Quantedav1

Michael Pearson

02/20/2018

Quanteda work

This will create a corpus, clean it, and tokenize it using quanteda

```
## count the lines in the twitter, news, and blog files
newslines <- countLines("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
bloglines <- countLines("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
tweetlines <- countLines("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
## use that to read the files
tweet_us <- file("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/files
tweet_all <- readLines(tweet_us, n= tweetlines, warn = FALSE, encoding = "UTF=8", skipNul = TRUE)
close(tweet_us)
blog_us <- file("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/files/
blog_all <- readLines(blog_us, n= bloglines, warn = FALSE, encoding = "UTF=8", skipNul = TRUE)
close(blog_us)
news_us <- file("/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/files/
news_all <- readLines(news_us, n = newslines, warn = FALSE, encoding = "UTF=8", skipNul = TRUE)
close(news_us)
```

Sample 20% of the files to get a test sample corpus

```
set.seed(20180428)
samp_per <- 0.2
sam_twit <- tweet_all[sample(1:length(tweet_all),samp_per*length(tweet_all))]
write_lines(sam_twit, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/
sam_news <- news_all[sample(1:length(news_all),samp_per*length(news_all))]
write_lines(sam_news, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/
sam_blog <- blog_all[sample(1:length(blog_all),samp_per*length(blog_all))]
write_lines(sam_blog, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/
temp <- "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Project/files/samples/
samplename <- readtext(temp)
myCorpus <- corpus(samplename)
mytokens <- tokens(myCorpus)
mydfm <- dfm(myCorpus)
```

Now make the n-grams - with and without stems

```
## onegrams with stemming and stopwords
one_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbo
ns_one_gram <- tokens_remove(one_gram, stopwords("english"))
dfm_one_gram_stem_and_stop <- dfm(ns_one_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_one_gram_stem_and_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/C
one_gram_s_s <- sort(colSums(dfm_one_gram_stem_and_stop), decreasing = TRUE)
one_gram_s_s <- data.frame(one_gram_s_s)
one_gram_s_s <- setDT(one_gram_s_s, keep.rownames = TRUE)
write_csv(one_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
```

```
rm(dfm_one_gram_stem_and_stop)
rm(one_gram_s_s)
```

onegram with no stemming and with stopping

```
dfm_one_gram_nostem_and_stop <- dfm(ns_one_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_one_gram_nostem_and_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
one_gram_ns_s <- sort(colSums(dfm_one_gram_nostem_and_stop), decreasing = TRUE)
one_gram_ns_s <- data.frame(one_gram_ns_s)
one_gram_ns_s <- setDT(one_gram_ns_s, keep.rownames = TRUE)
write_csv(one_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(one_gram_ns_s)
rm(dfm_one_gram_nostem_and_stop)
```

onegram with stemming on no stopwords

```
dfm_one_gram_stem_and_nostop <- dfm(one_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_one_gram_stem_and_nostop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
one_gram_s_ns <- sort(colSums(dfm_one_gram_stem_and_nostop), decreasing = TRUE)
one_gram_s_ns <- data.frame(one_gram_s_ns)
one_gram_s_ns <- setDT(one_gram_s_ns, keep.rownames = TRUE)
write_csv(one_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(one_gram_s_ns)
rm(dfm_one_gram_stem_and_nostop)
```

no stemming or stopwords

```
dfm_one_gram_nostem_and_nostop <- dfm(one_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_one_gram_nostem_and_nostop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Course
one_gram_ns_ns <- sort(colSums(dfm_one_gram_nostem_and_nostop), decreasing = TRUE)
one_gram_ns_ns <- data.frame(one_gram_ns_ns)
one_gram_ns_ns <- setDT(one_gram_ns_ns, keep.rownames = TRUE)
write_csv(one_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(one_gram_ns_ns)
rm(dfm_one_gram_nostem_and_nostop)
```

Now we will do a bunch of bi_grams

```
##bigrams with stemming and stop words removed
bi_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbol
ns_bi_gram <- tokens(ns_one_gram, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_
dfm_bi_gram_stem_stop <- dfm(ns_bi_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
####saveRDS(dfm_bi_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
bi_gram_s_s <- sort(colSums(dfm_bi_gram_stem_stop), decreasing = TRUE)
bi_gram_s_s <- data.frame(bi_gram_s_s)
bi_gram_s_s <- setDT(bi_gram_s_s, keep.rownames = TRUE)
bi_gram_s_s <- separate(bi_gram_s_s, rn, c("word1", "word2"), sep = " ")
write_csv(bi_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Projec
```

```
rm(dfm_bi_gram_stem_stop)
rm(bi_gram_s_s)
```

bigrams with no stemming but stopwords

```
dfm_bi_gram_nostem_stop <- dfm(ns_bi_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"))
####saveRDS(dfm_bi_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Ca
bi_gram_ns_s <- sort(colSums(dfm_bi_gram_nostem_stop), decreasing = TRUE)
bi_gram_ns_s <- data.frame(bi_gram_ns_s)
bi_gram_ns_s <- setDT(bi_gram_ns_s, keep.rownames = TRUE)
bi_gram_ns_s <- separate(bi_gram_ns_s, rn, c("word1", "word2"), sep = " ")
write_csv(bi_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
rm(dfm_bi_gram_nostem_stop)
rm(bi_gram_ns_s)
rm(ns_bi_gram)
```

stemming, but no stopwords

```
dfm_bi_gram_stem_nostop <- dfm(bi_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_bi_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
bi_gram_s_ns <- sort(colSums(dfm_bi_gram_stem_nostop), decreasing = TRUE)
bi_gram_s_ns <- data.frame(bi_gram_s_ns)
bi_gram_s_ns <- setDT(bi_gram_s_ns, keep.rownames = TRUE)
bi_gram_s_ns <- separate(bi_gram_s_ns, rn, c("word1", "word2"), sep = " ")
write_csv(bi_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
rm(dfm_bi_gram_stem_nostop)
rm(bi_gram_s_ns)
```

neither stemming nor stop words

```
dfm_bi_gram_nostem_nostop <- dfm(bi_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_bi_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/C
bi_gram_ns_ns <- sort(colSums(dfm_bi_gram_nostem_nostop), decreasing = TRUE)
bi_gram_ns_ns <- data.frame(bi_gram_ns_ns)
bi_gram_ns_ns <- setDT(bi_gram_ns_ns, keep.rownames = TRUE)
bi_gram_ns_ns <- separate(bi_gram_ns_ns, rn, c("word1", "word2"), sep = " ")
write_csv(bi_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_bi_gram_nostem_nostop)
rm(bi_gram_ns_ns)
rm(bi_gram)
```

trigrams

```
##trigrams with stemming and stop words removed
tri_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbol
ns_tri_gram <- tokens(ns_one_gram, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_
dfm_tri_gram_stem_stop <- dfm(ns_tri_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
```

```
##saveRDS(dfm_tri_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
tri_gram_s_s <- sort(colSums(dfm_tri_gram_stem_stop), decreasing = TRUE)
tri_gram_s_s <- data.frame(tri_gram_s_s)
tri_gram_s_s <- setDT(tri_gram_s_s, keep.rownames = TRUE)
tri_gram_s_s <- separate(tri_gram_s_s, rn, c("word1", "word2", "word3"), sep = " ")
write_csv(tri_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
rm(dfm_tri_gram_stem_stop)
rm(tri_gram_s_s)
```

trigrams with no stemming but stopwords

```
dfm_tri_gram_nostem_stop <- dfm(ns_tri_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"
##saveRDS(dfm_tri_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Ca
tri_gram_ns_s <- sort(colSums(dfm_tri_gram_nostem_stop), decreasing = TRUE)
tri_gram_ns_s <- data.frame(tri_gram_ns_s)
tri_gram_ns_s <- setDT(tri_gram_ns_s, keep.rownames = TRUE)
tri_gram_ns_s <- separate(tri_gram_ns_s, rn, c("word1", "word2", "word3"), sep = " ")
write_csv(tri_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_tri_gram_nostem_stop)
rm(tri_gram_ns_s)
rm(ns_tri_gram)
```

stemming, but no stopwords

```
dfm_tri_gram_stem_nostop <- dfm(tri_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_tri_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Ca
tri_gram_s_ns <- sort(colSums(dfm_tri_gram_stem_nostop), decreasing = TRUE)
tri_gram_s_ns <- data.frame(tri_gram_s_ns)
tri_gram_s_ns <- setDT(tri_gram_s_ns, keep.rownames = TRUE)
tri_gram_s_ns <- separate(tri_gram_s_ns, rn, c("word1", "word2", "word3"), sep = " ")
write_csv(tri_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_tri_gram_nostem_stop)
```

```
## Warning in rm(dfm_tri_gram_nostem_stop): object 'dfm_tri_gram_nostem_stop'
## not found
```

```
rm(tri_gram_s_ns)
```

neither stemming nor stop words

```
dfm_tri_gram_nostem_nostop <- dfm(tri_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_tri_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
tri_gram_ns_ns <- sort(colSums(dfm_tri_gram_nostem_nostop), decreasing = TRUE)
tri_gram_ns_ns <- data.frame(tri_gram_ns_ns)
tri_gram_ns_ns <- setDT(tri_gram_ns_ns, keep.rownames = TRUE)
tri_gram_ns_ns <- separate(tri_gram_ns_ns, rn, c("word1", "word2", "word3"), sep = " ")
write_csv(tri_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_tri_gram_nostem_nostop)
```

```
rm(tri_gram_ns_ns)
rm(tri_gram)
```

quad grams

```
##quadgrams with stemming and stop words removed
quad_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbo
ns_quad_gram <- tokens(ns_one_gram, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove
dfm_quad_gram_stem_stop <- dfm(ns_quad_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
##saveRDS(dfm_quad_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
quad_gram_s_s <- sort(colSums(dfm_quad_gram_stem_stop), decreasing = TRUE)
quad_gram_s_s <- data.frame(quad_gram_s_s)
quad_gram_s_s <- setDT(quad_gram_s_s, keep.rownames = TRUE)
quad_gram_s_s <- separate(quad_gram_s_s, rn, c("word1", "word2", "word3", "word4"), sep = " ")
write_csv(quad_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quad_gram_stem)
```

```
## Warning in rm(dfm_quad_gram_stem): object 'dfm_quad_gram_stem' not found
```

```
rm(quad_gram_s_s)
```

quadgrams with no stemming but stopwords

```
dfm_quad_gram_nostem_stop <- dfm(ns_quad_gram, tolower = TRUE, stem = FALSE, remove = stopwords("englis
##saveRDS(dfm_quad_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Ca
quad_gram_ns_s <- sort(colSums(dfm_quad_gram_nostem_stop), decreasing = TRUE)
quad_gram_ns_s <- data.frame(quad_gram_ns_s)
quad_gram_ns_s <- setDT(quad_gram_ns_s, keep.rownames = TRUE)
quad_gram_ns_s <- separate(quad_gram_ns_s, rn, c("word1", "word2", "word3", "word4"), sep = " ")
write_csv(quad_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_quad_gram_nostem_stop)
rm(quad_gram_ns_s)
rm(ns_quad_gram)
```

stemming, but no stopwords

```
dfm_quad_gram_stem_nostop <- dfm(quad_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_quad_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/
quad_gram_s_ns <- sort(colSums(dfm_quad_gram_stem_nostop), decreasing = TRUE)
quad_gram_s_ns <- data.frame(quad_gram_s_ns)
quad_gram_s_ns <- setDT(quad_gram_s_ns, keep.rownames = TRUE)
quad_gram_s_ns <- separate(quad_gram_s_ns, rn, c("word1", "word2", "word3", "word4"), sep = " ")
write_csv(quad_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_quad_gram_nostem_stop)
```

```
## Warning in rm(dfm_quad_gram_nostem_stop): object
## 'dfm_quad_gram_nostem_stop' not found
```

```
rm(quad_gram_s_ns)
```

neither stemming nor stop words

```
dfm_quad_gram_nostem_nostop <- dfm(quad_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_quad_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
quad_gram_ns_ns <- sort(colSums(dfm_quad_gram_nostem_nostop), decreasing = TRUE)
quad_gram_ns_ns <- data.frame(quad_gram_ns_ns)
quad_gram_ns_ns <- setDT(quad_gram_ns_ns, keep.rownames = TRUE)
quad_gram_ns_ns <- separate(quad_gram_ns_ns, rn, c("word1", "word2", "word3", "word4"), sep = " ")
write_csv(quad_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pr
rm(dfm_quad_gram_nostem_nostop)
rm(quad_gram_ns_ns)
rm(quad_gram)
```

quin grams

```
##quingrams with stemming and stop words removed
quin_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symb
ns_quin_gram <- tokens(ns_one_gram, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remov
dfm_quin_gram_stem_stop <- dfm(ns_quin_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
##saveRDS(dfm_quin_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
quin_gram_s_s <- sort(colSums(dfm_quin_gram_stem_stop), decreasing = TRUE)
quin_gram_s_s <- data.frame(quin_gram_s_s)
quin_gram_s_s <- setDT(quin_gram_s_s, keep.rownames = TRUE)
quin_gram_s_s <- separate(quin_gram_s_s, rn, c("word1", "word2", "word3", "word4", "word5"), sep = " ")
write_csv(quin_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quin_gram_stem)
```

```
## Warning in rm(dfm_quin_gram_stem): object 'dfm_quin_gram_stem' not found
```

```
rm(quin_gram_s_s)
```

quingrams with no stemming but stopwords

```
dfm_quin_gram_nostem_stop <- dfm(ns_quin_gram, tolower = TRUE, stem = FALSE, remove = stopwords("englis
##saveRDS(dfm_quin_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/C
quin_gram_ns_s <- sort(colSums(dfm_quin_gram_nostem_stop), decreasing = TRUE)
quin_gram_ns_s <- data.frame(quin_gram_ns_s)
quin_gram_ns_s <- setDT(quin_gram_ns_s, keep.rownames = TRUE)
quin_gram_ns_s <- separate(quin_gram_ns_s, rn, c("word1", "word2", "word3", "word4", "word5"), sep = "
write_csv(quin_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_quin_gram_nostem_stop)
rm(quin_gram_ns_s)
rm(ns_quin_gram)
```

stemming, but no stopwords

```
dfm_quin_gram_stem_nostop <- dfm(quin_gram, tolower = TRUE, stem = TRUE)
##saveRDS(dfm_quin_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/C
quin_gram_s_ns <- sort(colSums(dfm_quin_gram_stem_nostop), decreasing = TRUE)
quin_gram_s_ns <- data.frame(quin_gram_s_ns)
quin_gram_s_ns <- setDT(quin_gram_s_ns, keep.rownames = TRUE)
quin_gram_s_ns <- separate(quin_gram_s_ns, rn, c("word1", "word2", "word3", "word4", "word5"), sep = " "
write_csv(quin_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quin_gram_stem_nostop)
rm(quin_gram_s_ns)
```

neither stemming nor stop words

```
dfm_quin_gram_nostem_nostop <- dfm(quin_gram, tolower = TRUE, stem = FALSE)
##saveRDS(dfm_quin_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
quin_gram_ns_ns <- sort(colSums(dfm_quin_gram_nostem_nostop), decreasing = TRUE)
quin_gram_ns_ns <- data.frame(quin_gram_ns_ns)
quin_gram_ns_ns <- setDT(quin_gram_ns_ns, keep.rownames = TRUE)
quin_gram_ns_ns <- separate(quin_gram_ns_ns, rn, c("word1", "word2", "word3", "word4", "word5"), sep =
write_csv(quin_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_quin_gram_nostem_nostop)
rm(quin_gram_ns_ns)
rm(quin_gram)
```

This is the end of the line