[WS](https://www.coursera.org/learn/data-science-project/profiles/371f4e9fc3ea1208226cc45c3342fc54)

Cleaning data with quanteda

wendy sarrett

[Week 3](https://www.coursera.org/learn/data-science-project/discussions/weeks/3) · [10 months ago](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A)

I switched to quanteda but now I'm having issues cleaning the data. If I remove punctuation when tokenaizing and then remove stopwords it doesn't fully work because the removed punctuation puts in symbols. I haven't found how to remove stopwords before tokenizing. This doesn't seem to be an issue with tm. Perhaps one option is to create the corpus in tm originally and switch to quanteda or just use tm.

Thoughts?

Here is what I'm seeing

text\_corpus<-corpus(text\_string)

> tok<-quanteda::tokenize(text\_corpus,remove\_numbers = TRUE,remove\_punct = TRUE)

> tok<-removeFeatures(tok,stopwords("english"))

> ng<-tokens\_ngrams(tok,n=3L)

> myDfm<-dfm(ng)

> DT\_freq<-sort(colSums(myDfm), decreasing = TRUE)

> DT\_freq[1:25]

donâ\_t\_know iâ\_m\_sure donâ\_t\_want iâ\_m\_going donâ\_t\_think itâ\_s\_just ã\_ã\_ã didnâ\_t\_know donâ\_t\_like

3494 2217 1955 1834 1785 1382 1360 1037 931

donâ\_t\_get â\_œitâ\_s â\_itâ\_s â\_â\_â itâ\_s\_like â\_œiâ\_m think\_itâ\_s didnâ\_t\_want new\_york\_city

924 912 897 895 888 876 849 844 811

itâ\_s\_time iâ\_d\_like iâ\_ve\_never now\_iâ\_m iâ\_m\_just itâ\_s\_hard donâ\_t\_need

802 751 736 726 685 682 675

0 UpvotesReplyFollow this discussion

* [**Earliest**](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A?sort=createdAtAsc&page=1)
* [**Top**](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A?sort=upvotesDesc&page=1)
* [**Most Recent**](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A?sort=createdAtDesc&page=1)

1. wendy sarrett's Post

[WS](https://www.coursera.org/learn/data-science-project/profiles/371f4e9fc3ea1208226cc45c3342fc54)

wendy sarrett

 · [10 months ago](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A/replies/lad4Gzm6EeepzgojxZP20g)

So it looks like I'm having an issue with readLines.....I'm going to have to go for another way to read in the file. When I look at the text it looks fine but when I look at the text after it's read in it's a mess...weird characters whenever there is a ' in the text:

[1000] "Apparently this is available on a DVD of Frank Tashlinâ€™s THE GLASS BOTTOM BOAT, which is vaguely apt, but it should really be an extra with VERTIGO. Both because of the ways in which Jonesâ€™s visuals approach Saul Bassâ€™s (the YouTuber who posted it apparently thinks itâ€™s by Norman McLaren â€” a fair guess, but WRONG), and in the way the short reverses the sympathies engendered in Hitchcockâ€™s film â€” a woman trapped and torn and manipulated and molded between two horrible men is replaced by a female manipulator who remodels the men in her life, rejecting the less adaptable model in favour of the one who can literally be bent to her will.

0 UpvotesHide 1 Reply

Bjenk Ellefsen's Post

[Bjenk Ellefsen](https://www.coursera.org/learn/data-science-project/profiles/96f23f6f6958c4e5c2605ed8c23f3326)

Bjenk Ellefsen

 · [10 months ago](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A/replies/lad4Gzm6EeepzgojxZP20g/comments/Etjd80LREeesRQ4YCzdAEg)

read\_lines from readr (tidyverse package) is the fastest and least trouble way of reading the txt files I have found. Ridiculously fast, in fact.



1

2

3

library(tidyverse)

blogs <- read\_lines("./data/final/en\_US/en\_US.blogs.txt")

3 Upvotes

[MV](https://www.coursera.org/learn/data-science-project/profiles/a1bbc309a9279da69c6b3f9a9fc38ce7)

Reply

Reply

1. Albert de Roos's Post

[Albert de Roos](https://www.coursera.org/learn/data-science-project/profiles/78d057de50a9fe84148a4296d654072a)

[M](https://www.coursera.org/learn/data-science-project/profiles/78d057de50a9fe84148a4296d654072a)

Albert de Roos

Mentor · [10 months ago](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A/replies/ZwSOfjlTEeeVrA7fHHMkEg)

Hi Wendy, if I glance over your code it seems that you use different packages but I think you can use only quanteda functions for the dfm and sorted lists. This may explain the strange results you get. The dfm function in quanteda lets you specify the ngrams you want (e.g. ngrams = 4) as well as remove punctuation and even stopwords. Also, you may want to have a look at the 'topfeatures' function in quanteda. As an aside, why would you want to remove the stopwords, as the input of your app probably will include stop words.

0 UpvotesHide 1 Reply

wendy sarrett's Post

[WS](https://www.coursera.org/learn/data-science-project/profiles/371f4e9fc3ea1208226cc45c3342fc54)

wendy sarrett

 · [10 months ago](https://www.coursera.org/learn/data-science-project/discussions/weeks/3/threads/GPnQ8DjpEeeSvgpcGYcJ8A/replies/ZwSOfjlTEeeVrA7fHHMkEg/comments/VG1FljlVEeeSURIzKx1jwA)

Thanks Albert, good point! Thinking about what you said, removing stop words only makes sense when you're trying to get meaningful word frequencies (as with the previous assignment.)

0 Upvotes

[MV](https://www.coursera.org/learn/data-science-project/profiles/a1bbc309a9279da69c6b3f9a9fc38ce7)