraseBaby))

## Observations: 124,005  
## Variables: 3  
## $ ngrams <chr> "president donald trump ", "ap <U+0097> the ", "president dona...  
## $ freq <int> 407, 128, 120, 117, 115, 110, 109, 105, 103, 97, 77, 74...  
## $ prop <dbl> 0.0024875774, 0.0007823339, 0.0007334381, 0.0007151021,...

check out a random babble

babbler <-babble(nope, genlen=15, seed=1234)  
babbler

## [1] "not a huge fan of the american fabric is historical revisionism with a political agenda "