Ezra Cohen lab 11

August 4, 2021

task 1

[1]:
```
#install.packages("quanteda")
#install.packages("quanteda.textplots")
#install.packages("tm")
library(quanteda)
library(quanteda.textplots)
library(readr)
library(tm)
tweetDF<-data.frame(read_csv("https://ist387.s3.us-east-2.amazonaws.com/lab/
 ↪ClimatePosts.csv"))
str(tweetDF)
tweetDF
#It is a data frame with 3 columns and 18 rows, first is ID which on the first␣
 ↪row says what the issue is about and on every row after that says the name␣
 ↪of the person responding, skeptic is a zero or a one depending on whether␣
 ↪they agree or disagree with climate change (I don't know why science would␣
 ↪have to be something you have to agree on but This has nothing to do with␣
 ↪what we're talking about so I'll just move on) a 0 indicating that they do␣
 ↪and 1 indicating that they don't, and tweet is first the original tweet and␣
 ↪then people's replies
```

```
Package version: 3.0.0
Unicode version: 13.0
ICU version: 67.1

Parallel computing: 32 of 32 threads used.

See https://quanteda.io for tutorials and examples.

Loading required package: NLP


Attaching package: ‘NLP’


The following objects are masked from ‘package:quanteda’:
    meta, meta<-
```

```
Attaching package: ‘tm’
```

```
The following object is masked from ‘package:quanteda’:

    stopwords
```

```
  Column specification

cols(
  ID = col_character(),
  Skeptic = col_double(),
  Tweet = col_character()
)
```

```
'data.frame':   18 obs. of  3 variables:
 $ ID     : chr  "climatechange" "billmckibben" "megancollins" "neiltyson" …
 $ Skeptic: num  0 0 0 0 0 0 0 0 0 1 …
 $ Tweet  : chr  "BREAKING: Iran soars to record of 129 degrees - near hottest
ever reliably measured on Earth" "I know you're Mr. America-is-all-that-matters,
but climate is actually a global phenomenon. Here's today's glob"| __truncated__
"Could reporters stop asking if political leaders believe in climate change and
start asking if they understand it instead" "If I were ever abducted by aliens,
the first thing I’d ask is whether they came from a planet where people also
deny science." …
```

A data.frame: 18 × 3

ID	Skeptic	Tweet
<chr>	<dbl>	<chr>
climatechange	0	BREAKING: Iran soars to record of 129 degrees - near hotte
billmckibben	0	I know you're Mr. America-is-all-that-matters, but climate is
megancollins	0	Could reporters stop asking if political leaders believe in clin
neiltyson	0	If I were ever abducted by aliens, the first thing I'd ask is wh
johnbiehl	0	Alien: why should I not blow up this planet? Human: we are
ScottWesterfeld	0	Plot idea: 97% of the world's scientists contrive an environm
StephenAtHome	0	Global warming isn't real because I was cold today! Also gre
Johngcole	0	Scientist: The eclipse will be just like this... People: Wow, yo
MrGeorgeWallace	0	Global warming's for real y'all. Someday there won't be any
patrickmoore	1	The whole bogus climate crisis is not only Fake News, it's Fa
americanthinker	1	Climate alarmism is all about killing capitalism and replacin
BortSnrub	1	I honestly feel for some of these stupid people. They are livi
bfwilley	1	'We had expected more melting': Thick Arctic ice forces Nor
codenaught	1	Article claims elephants fight climate change while ignoring t
Kim147	1	Denying Climate change: Denying 2000 years of the Medieva
rumbletubble	1	If global warming was real and that the world was going to e
logicalprogressive	1	Air-conditioner maker Lennox cuts forecast, citing significant
3dogNapt	1	Another cool morning in Seattle? The Seattle PI neglects rec

task 2

```
[2]: tweetCorpus <- corpus(tweetDF$Tweet, docnames=tweetDF$ID)
```

task 3

```
[3]: tweetDFM <- dfm(tweetCorpus, remove_punct=TRUE, remove=stopwords("english"), )
```

```
Warning message:
"'dfm.corpus()' is deprecated. Use 'tokens()' first."
Warning message:
"'…' should not be used for tokens() arguments; use 'tokens()' first."
Warning message:
"'remove' is deprecated; use dfm_remove() instead"
```

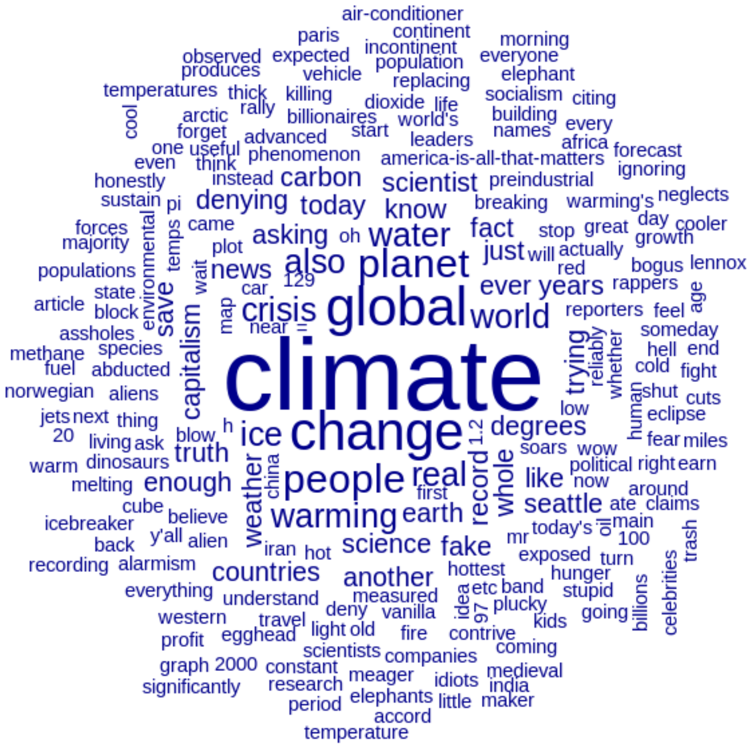
task 4

```
[4]: tweetDFM
     #It is 93.2% sparse, and it is taking all of the words and seeing how much they were used with a 1 being for every time they were used and a 0 for every word that isn't the specific word
```

```
Document-feature matrix of: 18 documents, 224 features (93.20% sparse) and 0 docvars.
                 features
docs               breaking iran soars record 129 degrees near hottest ever
  climatechange           1    1     1      1   1       1    1       1    1
  billmckibben            0    0     0      0   0       1    0       0    0
  megancollins            0    0     0      0   0       0    0       0    0
  neiltyson             0    0     0      0   0       0    0       0    1
  johnbiehl             0    0     0      0   0       0    0       0    0
  ScottWesterfeld       0    0     0      0   0       0    0       0    0
                 features
docs              reliably
  climatechange         1
  billmckibben          0
  megancollins          0
  neiltyson             0
  johnbiehl             0
  ScottWesterfeld       0
[ reached max_ndoc … 12 more documents, reached max_nfeat … 214 more features ]
```

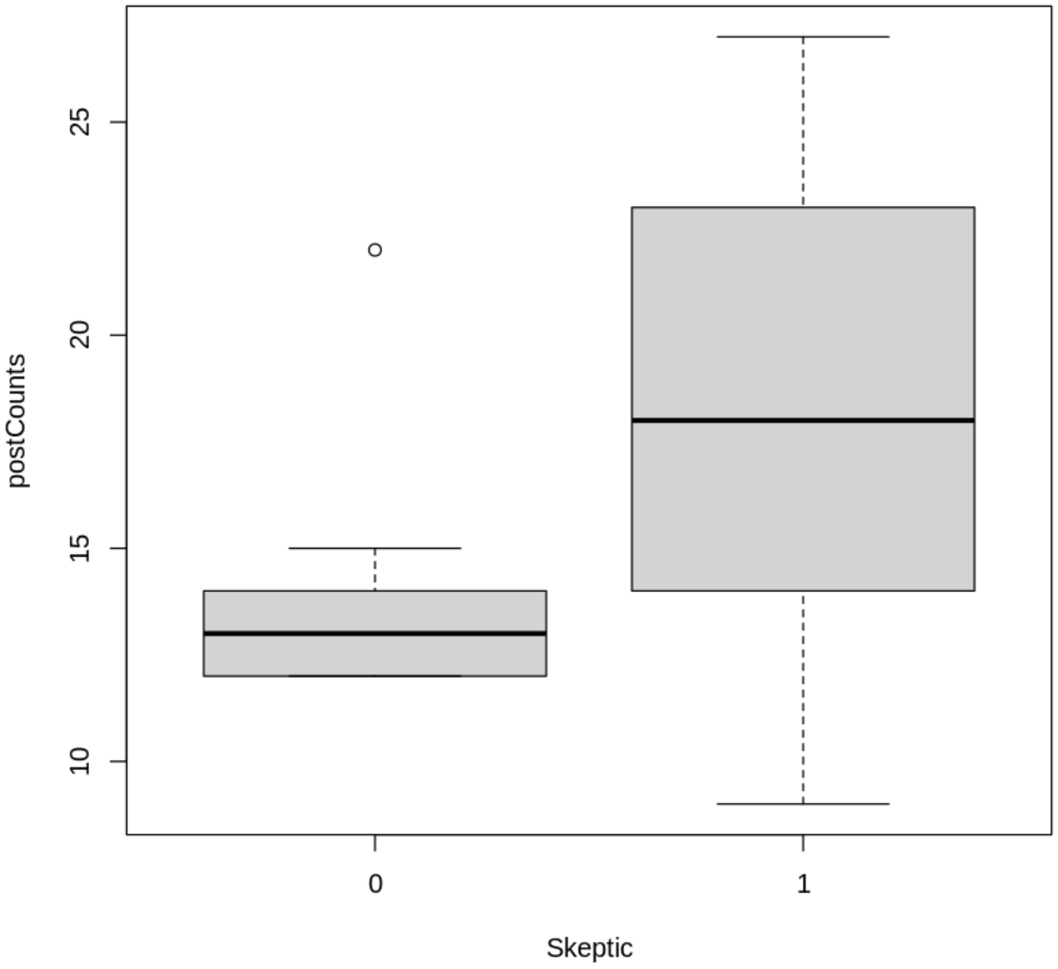
task 5

[5]:
```
textplot_wordcloud(tweetDFM, min_count = 1)
#The biggest words and therefore most used words are climate, change, global,␣
 ↪warming, people, and planet, the other notable feature other than the fact␣
 ↪that the words are organized by size depending on usage and the fact that it␣
 ↪seems that the more used words are closer to the center is just that it's␣
 ↪very nice to look at
```



task 6

[6]:
```
m <- as.matrix(tweetDFM)
postCounts <- rowSums(m)
tweetDF$postCounts <- postCounts
boxplot(postCounts ~ Skeptic, data=tweetDF)
```



task 7

[11]:

```r
URL <- "https://ist387.s3.us-east-2.amazonaws.com/data/positive-words.txt"
posWords <- scan(URL, character(0), sep = "\n")
posWords <- posWords[-1:-34] #I remembered to do this from the textbook
URL2 <- "https://ist387.s3.us-east-2.amazonaws.com/data/negative-words.txt"
negWords <- scan(URL2, character(0), sep = "\n")
negWords <- negWords <- negWords[-1:-34]
length(posWords)
length(negWords)
```

```
2006
```

```
4783
```

task 8

[12]:
```
wordCounts <- colSums(m)
wordCounts <- sort(wordCounts, decreasing=TRUE)
```

task 9

[13]:
```
str(wordCounts)
head(wordCounts)
#This has almost all of the words I specified as being the largest and␣
 ↪therefore the most common, with the addition of also and the exclusion of␣
 ↪warming, and it has all of the words and a number for each saying how many␣
 ↪times each word was used in total
```

```
 Named num [1:224] 13 6 6 4 4 3 3 3 3 3 ...
 - attr(*, "names")= chr [1:224] "climate" "global" "change" "planet" ...
```

climate 13 **global** 6 **change** 6 **planet** 4 **people** 4 **also** 3

task 10

[14]:
```
matchedP <- match(names(wordCounts), posWords, nomatch = 0)
```

task 11

[19]:
```
matchedP
length(which(matchedP!=0))
posWords[matchedP[which(matchedP!=0)]] #This last part is not specifically␣
 ↪asked for but I do it just because I can, and looking at this I actually␣
 ↪found that a bunch of these words are not used in the context that the list␣
 ↪would think of them in order to Define them as a positive word, words like␣
 ↪warm hot and cool are being used in their literal definitions and not as␣
 ↪positive descriptors
```

1. 0 2. 0 3. 0 4. 0 5. 0 6. 0 7. 0 8. 0 9. 0 10. 0 11. 0 12. 0 13. 0 14. 0 15. 0 16. 0 17. 0 18. 0 19. 0 20. 0 21. 0 22. 0 23. 0 24. 0 25. 0 26. 1088 27. 0 28. 0 29. 0 30. 0 31. 0 32. 0 33. 560 34. 0 35. 0 36. 0 37. 0 38. 0 39. 0 40. 0 41. 0 42. 0 43. 0 44. 0 45. 930 46. 1477 47. 0 48. 0 49. 0 50. 0 51. 0 52. 0 53. 0 54. 0 55. 0 56. 0 57. 927 58. 0 59. 0 60. 0 61. 0 62. 0 63. 0 64. 0 65. 0 66. 0 67. 0 68. 0 69. 0 70. 0 71. 0 72. 0 73. 0 74. 0 75. 0 76. 0 77. 0 78. 0 79. 0 80. 48 81. 0 82. 0 83. 0 84. 0 85. 0 86. 0 87. 0 88. 0 89. 0 90. 0 91. 0 92. 0 93. 0 94. 0 95. 0 96. 0 97. 0 98. 0 99. 0 100. 0 101. 0 102. 857 103. 0 104. 0 105. 0 106. 0 107. 1997 108. 1533 109. 0 110. 0 111. 0 112. 0 113. 0 114. 0 115. 0 116. 0 117. 0 118. 0 119. 0 120. 0 121. 0 122. 0 123. 0 124. 0 125. 0 126. 0 127. 0 128. 0 129. 0 130. 0 131. 0 132. 0 133. 1901 134. 0 135. 0 136. 0 137. 0 138. 0 139. 0 140. 0 141. 0 142. 0 143. 0 144. 0 145. 0 146. 0 147. 0 148. 0 149. 0 150. 0 151. 0 152. 0 153. 0 154. 0 155. 0 156. 0 157. 0 158. 0 159. 0 160. 0 161. 0 162. 0 163. 0 164. 0 165. 0 166. 0 167. 0 168. 0 169. 0 170. 0 171. 0 172. 0 173. 0 174. 0 175. 0 176. 0 177. 0 178. 0 179. 0 180. 0 181. 0 182. 0 183. 0 184. 0 185. 0 186. 1929 187. 0 188. 0 189. 0 190. 0 191. 0 192. 0 193. 0 194. 0 195. 0 196. 0 197. 0 198. 0 199. 0 200. 0 201. 0 202. 0 203. 0 204. 0 205. 0 206. 0 207. 0 208. 0 209. 0 210. 0 211. 0 212. 0 213. 0 214. 0 215. 0 216. 360 217. 0 218. 0 219. 0 220. 0 221. 0 222. 0 223. 0 224. 0

12

1. 'like' 2. 'enough' 3. 'hottest' 4. 'reliably' 5. 'hot' 6. 'advanced' 7. 'great' 8. 'wow' 9. 'right' 10. 'useful' 11. 'warm' 12. 'cool'

task 12

```
[20]: matchedN <- match(names(wordCounts), negWords, nomatch = 0)
      matchedN
      length(which(matchedN!=0))
      negWords[matchedN[which(matchedN!=0)]]#I did it here for the exact same reason␣
       ↪of just because I can, but in this case pretty much all of the words␣
       ↪actually are being used in the negative sense
```

1. 0 2. 0 3. 0 4. 0 5. 0 6. 0 7. 763 8. 0 9. 0 10. 0 11. 0 12. 0 13. 0 14. 0 15. 0 16. 0 17. 0 18. 0 19. 0 20. 0 21. 0 22. 0 23. 0 24. 0 25. 0 26. 0 27. 1574 28. 0 29. 0 30. 0 31. 0 32. 0 33. 0 34. 0 35. 0 36. 959 37. 0 38. 0 39. 0 40. 431 41. 0 42. 0 43. 0 44. 0 45. 0 46. 0 47. 0 48. 0 49. 0 50. 0 51. 0 52. 0 53. 0 54. 0 55. 0 56. 0 57. 0 58. 0 59. 0 60. 0 61. 0 62. 0 63. 0 64. 0 65. 0 66. 0 67. 0 68. 0 69. 0 70. 0 71. 0 72. 0 73. 0 74. 0 75. 0 76. 958 77. 0 78. 377 79. 0 80. 0 81. 0 82. 0 83. 0 84. 0 85. 0 86. 0 87. 0 88. 3317 89. 0 90. 0 91. 0 92. 0 93. 697 94. 0 95. 0 96. 0 97. 0 98. 0 99. 0 100. 0 101. 599 102. 0 103. 0 104. 0 105. 0 106. 0 107. 0 108. 0 109. 0 110. 0 111. 0 112. 0 113. 0 114. 0 115. 0 116. 0 117. 0 118. 0 119. 391 120. 0 121. 0 122. 0 123. 0 124. 0 125. 0 126. 0 127. 2629 128. 0 129. 0 130. 2023 131. 0 132. 0 133. 0 134. 0 135. 0 136. 0 137. 0 138. 0 139. 0 140. 4101 141. 0 142. 0 143. 0 144. 1634 145. 0 146. 0 147. 0 148. 0 149. 2831 150. 0 151. 0 152. 0 153. 0 154. 0 155. 0 156. 0 157. 0 158. 0 159. 0 160. 0 161. 0 162. 0 163. 0 164. 0 165. 0 166. 0 167. 0 168. 0 169. 2108 170. 0 171. 0 172. 0 173. 0 174. 0 175. 0 176. 0 177. 0 178. 0 179. 0 180. 0 181. 0 182. 0 183. 0 184. 0 185. 0 186. 0 187. 0 188. 0 189. 0 190. 0 191. 0 192. 0 193. 0 194. 0 195. 0 196. 0 197. 0 198. 0 199. 0 200. 0 201. 0 202. 0 203. 0 204. 4307 205. 0 206. 0 207. 0 208. 0 209. 0 210. 0 211. 0 212. 0 213. 0 214. 0 215. 0 216. 0 217. 0 218. 0 219. 0 220. 0 221. 0 222. 0 223. 0 224. 0

17

1. 'crisis' 2. 'fake' 3. 'denying' 4. 'breaking' 5. 'deny' 6. 'blow' 7. 'plot' 8. 'contrive' 9. 'cold' 10. 'bogus' 11. 'killing' 12. 'hell' 13. 'stupid' 14. 'fear' 15. 'meager' 16. 'idiots' 17. 'trash'