Quantedav1

Michael Pearson

1/15/2018

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(1152018)
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/qsamples
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
dfm_one_gram_stem_and_stop <- dfm(one_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
saveRDS(dfm_one_gram_stem_and_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
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(one_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english
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, ignoredFeatures = stopwords(
```

```
## Warning: Argument ignoredFeatures not used.
```

```
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, ignoreFeatures = stopwords
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_one_gram_nostem_and_nostop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera
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 Proj
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_symbols
dfm_bi_gram_stem_stop <- dfm(bi_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
saveRDS(dfm_bi_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone
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(bi_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"))
saveRDS(dfm_bi_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capst
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 = "_")
```

```
## Warning: Too many values at 392 locations: 626797, 930191, 948680, 1260504,
## 1260505, 1260507, 1311857, 1311858, 1368126, 1368127, 1390523, 1390525,
## 1409024, 1409066, 1417439, 1417440, 1433529, 1433530, 1435738, 1435740, ...
```

```
## Warning: Too few values at 4782228 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9,
## 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...
```

```
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)
```

stemming, but no stopwords

```
dfm_bi_gram_stem_nostop <- dfm(bi_gram, tolower = TRUE, stem = TRUE, ignoreFeatures = stopwords("englis
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_bi_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
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 = "_")
```

```
## Warning: Too many values at 392 locations: 601106, 875118, 891822, 1170455,
## 1170456, 1170458, 1211528, 1211529, 1256906, 1256907, 1274930, 1274932,
## 1289845, 1289884, 1296580, 1296581, 1309360, 1309361, 1311133, 1311134, ...
```

```
## Warning: Too few values at 3984060 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9,
## 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...
```

```
write_csv(bi_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
rm(dfm_bi_gram_nostem_stop)
```

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

```
rm(bi_gram_s_ns)
```

neither stemming nor stop words

```
dfm_bi_gram_nostem_nostop <- dfm(bi_gram, tolower = TRUE, stem = FALSE, ignoreFeatures = stopwords("engl
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_bi_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
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 = "_")
```

```
## Warning: Too many values at 392 locations: 626797, 930191, 948680, 1260504,
## 1260505, 1260507, 1311857, 1311858, 1368126, 1368127, 1390523, 1390525,
## 1409024, 1409066, 1417439, 1417440, 1433529, 1433530, 1435738, 1435740, ...
```

```
## Warning: Too few values at 4782236 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9,
## 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...
```

```
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)
```

trigrams

```
##trigrams with stemming and stop words removed
tri_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbol
dfm_tri_gram_stem_stop <- dfm(tri_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
saveRDS(dfm_tri_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstor
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 = "_")
```

```
## Warning: Too many values at 2404 locations: 175958, 457925, 594513, 918050,
## 931298, 999886, 1118804, 1435705, 1456240, 1575762, 1717904, 1717905,
## 1717911, 1717912, 1717913, 1717916, 1721257, 1721259, 1721262, 1733499, ...
```

```
write_csv(tri_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
rm(dfm_tri_gram_stem)
```

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

```
rm(tri_gram_s_s)
```

trigrams with no stemming but stopwords

```
dfm_tri_gram_nostem_stop <- dfm(tri_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"))
saveRDS(dfm_tri_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
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 = "_")
```

```
## Warning: Too many values at 2432 locations: 161440, 422951, 550168, 855654,
## 868327, 934036, 1047580, 1351051, 1370718, 1484979, 1620974, 1620975,
## 1620981, 1620982, 1620983, 1620986, 1624528, 1624530, 1624533, 1637619, ...
```

```
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)
```

stemming, but no stopwords

```
dfm_tri_gram_stem_nostop <- dfm(tri_gram, tolower = TRUE, stem = TRUE, ignoreFeatures = stopwords("engli
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_tri_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
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 = "_")
```

```
## Warning: Too many values at 2407 locations: 175958, 457925, 594513, 918050,
## 931298, 999886, 1118804, 1435705, 1456240, 1575762, 1717904, 1717905,
## 1717911, 1717912, 1717913, 1717916, 1721257, 1721259, 1721262, 1733499, ...
```

```
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, ignoreFeatures = stopwords("en
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_tri_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Ca
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 = "_")
```

```
## Warning: Too many values at 2435 locations: 161440, 422951, 550168, 855654,
## 868327, 934036, 1047580, 1351051, 1370718, 1484979, 1620974, 1620975,
## 1620981, 1620982, 1620983, 1620986, 1624528, 1624530, 1624533, 1637619, ...
```

```
write_csv(tri_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_tri_gram_nostem_nostop)
rm(tri_gram_ns_ns)
```

quad grams

```
##quadgrams with stemming and stop words removed
quad_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symbo
dfm_quad_gram_stem_stop <- dfm(quad_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
saveRDS(dfm_quad_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capsto
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 = "_")
```

```
## Warning: Too many values at 3190 locations: 132529, 164177, 200666, 421863,
## 762435, 775648, 947433, 947434, 947435, 947439, 947440, 947441, 947442,
## 952779, 952782, 952785, 952788, 972645, 972647, 972649, ...
```

```
write_csv(quad_gram_s_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proje
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(quad_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"))
saveRDS(dfm_quad_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
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 = "_")
```

```
## Warning: Too many values at 3215 locations: 122009, 151026, 184470, 388758,
## 705220, 717450, 876963, 876964, 876965, 876969, 876970, 876971, 876972,
## 882368, 882371, 882374, 882377, 902551, 902553, 902555, ...
```

```
write_csv(quad_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quad_gram_nostem_stop)
rm(quad_gram_ns_s)
```

stemming, but no stopwords

```
dfm_quad_gram_stem_nostop <- dfm(quad_gram, tolower = TRUE, stem = TRUE, ignoreFeatures = stopwords("eng
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_quad_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
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 = "_")
```

```
## Warning: Too many values at 3194 locations: 132529, 164177, 200666, 421863,
## 762435, 775648, 947433, 947434, 947435, 947439, 947440, 947441, 947442,
```

```
## 952779, 952782, 952785, 952788, 972645, 972647, 972649, ...
```

```
write_csv(quad_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
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, ignoreFeatures = stopwords(
```

```
## Warning: Argument ignoreFeatures not used.
```

```
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 = "_")
```

```
## Warning: Too many values at 3219 locations: 122009, 151026, 184470, 388758,
## 705220, 717450, 876963, 876964, 876965, 876969, 876970, 876971, 876972,
## 882368, 882371, 882374, 882377, 902551, 902553, 902555, ...
```

```
write_csv(quad_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pro
rm(dfm_quad_gram_nostem_nostop)
rm(quad_gram_ns_ns)
```

quin grams

```
##quingrams with stemming and stop words removed
quin_gram <- tokens(myCorpus, what = c("word"), remove_numbers = TRUE, remove_punct = TRUE, remove_symb
dfm_quin_gram_stem_stop <- dfm(quin_gram, tolower = TRUE, stem = TRUE, remove = stopwords("english"))
saveRDS(dfm_quin_gram_stem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capst
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 = "_")
```

```
## Warning: Too many values at 3961 locations: 134181, 289997, 353604, 353605,
## 353606, 353609, 353610, 353611, 353612, 353613, 364682, 364683, 364684,
## 364685, 364686, 390899, 390900, 390901, 390902, 390903, ...
```

```
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(quin_gram, tolower = TRUE, stem = FALSE, remove = stopwords("english"))
saveRDS(dfm_quin_gram_nostem_stop, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Caps
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 = "_
```

```
## Warning: Too many values at 3986 locations: 124670, 270787, 330413, 330414,
## 330415, 330418, 330419, 330420, 330421, 330422, 341532, 341533, 341534,
## 341535, 341536, 367834, 367835, 367836, 367837, 367838, ...
```

```
write_csv(quin_gram_ns_s, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quin_gram_nostem_stop)
rm(quin_gram_ns_s)
```

stemming, but no stopwords

```
dfm_quin_gram_stem_nostop <- dfm(quin_gram, tolower = TRUE, stem = TRUE, ignoreFeatures = stopwords("eng
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_quin_gram_stem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Cap
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 = "_
```

```
## Warning: Too many values at 3971 locations: 134181, 289998, 353605, 353606,
## 353607, 353610, 353611, 353612, 353613, 353614, 364683, 364684, 364685,
## 364686, 364687, 390900, 390901, 390902, 390903, 390904, ...
```

```
write_csv(quin_gram_s_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Proj
rm(dfm_quin_gram_nostem_stop)
```

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

```
rm(quin_gram_s_ns)
```

neither stemming nor stop words

```
dfm_quin_gram_nostem_nostop <- dfm(quin_gram, tolower = TRUE, stem = FALSE, ignoreFeatures = stopwords(
```

```
## Warning: Argument ignoreFeatures not used.
```

```
saveRDS(dfm_quin_gram_nostem_nostop , "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/C
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 =
```

```
## Warning: Too many values at 3996 locations: 124670, 270788, 330414, 330415,
```

```
## 330416, 330419, 330420, 330421, 330422, 330423, 341533, 341534, 341535,
## 341536, 341537, 367835, 367836, 367837, 367838, 367839, ...
```

```
write_csv(quin_gram_ns_ns, "/Users/mutecypher/Documents/Documents - Michael's iMac/Coursera/Capstone Pr
rm(dfm_quin_gram_nostem_nostop)
rm(quin_gram_ns_ns)
```

This is the end of the line