R Notebook

Code

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
library(stringr)
library(httr)
library(readr)
library(rvest)
library(ggplot2)
library(dplyr)
library(magrittr)
library(syuzhet)
library(tm)
library(wordcloud)
library(RColorBrewer)
library(tidytext)
library(tidyr)
library(topicmodels)
library(xml2)
library(knitr)
set_config(config(ssl_verifypeer=0L))
```

Step 1

Hide

```
setwd("~/Fall 2019/Oidd 245/Homework/News Analytics")
cnbcdata = read.csv("NewsArticles.csv")

cnbccorp = VCorpus(VectorSource(cnbcdata$content))
cnbccorp = tm_map(cnbccorp, removePunctuation)
cnbccorp = tm_map(cnbccorp, removeNumbers)
cnbccorp = tm_map(cnbccorp, content_transformer(removeWords), stopwords("SMART"), lazy=TRUE)
cnbccorp = tm_map(cnbccorp, content_transformer(tolower), lazy=TRUE)
cnbccorp = tm_map(cnbccorp, content_transformer(stemDocument), lazy=TRUE)
cnbccorp = tm_map(cnbccorp, stripWhitespace)

dtmatrix = DocumentTermMatrix(cnbccorp)
dtmatrix = removeSparseTerms(dtmatrix , .995)
cnbcmatrix = as.matrix(dtmatrix)

traindata = cnbcmatrix[1:5000,]
topicmodels = LDA(traindata, k = 10, control = list(seed = 124))
```

Hide

```
topicresult = terms(topicmodels, 10)
topicresult
```

```
      Topic 1   Topic 2   Topic 3   Topic 4   Topic 5
 [1,] "report"  "oil"     "home"    "compani" "percent"
 [2,] "bank"    "price"   "sale"    "percent" "rate"
 [3,] "read"    "energi"  "year"    "share"   "year"
 [4,] "secur"   "read"    "price"   "billion" "month"
 [5,] "compani" "state"   "retail"  "year"    "read"
 [6,] "case"    "water"   "read"    "quarter" "increas"
 [7,] "inform"  "year"    "car"     "revenu"  "report"
 [8,] "court"   "gas"     "percent" "earn"    "job"
 [9,] "feder"   "product" "market"  "million" "growth"
[10,] "million" "weather" "hous"    "sale"    "week"
      Topic 6   Topic 7    Topic 8      Topic 9
 [1,] "compani" "market"   "state"      "percent"
 [2,] "busi"    "stock"    "obama"      "year"
 [3,] "read"    "percent"  "presid"     "financi"
 [4,] "year"    "fund"     "republican" "plan"
 [5,] "million" "year"     "tax"        "retir"
 [6,] "time"    "bank"     "read"       "pay"
 [7,] "ceo"     "rate"     "hous"       "loan"
 [8,] "make"    "investor" "senat"      "read"
 [9,] "game"    "fed"      "democrat"   "peopl"
[10,] "deal"    "invest"   "congress"   "money"
      Topic 10
 [1,] "apple"
 [2,] "read"
 [3,] "compani"
 [4,] "peopl"
 [5,] "app"
 [6,] "googl"
 [7,] "user"
 [8,] "ebola"
 [9,] "devic"
[10,] "servic"
```

Hide

```
eyeballedtopics = c("Legal", "Energy", "Homes", "Company Financials", "Macroeconomics", "Deals",
"Investing", "Politics", "Financing", "Trending")
```

Step 2

Hide

```
url = read_html("https://www.cnbc.com/us-news/")
cnbcurl = read_html("https://www.cnbc.com/us-news/")
cnbcurl = html_nodes(cnbcurl, ".Card-title")
cnbcurl = html_attr(cnbcurl, "href")

articletitle = html_nodes(url, ".Card-title")
articletitle = html_text(articletitle)

scrapedarticles = data.frame(title = articletitle, url = cnbcurl, stringsAsFactors = FALSE)

for (m in 1:nrow(scrapedarticles)) {

cnbclink = read_html(scrapedarticles$url[m],sep="")

scrapedarticles[m, "text"] = cnbclink %>% html_nodes(".group p") %>% html_text() %>% paste(colla
pse = '\n')


scrapedarticles[m, "text"] <- gsub("\n", " ", scrapedarticles[m, "text"])
scrapedarticles[m, "text"] <- gsub('\t', " ", scrapedarticles[m, "text"])
scrapedarticles[m, "text"] <- gsub('`\`', " ", scrapedarticles[m, "text"])
scrapedarticles[m, "text"] <- gsub('[[:punct:]]', " ", scrapedarticles[m, "text"])
scrapedarticles[m, "text"] <- gsub('[[:cntrl:]]', " ", scrapedarticles[m, "text"])

}

scrapedarticles[3,3]
```

```
[1] "Shares of French luxury goods makers  including LVMH and Hermes  dropped in European tradin
g on Tuesday after President Donald Trump s administration said it may place heavy tariffs on se
veral goods  LVMH – which includes the brands Louis Vuitton and Hennessy – and Hermes each fell
more than 2  following the announcement    The move by the U S  comes in retaliation to France s
implementation of a Digital Services Tax  Under the new tariffs  which may begin in late January
the U S  Trade Representative could add levies up to 100  on around  2 4 billion imports from Fr
ance   This would essentially double taxes on French companies and citizens with U S  income   C
owen said in a note to investors   These tariffs would be in addition to the WTO sanctioned Octo
ber tariffs stemming from the Airbus subsidies that target   25B in European goods  mainly Franc
e  Spain  Germany and UK    Shares of Christian Dior and Kering – the luxury group that owns bra
nds like Gucci  Yves Saint Laurent  Balenciaga – also each fell more than 2      Once again  key
French consumer agricultural brands have been targeted including cheese and champagne   Cowen sa
id  Verallia  the glass bottle maker for Dom Perignon and others  was also among the French stoc
ks falling  as shares slipped more than 1    You can read the full list here of proposed tariffs
on French imports  – CNBC s Gina Francolla contributed to this report "
```

Hide

```
cleantext = VCorpus(VectorSource(scrapedarticles$text))
cleantext = tm_map(cleantext, removePunctuation)
cleantext = tm_map(cleantext, removeNumbers)
cleantext = tm_map(cleantext, content_transformer(removeWords), stopwords("SMART"), lazy=TRUE)
cleantext = tm_map(cleantext, content_transformer(tolower), lazy=TRUE)
cleantext = tm_map(cleantext, content_transformer(stemDocument), lazy=TRUE)
cleantext = tm_map(cleantext, stripWhitespace)
```

Step 3

Hide

```
dic = Terms(dtmatrix)
dtmnews = DocumentTermMatrix(cleantext, control=list(dictionary = dic))
dtmnews = dtmnews[rowSums(as.matrix(dtmnews))!=0,]
topic_probabilities = posterior(topicmodels, dtmnews)

clustered = as.data.frame(topic_probabilities$topics)
colnames(clustered) = c(eyeballedtopics)

clustered$topic <- colnames(clustered)[max.col(clustered, ties.method="first")]
scrapedarticles$topic = clustered$topic

scrapedarticles$text = substr(scrapedarticles$text, start = 1, stop = 80)

kable(head(scrapedarticles,10))
```

title	url	text	topic
Here’s what police say you can do to to prevent packages from getting stolen from your porch	https://www.cnbc.com/2019/12/03/police-tips-to-prevent-package-theft-during-the-holidays.html	My local police department recently posted some holiday crime prevention tips to	Trending
Trump loses appeal to block banks from handing his financial records to Congress	https://www.cnbc.com/2019/12/03/trump-loses-appeal-to-block-deutsche-bank-capital-one-from-handing-his-financial-records-to-congress.html	A federal appeals court ruled Tuesday that Deutsche Bank and Capital One can han	Legal
French luxury stocks drop as Trump threatens heavy new tariffs	https://www.cnbc.com/2019/12/03/french-luxury-stocks-drop-as-trump-threatens-heavy-new-tariffs.html	Shares of French luxury goods makers including LVMH and Hermes dropped in Euro	Company Financials
Delaying trade deal until after 2020 election takes leverage away from China, Wilbur Ross says	https://www.cnbc.com/2019/12/03/wilbur-ross-says-china-trade-deal-after-2020-election-takes-away-leverage.html	Waiting until after the 2020 election to strike a China trade deal takes away so	Politics

title	url	text	topic
Google’s ‘Thanksgiving Four’ plan to file unfair labor practice charges with a federal agency	https://www.cnbc.com/2019/12/03/googles-thanksgiving-four-file-unfair-labor-practice-nlrb-charges.html	Four former Google employees who were fired days before Thanksgiving plan to fil	Legal
Here’s how ETF investors are positioned for the historically strongest month of the year	https://www.cnbc.com/2019/12/03/etfs-in-2019-how-investors-are-positioned-for-december.html	The year end rush has begun December which has been the strongest trading mont	Investing
Cyber Monday online sales hit record $9.4 billion, boosted by late-night spending spree, Adobe says	https://www.cnbc.com/2019/12/03/cyber-monday-online-sales-hit-record-9point4-billion-adobe-says.html	Cyber Monday shoppers spent a record 9 4 billion online up 19 7 from a year a	Homes
The other college debt crisis: Schools are going broke	https://www.cnbc.com/2019/12/03/the-other-college-debt-crisis-schools-are-going-broke.html	HIRAM Ohio — Small private liberal arts colleges — a staple of American academ	Financing
How much you should donate on Giving Tuesday (and other money-smart tips)	https://www.cnbc.com/2019/12/03/how-to-make-the-most-of-giving-tuesday.html	Charitable donations generally kick into high gear on Giving Tuesday a single	Financing
‘Obsession,’ ‘dangerous,’ ‘basement bunker’: GOP impeachment report rips Democrats’ inquiry	https://www.cnbc.com/2019/12/03/obsession-dangerous-basement-bunker-gop-impeachment-report-rips-democrats-inquiry.html	House Republicans argue in a 123 page minority report that Democrats have failed	Politics

Code