Assignment 1

Max Eckert

2023-09-24

Task 1.1

```
texts <- c("I don't like cricket", "You like cricket")
dfmat_diff <- texts %>% tokens() %>% dfm()
dfmat_diff
```

Text Processing Pipeline that Preserves Differences

```
## Document-feature matrix of: 2 documents, 5 features (30.00% sparse) and 0 docvars.
##        features
## docs    i don't like cricket you
##   text1 1     1    1       1   0
##   text2 0     0    1       1   1
```

```
dfmat_nodiff <- texts %>% tokens %>% tokens_remove(pattern=stopwords("en")) %>% dfm()
dfmat_nodiff
```

Text Processing Pipeline that Doesn't Preserve Differences

```
## Document-feature matrix of: 2 documents, 2 features (0.00% sparse) and 0 docvars.
##        features
## docs    like cricket
##   text1    1       1
##   text2    1       1
```

1.2

The stopword removal function of the quanteda package in the latter text processing pipeline deletes the words "I","don't", and "You". Stopword removal is a commonly employed strategy to reduce noise; considering the small number of features and the importance of the stopwords to understand the meaning of the sentences, it comes at the cost of removing too much signal as well, thereby making the two texts indistinguishable.

1.3

For a document classification task that sorts texts into two categories, those that mention cricket and those that don't, the former pipeline would be adequate. For the purposes of sentiment analysis task we would want to preserve the differences since the two documents have very different sentiments towards cricket "like" versus "don't like".

2.1

```
climatext <- c(
"Climatic change is causing adverse impacts",
```

```
"Changes in the climate have caused impacts to human systems",
"Chelsea have a goal difference of zero in the premier league this season"
)
dfm_climate <- climatext %>% tokens() %>% dfm()
```

2.2

```
sums_12 <- colSums(dfm_climate[c("text1","text2"),])
sums_12[order(sums_12, decreasing=TRUE)]
```

```
##    impacts   climatic     change         is    causing    adverse    changes
##          2          1          1          1          1          1          1
##         in        the    climate       have     caused         to      human
##          1          1          1          1          1          1          1
##    systems    chelsea          a       goal difference         of       zero
##          1          0          0          0          0          0          0
##    premier     league       this     season
##          0          0          0          0
```

When comparing texts 1 and 2, only the column 'impacts' contains a non-zero value for both texts, hence we would not estimate these texts to be very similar.

```
sums_13 <- colSums(dfm_climate[c("text1","text3"),])
sums_13[order(sums_13, decreasing=TRUE)]
```

```
##   climatic     change         is    causing    adverse    impacts         in
##          1          1          1          1          1          1          1
##        the       have    chelsea          a       goal difference         of
##          1          1          1          1          1          1          1
##       zero    premier     league       this     season    changes    climate
##          1          1          1          1          1          0          0
##     caused         to      human    systems
##          0          0          0          0
```

A simplistic measure of similarity like the one constructed above reveals no similarities between Texts 1 and 3.

```
sums_23 <- colSums(dfm_climate[c("text2","text3"),])
sums_23[order(sums_23, decreasing=TRUE)]
```

```
##         in        the       have    impacts    changes    climate     caused
##          2          2          2          1          1          1          1
##         to      human    systems    chelsea          a       goal difference
##          1          1          1          1          1          1          1
##         of       zero    premier     league       this     season   climatic
##          1          1          1          1          1          1          0
##     change         is    causing    adverse
##          0          0          0          0
```

Lastly, texts 2 and 3 are revealed to have most similarities in terms of features contained in the dfm without any additional preprocessing steps.

2.3

```
dfm_preserv <- climatext %>% tokens() %>% tokens_remove(pattern=stopwords("en")) %>% tokens_wordstem() %
print(dfm_preserv, max_ndoc = 3, max_nfeat = 14)
```

```
## Document-feature matrix of: 3 documents, 14 features (57.14% sparse) and 0 docvars.
##        features
## docs    climat chang caus advers impact human system chelsea goal differ zero
##   text1      1     1    1      1      1     0      0       0    0      0    0
##   text2      1     1    1      0      1     1      1       0    0      0    0
##   text3      0     0    0      0      0     0      0       1    1      1    1
##        features
## docs    premier leagu season
##   text1       0     0      0
##   text2       0     0      0
##   text3       1     1      1
```

2.4

The above text processing pipeline includes more steps, such as stopword removal, and word stemming, which in turn reduces the vocabulary, since words like "in", "the" and "have" are removed and words with the same meaning are combined. In this case, the removal creates a more useful representation since the mentioned stopwords do not only add noise, but could in fact be misleading, since the high numerical value revealed a supposed similarity between Text 2 and 3, which we know to not be thematically related (Premier League Football vs. Climate Change). After removing the stopwords, we do not observe any columns that have a non-zero value between Texts 1 and 2 vis-a-vis Text 3, which confirms the similarity we we observed intuitively.

In addition, stemming generated useful roots for the worlds "climate", "change", and "cause" which were previously dismissed by the dfm as unrelated, because they appeared in its adjective or plural form, such as "climatic" or "changes". As a result, the most closely related texts are now Text 1 and Text 2 with 4 non-zero feature columns.